Characterisation of a Susceptibility Locus for Inflammatory Arthritis

A thesis submitted to The University of Manchester for the degree of PhD

In the Faculty of Medical and Human Sciences

2014

Kathryn Jean Audrey Steel

School of Medicine

Table of Contents

List of Figures 8
List of Tables 10
Abbreviations 11
Abstract 15
Acknowledgements 17
1.1.1 Rheumatoid Arthritis 20
1.1.2 Juvenile Idiopathic Arthritis 20
1.1.3 Psoriatic Arthritis 21
1.2 Pathogenesis of Inflammatory Arthritis 22
1.2.1 Cellular Infiltrate 25
1.2.2 Autoantibodies 26
1.2.3 Cytokine Disequilibrium 28
1.2.4. Animal models and the pathogenesis of IA 30
1.3 Aetiology of Inflammatory Arthritis 31
1.4 Identification of Disease Susceptibility Genes 35
1.4.1 Case Control Studies 35
1.4.2 Linkage Disequilibrium 37
1.4.3 Genome Wide Association Studies 38
1.4.4 Meta-analysis 41
1.4.5 Population stratification 42
1.4.6 Common Disease Common Variant Hypothesis 44
1.4.7 Missing Heritability 45
1.5 Identification of a Causal Variant 48
1.5.1 Imputation 49
1.5.2 DNA Resequencing 49
1.5.3 Fine Mapping 50
1.6 From Genotype to Phenotype 51
1.7 Genetics of inflammatory arthritis 54
1.7.1 RA Genetics 54
1.7.1.1 The Major Histocompatibility Complex (MHC) 55
1.7.1.2 Non-MHC RA Loci 56
1.7.2 JIA genetics 59
1.7.2.1 The Major Histocompatibility Complex (MHC) 60
1.7.2.2 Non-MHC JIA Loci 61
1.7.2.2.1 Candidate gene studies 61
1.7.2.2.2 GWA studies 61

1.7.3 PsA Genetics 63
1.7.3.1 The Major Histocompatibility Complex (MHC) 63
1.7.3.2 Non-MHC PsA loci 64
1.7.4 Overlap of Susceptibility Loci 66
1.7.4.1 The Concept of Shared Loci 66
1.8 Inflammatory arthritis overlapping regions 70
1.8.1 PTPN22 70
1.8.2 STAT4 71
1.8.3 ATXN2/SH2B3 72
1.8.4 TNFAIP3 72
1.8.5 TRAF1/C5 73
1.8.6 IL2RA 74
1.8.7 IL2/IL21 74
1.9 Summary 75
1.10 The Immunochip 76
1.11 Aims of study 78
1.12 Objectives 78
2.1.1 Inflammatory arthritis overlap 80
2.1.2 Subjects 80
2.1.3 Illumina Infinium HD assay genotyping 81
2.1.3.1 DNA quality control 84
2.1.3.2 Amplification of DNA 85
2.1.3.3 Fragmentation and precipitation of DNA 85
2.1.3.4 Resuspension of DNA and hybridisation to bead chip array 85
2.1.3.5 Washing of bead chip array 86
2.1.3.6 Single base extension and bead chip staining 86
2.1.3.7 Imaging of bead chip array on iScan system 88
2.1.4 Immunochip SNP and sample QC 88
2.1.4.1 Identity by descent analysis 89
2.1.4.2 Principal Components Analysis 89
2.1.4.3 Hardy-Weinberg Equilibrium 90
2.1.4.4 Association analysis 91
2.1.5 Power of each disease to detect genetic effects 91
2.1.6 Calculating the number of inflammatory arthritis overlapping regions 92
2.1.7 Identifying correlation between SNPs in overlapping regions 92
2.1.8 Selecting a functionally promising region for further analysis 92
2.1.9 Functional annotation 93
2.2 Replication of overlapping associations 94

2.2.1 SNP Assay Design 95
2.2.2 Subjects 96
2.2.3 Genotyping using the Sequenom MassARRAY Platform 96
2.2.3.1 Amplifying DNA for genotyping 97
2.2.3.2 Agarose gel electrophoresis 99
2.2.3.3 SAP Treatment 99
2.2.3.4 IPlex Reaction 100
2.2.3.5 Conditioning the iPlex reaction products — clean resin 102
2.2.3.6 Dispensing sample onto the SpectroChip arrays 103
2.2.4 Calling SNP genotypes 103
2.2.5 Sample and SNP QC 103
2.2.6 Association testing 104
2.3 *RUNX1* replication and fine mapping 104
2.3.1 Defining the region for fine mapping 104
2.3.2. Calculation of coverage for the selected region on the Immunochip array 104
2.3.3 Subjects 105
2.3.4 Tag SNP selection and assay design 105
2.3.5 Genotyping using the Sequenom MassARRAY system 105
2.3.6 Calling SNP genotypes 105
2.3.7 Sample and SNP QC 106
2.3.8 Association testing 106
2.3.9 Identification of multiple effects in the selected region 106
2.4 Functional Analysis of the selected region 106
2.4.1.1 Subjects 107
2.4.1.2 SNP genotyping using Taqman allelic discrimination assays 108
2.4.1.2.1 Extraction of DNA for genotyping 108
2.4.1.2.2 Plating out of DNA for genotyping 109
2.4.1.2.3 Preparing the reaction mastermix 109
2.4.1.3. Calling of genotypes using the Quant studio RT-PCR software 110
2.3.1.4 Whole blood gene expression analysis 110
2.3.1.5 Design of selected gene and endogenous controls gene expression assays 112
2.3.1.6 Subjects for gene expression analysis 112
2.3.1.7 Total RNA quality control 112
2.3.1.8.1 Preparing the gel and gel dye mix 113
2.3.1.8.2 Loading the gel dye mix 114
2.3.1.8.3 Loading the Nanomarker, RNA ladder and samples onto the chip . 114
2.3.1.8.4 Running the Nanochip 114

2.3.1.9 cDNA conversion using High-capacity cDNA Reverse Transcription Kit 114
2.4.1.10 Gene expression analysis of selected gene and endogenous controls 115
2.4.1.11 Whole blood eQTL analysis 117
2.4.2.1 Subjects 118
2.4.2.2 Genotyping of samples 118
2.4.2.3 Sample collection for PBMC extraction 119
2.4.2.4 Cell counts and viability checks 120
2.4.2.4.2 Trypan blue exclusion 122
2.4.2.4.2.1 Assessing viability using Trypan Blue 122
3.4.2.5 Cryopreservation and thawing of PBMCs 122
3.4.2.5.1 Freezing of PBMC samples 123
3.4.2.5.2 Thawing of PBMC samples 123
2.4.2.6 Separation of PBMCs into T lymphocyte subsets 124
2.4.2.6.1 Positive selection 124
2.4.2.6.2 Negative selection 124
2.4.2.7 Assessment of cell viability and purity using flow cytometry 126
2.4.2.7.1 Using flow cytometry to analyse T lymphocyte subsets 126
2.4.2.7.2 Staining of cells for flow cytometry 127
2.4.2.7.3 Flow cytometry analysis 128
overlap 131
2.4.2.8 Extracting total RNA from cell subset suspensions 132
2.4.2.9.1 RNA quality control using the Agilent bioanalyzer 2100 133
2.4.2.10 DNase treatment of total RNA 133
2.4.2.11.1 Reverse transcription to synthesize First Strand cDNA 136
2.4.2.11.2 Second strand cDNA synthesis 137
2.4.2.11.3 cDNA purification 138
2.4.2.12 Illumina Gene Expression Direct Hybridization Assay 139
2.4.2.12.1 Hybridization to the bead chip 140
2.4.2.12.2 Washing beadchip 140
2.4.2.13 Detecting differential signals on array 141
2.4.2.14 Gene expression data normalisation 141
2.4.2.15 Calculation of the signal to noise ratio across arrays 142
2.4.2.16 Calculation of intensity signals across probes 142
2.4.2.17 Calculation of the proportion of probes expressed by each sample 142
2.4.2.18 Matching probes to hg19 genome build 143
2.4.2.19 Identification of sample outliers 143
2.4.2.20 Principal components analysis 143

2.4.2.21 Array weighting 143
2.4.2.22 Cell specific eQTL analysis 144
3.0 Results 146
3.1 Inflammatory arthritis overlap 146
3.1.1 Immunochip SNP and sample QC 146
3.1.2 Power of each cohort to detect genetic effects 151
3.1.3 Calculating the number of inflammatory arthritis overlapping regions 151
3.1.4 Identification of correlation between SNPs in overlapping regions 157
3.1.5 Selecting a functionally promising region for further analysis 163
3.1.6 RUNX1 functional annotation 163
3.1.6.1 RUNX1 eQTL analysis 164
3.2 Replication of overlapping associations 170
3.2.1 Selection of genetic regions for replication 170
3.2.2 SNP assay design 170
3.2.3 Subjects 171
3.2.4 Power calculations pre-QC 174
3.2.5 Genotyping using the Sequenom MassARRAY platform 174
3.2.6.1 Issues with DNA quality 176
3.2.7 Sample and SNP QC 178
3.2.8 Post-QC power calculations 180
3.2.9 Association testing 180
3.3 RUNX1 replication and fine mapping 183
3.3.1 Defining the region for fine mapping 183
3.3.2 Calculating *RUNX1* SNP coverage on the Immunochip array 184
3.3.3 Subjects 185
3.3.4 Pre-QC power calculations 185
3.3.5 Tag SNP selection and assay design 185
3.3.6 Genotyping using the Sequenom MassARRAY system 187
3.3.7 Calling SNP genotypes 187
3.3.7.1 Issues with DNA quality 188
3.3.8 Sample and SNP QC 189
3.3.9 Post-QC power calculations 191
3.3.10 Association testing 191
3.3.11 Identification of multiple effects in the *RUNX1* region 197
3.4 Functional analysis of the *RUNX1* region 200
3.4.1 eQTL analysis of the *RUNX1* region in whole blood 200
3.4.1.1 Subjects 200
3.4.1.2 SNP genotyping using Taqman allelic discrimination assays 201

3.4.1.3 Calling of genotypes using the Quant studio RT-PCR software 201
3.4.1.4 Design of RUNX1 and control gene expression assays 202
3.4.1.5 Subjects for gene expression analysis 203
3.4.1.6 Total RNA quality control 203
3.5.1.7 cDNA conversion of RNA samples 205
3.4.1.8 Gene expression analysis of *RUNX1* and endogenous controls 205
3.4.1.9 Whole blood eQTL analysis 206
3.4.2 eQTL analysis of the *RUNX1* region in T lymphocytes 208
3.4.2.1 Subjects 208
3.4.2.2 Genotyping of samples 208
3.4.2.3 Sample collection for PBMC extraction 209
3.4.2.4 Cell count and viability checks 209
3.4.2.5 Cryopreservation and thawing of PBMCs 211
3.4.2.6 Separation of PBMCs into T lymphocyte subsets 211
3.4.2.7 Assessment of viability and cell purity using flow cytometry 213
3.4.2.8 Purity of cell populations 218
3.4.2.9. Extracting total RNA from cell subset suspensions 222
3.4.2.10 RNA quality control 222
3.4.2.11 DNase treatment of Total RNA 226
3.4.2.12 RNA amplification using Illumina TotalPrep Amplification Kit 226
3.4.2.13 Illumina Gene Expression Direct Hybridization Assay 227
3.4.2.14 Detecting differential signals on array 227
3.4.2.15 Gene expression normalization and QC 227
3.4.2.16 Calculation of the signal to noise ratio across arrays 228
3.4.2.17 Calculation of the intensity signals across probes 229
3.4.2.18 Calculation of the proportion of probes expressed by each sample 231
3.4.2.19 Matching probes to hg19 transcripts 231
3.4.2.20 Identification of sample outliers 231
3.4.2.21 Principal components analysis 232
3.4.2.22 Array weighting 233
3.4.2.23 Cell specific eQTL analysis 234
4.0 Discussion 240
4.1 Summary of findings 240
4.2. Findings, strengths and weaknesses of the study 241
4.2.1 Immunochip overlap 241
4.2.2 Immunochip replication 248
4.2.3 RUNX1 fine mapping and replication 251
4.2.4 Whole blood eQTL analysis 253

4.2.5 eQTL analysis in T lymphocytes 256
4.3. Implications of study 262
4.3.1 Immunochip overlap 262
4.3.2 Immunochip replication 264
4.3.3 RUNX1 replication and fine mapping 265
4.3.4 RUNX1 eQTL analysis in whole blood and T lymphocytes 266
4.4 Future Work 268
4.5 Conclusion 277
5. Appendix 279
5.1 Tempus spin RNA isolation kit Tempus Spin RNA Isolation Kit 279
5.1.1 Processing of stabilized blood 279
5.1.2 Purification of RNA 279
5.1.3 RNA QC 280
5.1.3.1 RNA quality control using Nanodrop N-1000 280
6.0 References 282

Word count: 78,127 words

List of Figures

Figure 1 - Joint inflammation in RA, JIA and PsA....19
Figure 2 - Hand x-rays of patients with RA, JIA and PsA....25
Figure 3 - Diagram of normal and disease affected joint (Strand et al. 2007).....29
Figure 4 – Induction of autoimmunity in rheumatic diseases....33
(Deane and El-Gabalawy 2014). 33
Figure 5 - Summary of stages involved in identifying and characterising a disease-associated locus....34
Figure 6 – SNP polymorphism....36
Figure 7 – Associations identified by GWA studies....40
Figure 8– PCA analysis of Hapmap populations (Heath et al. 2008)....44
Figure 9 - Overlapping regions prior to the Immunochip study....69
Figure 10 – Immune mediated diseases which contributed to the Immunochip....77
Figure 11 – Illumina workflow 83
Figure 12 – Formula to determine DNA sample concentration and purity....84
Figure 13– Sequenom assay workflow....97
Figure 14 – Taqman allelic discrimination workflow....108
Figure 15 – Taqman gene expression chemistry....111
Figure 16– Ficoll separation layers....119
Figure 17– CASY cell counter current exclusion....121
Figure 18– Cell viability equation....122
Figure 19– Gating for analysis of purity of CD4+ and CD8+ samples....130
Figure 20– TotalPrep amplification workflow....135
Figure 21– SNP QC for each disease....146
Figure 22– Sample QC for each disease....149
Figure 23– Distribution of overlap between diseases....157
Figure 24 – *RUNX1* region association plots....159
Figure 25– IL2RA region association plots....160
Figure 26– rs9979383 region TF binding....166
Figure 27– rs8129030 region TF binding....167
Figure 28– SNP calls from Typer 4.0....175
Figure 29 – High quality DNA gel and genotyping....176
Figure 30– Low quality DNA gel and genotyping....177
Figure 31 – SNP and sample QC summary....179
Figure 32 – *RUNX1* region selected for fine mapping....183
Figure 33– Haploview tagger results for 51 SNPs....186
Figure 34 – SNP calls from Typer 4.0....187
Figure 35– High quality DNA gel and genotyping....188
Figure 36 – Low quality gel and genotyping....189
Figure 37 – Genotyping QC stage I and stage II....190
Figure 38 – Locus zoom plot of the *RUNX1* region....195
Figure 39– Odds ratio forest plot....196
Figure 40– Association plot when conditioned on rs9979383 showing no independent effects....199
Figure 41– Genotype calls using Quant studio RT-PCR software....201
Figure 42– ENSEMBL gene browser showing *RUNX1* splice variants....203
Figure 43– Electropherogram and gels from 2 healthy controls samples. 204
Figure 44 – Amplification plot for *RUNX1* and housekeeping genes in whole blood 205
Figure 45– QC summary for eQTL analysis....206
Figure 46– eQTL analysis of *RUNX1* region in whole blood....207

Figure 47– PBMC yield from healthy control bloods 210
Figure 48– CD4+ lymphocyte yield 212
Figure 49 CD8+ lymphocyte yield 213
Figure 50– Plots showing viability of CD8+ and CD4+ lymphocytes 215
Figure 51- CD8+ lymphocyte viability across all samples 216
Figure 52- CD4+ lymphocyte viability across all samples 217
Figure 53– Histogram plots showing CD8+ population purity 218
Figure 54– Plots showing CD4+ population purity using PE and Vioblue flourochromes 220
Figure 55 – CD8+ lymphocyte purity 221
Figure 56– CD4+ lymphocyte purity 222
Figure 57– Bioanalyzer traces of 2 healthy control samples 226
Figure 58– Gene expression normalisation and QC 228
Figure 59 – Signal to noise ratios for 45 samples 229
Figure 60– Average signal intensity in raw and normalized data 230
Figure 61– MDS plot showing clustering by sample type 232
Figure 62– Contribution of principal components to sample variance 233
Figure 63- QC summary for eQTL analysis 235
Figure 64– *RUNX1* region eQTL analysis in CD8+ lymphocytes 237
Figure 65– *RUNX1* region eQTL analysis in CD4+ lymphocytes 238

List of Tables

Table 1 – Summary of similarities and differences between RA, JIA and PsA 23
Table 2 – Total number of samples included in Immunochip analysis 81
Table 3 – Beadchip xStain stages 87
Table 4– PCR reaction mastermixes 98
Table 5 – PCR reaction cycles 99
Table 6 – SAP enzyme mastermixes 100
Table 7 – SAP reaction cycles 100
Table 8– iPlex reaction mastermixes 101
Table 9 – iPlex reaction cycles 102
Table 10 – Taqman genotyping reagents 109
Table 11 - Allelic discrimination assay reaction times 110
Table 12– cDNA conversion volumes 115
Table 13– cDNA conversion thermo-cycling 115
Table 14–Gene expression reaction mastermix 116
Table 15- Gene expression reaction thermo cycling conditions 116
Table 16 - Antibody cocktails added to first group of samples (n=6) 128
Table 17 - Antibody cocktails added to remaining group of samples (n=17) 132
Table 18– Reverse transcription mastermix 136
Table 19– Reverse transcription reaction times 136
Table 20– Second strand transcription mastermix 137
Table 21– Second strand transcription reaction times 137
Table 22– IVT transcription mastermix 138
Table 23 – IVT reaction times 139
Table 24– Beadchip wash steps 141
Table 25– Summary of total SNP and samples available for each disease post QC 150
Table 26 – Power to detect genetic effects with OR = 1.2 151
Table 27 – Regions associated with multiple types of Inflammatory Arthritis 153
Table 28 – Correlation between index SNPs in overlapping regions 161
Table 29– *RUNX1* eQTL analysis 164
Table 30– Demographics for 3879 cases and 2561 controls 171
Table 31 –Immunochip regions selected for overlap replication 172
Table 32– SNPs included on Immunochip overlap replication 173
Table 33 – Allelic association results from Overlap Replication 181
Table 34 – SNP capture of the RUNX1 region on the Immunochip array 184
Table 35 – SNP capture of the *RUNX1* fine mapping region with 2 assays 186
Table 36 – Allelic association testing for *RUNX1* fine mapping genotyping 192
Table 37– Association statistics for rs9979383 compared to the Immunochip Study 194
Table 38– Conditional logistic regression results 197
Table 39– Demographics for 75 subjects 200
Table 40- Genotypic distribution of genotype calls in healthy controls 202
Table 41– Demographics of 23 samples from the NRHV cohort 208
Table 42– Genotype distribution for 23 healthy controls 209
Table 43- Characteristics of extracted RNA 224
Table 44– eQTL results for *RUNX1* region 236

Abbreviations

A proliferation inducing ligand (APRIL)

American College of Rheumatology (ACR)

Ankylosing Spondylitis (AS).

Anti-carbamylated antibodies (anti-CarP)

Anti-citrullinated peptide antibodies (ACPA)

Antigen presenting cells (APC)

Anti-nuclear antibodies (ANAs)

B lymphocyte stimulator (BLyS)

Beta actin (ACTNB)

Bovine serum albumin (BSA),

Celiac Disease (CD)

Chromatin immunoprecipitation (ChIP)

Chromatin immunoprecipitation resequencing (ChIP-seq)

Chromosome conformation capture (3C)

Common disease common variant hypothesis (CD/CV),

Complement component 5 (C5)

Complementary DNA (cDNA)

Complementary RNA (cRNA)

Copy number variations (CNVs)

Crohn's disease (CrD)

Cross phenotype meta-analysis (CPMA)

Cytotoxic T-Lymphocyte Antigen 4 (CTLA4)

Dendritic cells (DCs)

Dideoxynucleotide triphosphate (ddNTP).

Disease modifying anti-rheumatic drugs (DMARDS)

DNA resequencing (DNA-seq)

Encyclopaedia of DNA elements (ENCODE)

Ethylenediaminetetraacetic acid (EDTA)

European league against rheumatism (EULAR)

Expression quantitative trait locus (eQTL)

False discovery rate (FDR)

Fluorescence activated cell sorting (FACS)

Foetal bovine serum (FBS)

Fragment crystallisable portion (Fc portion)

Genome wide association studies (GWA studies)

Genomic control (GC)

Genotype database (GDB)

Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)

Hardy Weinberg Equilibrium (HWE).

Human leukocyte antigen (HLA)

Identity by descent (IBD)

Immunoglobulin G (IgG)

Inflammatory arthritis (IA)

Interleukin 1 (IL-1)

Interleukin 13 (IL13)

Interleukin 18 (IL-18)

Interleukin 2 (IL-2)

Interleukin 2 receptor alpha (IL2RA)

Interleukin 21 (IL-21)

Interleukin 33 (IL-33)

Interleukin 6 (IL-6)

International League of Associations for Rheumatology (ILAR)

Janus kinase-signal transducer and activator of transcription (JAK-STAT)#

Juvenile chronic arthritis (JCA),

Juvenile idiopathic arthritis (JIA)

Linkage disequilibrium (LD)

Magnetic activated cell sorting (MACS)

Major histocompatibility complex (MHC)

Matrix metalloproteinase (MMP)

Minor allele frequency (MAF)

Multiple Sclerosis (MS).

Myotubularin-related protein 3 (MTMR3)

National repository healthy volunteers (NRHV)

Nod like receptors (NLRs)

Non-steroidal anti-inflammatory drugs (NSAIDs)

Nuclear factor kappa-light-chain-enhancer of activated B cells (NFKB)

Odds ratio (OR)

Peptidyl deaminase (PAD)

Peripheral blood mononuclear cell (PBMC)

Phosphate buffer saline (PBS)

Polymerase chain reaction (PCR)

Principal components analysis (PCA)

Protein tyrosine phosphatase 22 (PTPN22)

Psoriasis Vulgaris (PsV)

Psoriatic arthritis (PsA)

Quality control (QC)

Receptor activator of nuclear factor kappa-B ligand (RANKL)

Retinol binding protein 5 (RBP5)

Reverse transcription (RT)

Rheumatoid arthritis (RA)

Rheumatoid factor (RF)

Ribosomal RNA (rRNA)

RNA integrity number (RIN)

RNA resequencing (RNA-seq)

Runt related transcription factor 3 region (RUNX3)

Sequence by synthesis (SBS)

Shrimp alkaline phosphatase (SAP)

Signal transducer and activator of transcription 2 (STAT2)

Signal transducer and activator of transcription 4 (STAT4)

Single nucleotide polymorphism (SNP)

Systemic Lupus Erythematosus (SLE)

Systemic sclerosis (Ssc)

T regulatory cells (Tregs)

The ataxin 2/SH2B adaptor protein 3 (ATXN2/SH2B3)

TNF Receptor Associated Family (TRAF)

Toll like receptors (TLRs)

Transcription factor (TF)

Transcription factor binding sites (TFBS)

Tumour necrosis factor (TNF)

Tumour necrosis factor associated factor 1 (TRAF1)

Tumour necrosis factor, alpha-induced protein 3 (TNFAIP3)

Type 1 diabetes (T1D)

Tyrosine kinase 2 (TYK2)

United Kingdom Rheumatoid Arthritis Genetics Consortium (UKRAG)

Utah residents with Northern and Western European ancestry (ceph/CEU)

Wellcome Trust case control consortium (WTCCC)

Abstract

The University of Manchester, Kathryn Jean Audrey Steel, PhD, Characterisation of a Susceptibility Locus for Inflammatory Arthritis, 2014.

Inflammatory arthritis (IA) types such as rheumatoid arthritis (RA), juvenile idiopathic arthritis (JIA) and psoriatic arthritis (PsA) have been shown to exhibit common clinical features. As complex diseases, they have a known genetic component, some of which is known to be shared. The aim of this study was to assess the genetic overlap between 3 types of IA (RA, JIA and PsA) using genotype data generated on the Immunochip array and to select a biologically promising overlapping region for further genetic and functional investigation.

Overlap analysis was performed using association data generated for a large cohort of inflammatory arthritis cases and shared controls (11,475 RA; 2816 JIA; 929 PsA respectively). 50 genetic regions were identified as being associated with more than 1 type of IA ($p<1x10^{-3}$), with several interesting similarities and differences observed between the diseases. As several of the overlapping regions detected represented novel disease associations, they required replication in an independent sample cohort. 12 variants were selected for replication in an independent RA cohort of 3879 cases and 2561 controls. Of these, 2 variants in the *CTLA4* and *MTMR3* regions were successfully replicated in RA at $p<0.05$.

Bioinformatics analysis was performed for the 50 overlapping regions, with one particularly promising region, *RUNX1,* selected for further investigation. In this region, the same variant (rs9979383) is associated across the 3 diseases, with similar odds ratios (OR 0.8-0.9) observed in each disease. As this region represented both a novel IA association and had not been, densely genotyped on the Immunochip array, fine mapping was performed by genotyping 51 SNPS in 3491 cases and 2359 controls. This resulted in replication of the association at rs9979383 ($p=0.02$) with no additional significant genetic effects detected; therefore, this variant was selected for further functional analysis.

As rs9979383 lies ~280kb upstream of the *RUNX1* gene, a cis-eQTL analysis was performed to identify if the variant acts by regulation of *RUNX1* gene expression. This was performed in whole blood, CD4+ and CD8+ lymphocytes from 75 (and a subset of 23) healthy volunteers respectively. No significant eQTLs were detected between rs9979383 and *RUNX1* in whole blood ($p =0.9$) or *RUNX1/LOC100506403* CD4+ and CD8+ lymphocytes ($p=0.1$).

This study has provided insight into the genetic similarities and differences between different types of inflammatory arthritis, which can be applied to further investigations into disease susceptibility. Although no significant cis-eQTL was detected in any of these tissues with either *RUNX1* or the nearby lnc-RNA *LOC100506403,* in cells from healthy volunteers under unstimulated conditions, these findings will direct future functional investigations into the role of this overlapping region in the susceptibility of IA.

Declaration

I declare that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning

Copyright Statement

i The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and she has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on Presentation of Theses

Acknowledgements

I would like to thank my supervisors Professor Wendy Thomson, Professor Anne Barton and especially Dr Anne Hinks who have both supported and inspired me. I can't describe how much I have learned from them. I would also like to thank my advisor Dr Rebecca Dearman for all her advice.

The time I have spent within the Arthritis Research UK centre for Genetics and Genomics has been truly wonderful. I would like to thank everyone for all their support but a special mention to Dr Steve Eyre, Dr Annie Yarwood and Amanda McGovern who have given me science advice in the most stressful of times whilst making me laugh along the way.

A big special thank you to all the PhD students within the department, I could not have asked for a better bunch of people to make my time in Manchester so fantastic. The biggest thanks should go to Nisha for sharing both an office and a home with me but also Simon deserves a special mention as he has made me smile every day I have been here.

Lastly, I would like to thank my parents and brother Joe who have never doubted me. I am so lucky to be surrounded by such great friends and a fantastic partner in Holly. Thank you so much for all your support and love.

1.0 Introduction

1.0 Introduction

1.1 Epidemiology of Inflammatory Arthritis

Inflammatory arthritis (IA) describes a group of diseases which share common clinical features such as articular joint manifestations (Figure 1) and response to treatments such as disease modifying anti-rheumatic drugs (DMARDS) and biologic therapies. Collectively, IA includes rheumatoid arthritis (RA), juvenile idiopathic arthritis and psoriatic arthritis (PsA) as well as other diseases but, throughout this thesis, the term will be restricted to encompass just the RA, JIA and PsA. As complex diseases, the presence of such common characteristics indicates that IA may also share aetiology via common genetic and environmental susceptibility factors. Identification of overlapping genetic factors would give a greater insight into the common pathways contributing to susceptibility. This in turn could be used to identify both biomarkers to improve disease classification and provide candidates which could be targeted by shared therapeutics. Essentially it would provide the opportunity for more targeted therapies and would reduce the need for the immunosuppressive treatments which are utilised currently to treat these diseases.

Figure 1 - Joint inflammation in RA, JIA and PsA

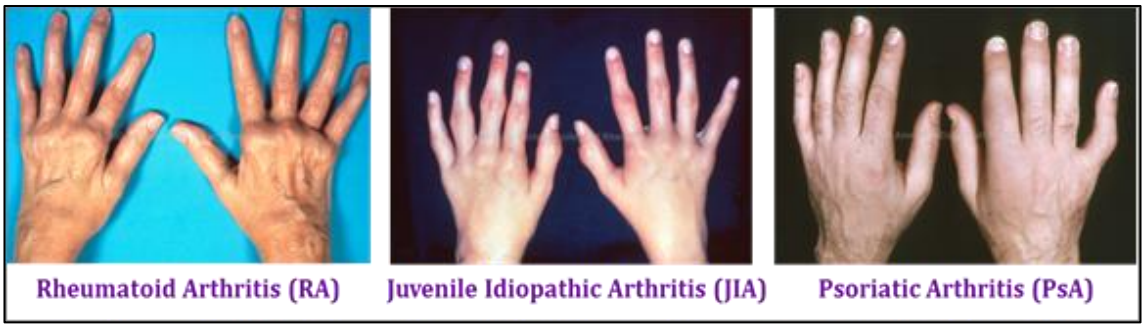


Figure 1 shows inflammation of the hand joints (L-R) in patients with RA, JIA and PsA (http://images.rheumatology.org/). Articular joint inflammation represents an overlapping clinical feature of these 3 types of IA.

1.1.1 Rheumatoid Arthritis

Rheumatoid arthritis (RA) is a chronic inflammatory disease initiated by a breach of immune tolerance and infiltration of autoimmune cells into self-tissue. Estimated to affect between 0.5% and 1% of the UK population, RA prevalence is approximately three times higher in women compared to men (Symmons et al. 2002) . Primarily affecting the synovial joints, RA is also associated with risk of several extra-articular comorbidities including coronary heart disease and pulmonary fibrosis (Arts et al. 2014;Young et al. 2007), with limitation of quality of life often representing the most debilitating feature of the disease. (Campbell et al. 2012). Like many autoimmune diseases, the heterogeneity of RA often makes accurate diagnosis and treatment challenging. Since 1958 several series of classification criteria have been generated by the American College of Rheumatology/European League against Rheumatism RA (ACR/EULAR). These were most recently updated in 2010 but the majority of patients have been classified using the earlier, 1987 ACR, criteria (Aletaha et al. 2010;Arnett et al. 1988;Ropes M.W et al. 1958). In the early stages of RA anti-inflammatory non-steroidal anti-inflammatory drugs (NSAIDs) and corticosteroids form the mainstay of treatment. In progressive and severe disease specific disease DMARDs and biologics are often used. Biologic treatments target specific molecules thought to be implicated in disease pathogenesis and include therapies which target the IL-6 and TNF-α pathways or the B cell marker CD20 (Emery et al. 2010;Taylor and Feldmann 2009). Furthermore, several of these therapies can be used in combination, allowing treatment to be tailored in response to patient disease activity and progression. Biological therapies have revolutionised the treatment of RA and has shown that if further therapeutic advancements are to be made, a greater understanding of the pathways contributing to disease is crucial.

1.1.2 Juvenile Idiopathic Arthritis

Juvenile idiopathic arthritis (JIA), formerly called juvenile rheumatoid arthritis (JRA) or juvenile chronic arthritis (JCA), is the most common arthritic disease of childhood (Ravelli and Martini 2007), affecting 1 in 1000 children in the United Kingdom (UK) (Symmons et al. 1996), with higher incidence reported in European

and North American populations (Adib et al. 2005). As JIA is a highly heterogeneous disease several classification criteria have been developed for phenotypic classification into more homogeneous subtypes. Currently 7 subtypes, defined by the International League of Associations for Rheumatology (ILAR) represent the most comprehensive and widely utilised guidelines (Petty et al. 2004). Although a minority of disease subtypes are considered well-defined, disparities exists throughout, with a single group who do not fulfil any other group called "undifferentiated arthritis." This is representative of the diverse pathology of JIA which can include several extra-articular manifestations including lymphadenopathy, hepatosplenomegaly and serositis. Like RA, JIA treatment has benefited significantly from the development of novel DMARDs and biologic therapies. Methotrexate is often considered the standard therapy for all JIA subtypes with particularly good responses seen in poly-articular disease subtypes (Cespedes-Cruz et al. 2008;Ruperto et al. 2004;Wallace et al. 2012) Furthermore response to TNF-α blockade through Etanercept and Adalimumab is variable amongst different subtypes (Giannini et al. 2009;Lovell et al. 2008;McErlane et al. 2013;Ruperto et al. 2010) This heterogeneity suggests that like RA a multi-therapy approach may be required for effective treatment of this disease.

1.1.3 Psoriatic Arthritis

Psoriatic arthritis (PsA) is a heterogeneous disease, encompassing several clinical entities but most prominently articular and dermatological manifestations (Kavanaugh and Ritchlin 2006). Although typically seronegative for auto-antibodies, PsA demonstrates significant clinical overlap with Psoriasis Vulgaris (PsV) and Ankylosing Spondylitis (AS). Furthermore 30-50% of PsV patients will develop arthritis compared to ~1% prevalence in the general population indicating common pathogenic pathways may exist between the diseases (Gladman et al. 2005). Additionally, articular disease presentations are often accompanied by cardiovascular disease risk and other systemic comorbidities (Han et al. 2006). It has also been shown that patients with PsA have significantly reduced quality of life compared to those with PsV alone (Rosen et al. 2012). Previously the heterogeneity of PsA has led to the generation of several contradictory classification criteria, resulting in restricted uniformity of disease

subsets (Dougados et al. 1991;Gladman et al. 1987;McGonagle et al. 1999;Moll and Wright 1973). In 2010, a series of classification criteria for PsA (CASPAR) guidelines have been established, providing a concise series of diagnostic criteria with a high disease specificity of 98.7% (Taylor et al. 2006). This has allowed for classification of disease into distinct clinical groups, increasing consistency of diagnosis and therefore the validity of research and reproducibility of clinical trials. PsA therapy often requires the use of one intervention which targets multiple aspects of disease or a combination of single target treatments which can effectively be used in combination. As with other types of IA, NSAIDs, DMARDs and biologic agents may be used to achieve symptomatic control of articular disease (Gossec et al. 2012;Smolen et al. 2014).

1.2 Pathogenesis of Inflammatory Arthritis

As with several autoimmune diseases RA, JIA and PsA share several aspects of their clinical presentation. The most prominent of these are articular manifestations, which are believed to be attributed to a series of inflammatory events resulting in joint specific damage and destruction. Examples of the joint space narrowing and destruction which occur as a consequence of this are shown in the hand x-rays in Figure 2.

Of these three types of IA, the pathogenesis of RA has been most well characterised, as it is the most prevalent and well-studied. Once mechanisms of pathogenesis are identified in RA, many are reproduced in JIA and PsA, indicating that a significant overlap exists between the diseases. Therefore, the majority of the mechanisms discussed in this section have been identified in RA but many can be applied to JIA and PsA inclusively. A summary of these features is given in Table 1.

Table 1 – Summary of similarities and differences between RA, JIA and PsA

	Clinical features						Immunological features				
	Articular	**Additional**	**Lab markers**	**Age of onset**	**Sex bias**	**Treatment**	**Role of MHC**	**Innate**	**Adaptive**	**Cytokines**	**Refs**
Similarities between RA, JIA and PsA	All Polyarticular, commonly affect multiple joints.	Some overlap e.g. PsA-JIA and PsA share features such as psoriasis. RF+ p-JIA and RA may both feature rheumatoid nodules.	Anaemia raised CRP and ESR often characteristics of inflammation across the 3 types of IA. RF and ACPA autoantibodies found in RA and RF+ p-JIA. Less frequently ANAs found across o-JIA and RA.	Some overlap between age on onset in PsA (30-55 years) and RA (40-50 years female/70-80 years male)	Female gender bias in RA, o-JIA and RF+ p-JIA. No gender bias in RF- p-JIA, PsA-JIA and PsA.	All types of IA treated with Glucocorticoids, NSAIDs, DMARDS e.g. methotrexate, sulfasalazine, Biologics e.g. Etanercept, Abatacept, Rituximab.	MHC Class I associated with RA, e-JIA and PsA although different antigens (RA = HLA-B, e-JIA = HLA-B27 and PsA = HLA-Cw6/HLA-B27) Class II HLA-DRB1] associated with RA and oJIA/pJIA.	Macrophages (particularly M1 type) involved in RA and o-JIA/p-JIA in producing inflammatory cytokines and inducible nitric oxide synthase. In s-JIA MAF key manifestation of disease. .	Plasma B cell involvement in RA and p-JIA as production of ACPA, ANA and RF autoantibodies. Primarily CD4+ Th1 T cell infiltrate in RA and o-JIA/p-JIA. In both Tregs are present but are limited in function/capable of pro-inflammatory activity. Th17 cells important in RA, JIA and PsA.	Produced by a variety of inflammatory cells. In some cases have been successfully targeted for therapy. Cytokines identified across all diseases include TNF-a, IL1, IL6, IL-12, and IL-23.	(Aletaha et al. 2010;Petty et al. 2004) I(Taylor et al. 2006), (Raychaudhuri et al. 2012),(Hinks et al. 2013) (Gladman et al. 2005)

	Clinical features						Immunological features				
	Articular	**Additional**	**Lab markers**	**Age of onset**	**Sex bias**	**Treatment**	**Role of MHC**	**Innate**	**Adaptive**	**Cytokines**	**Refs**
Differences between RA, JIA and PsA	Different number of joints and patterns observed. Differences in location, symmetry and severity of joint inflammation.	JIA subgroups are often distinguishable by extra-articular manifestations such as uveitis in o-JIA and serositis in s-JIA with limited overlap observed.	PsA and PsA-JIA usually seronegative for RF, ACPA and ANAs. HLA-B27 exclusive to e-JIA subtype not associated with other types of IA.	Diagnosis of JIA is relies on manifestations occurring under16 years although symptoms can continue into adulthood.	Male gender bias in e-JIA not found in others.	Treatment is dependent on disease and specific patients e.g. more likely to treat systemic JIA with Anakinra (anti-IL1R antagonist) and e-JIA with ustekinumab (IL-12 and IL-23 inhibitor).	Although same class of MHC, may be differences e.g. HLA-B in RA and HLA-C in PsA.	M2 macrophages present in PsA synovium but not characterised in RA and JIA.	No autoantibody production in PsA, indicating limited role for plasma B-cells. Although RA and JIA carry CD4+ signature, distinctive role for CD8+ T cells identified in PsA as enriched in joints compared to RA.	Although many similarities in cytokines present may be produced at different rates and have different consequences.	(Menon et al. 2014), (Seibl et al. 2003), (Wenink et al. 2012) (Trynka et al. 2013) (Brennan et al. 1989) (Feldmann 1996) (McInnes and Schett 2011)

Table 1 is a summary of the clinical and immunological features in common and different between RA, JIA and PsA. RA = Rheumatoid arthritis, JIA = Juvenile idiopathic arthritis, PsA = Psoriatic arthritis, RF = rheumatoid factor, ACPA = anti-citrillunated protein antibodies, ANA = anti-nuclear antibodies, p-JIA = polyarticular juvenile idiopathic arthritis, o-JIA = oligoarticular juvenile idiopathic arthritis, s-JIA = systemic juvenile idiopathic arthritis, e-JIA = enthesitis related juvenile idiopathic arthritis, PsA-JIA = psoriatic juvenile idiopathic arthritis, HLA = human leukocyte antigen, MHC = Major histocompatibility complex, NSAID = non-steroidal anti-inflammatory drugs, DMARDS = disease modifying anti-inflammatory drugs, TNF = tumour necrosis factor.

Overall, as a clinical syndrome RA is a chronic arthritis thought to be driven by a series of autoimmune inflammatory processes. Most characteristic of these is the establishment of chronic inflammation in the synovium, which normally occurs across multiple joint sites (polyarthritis) Figure 3. These synovial lesions are generated by influx and hyperplasia of a multitude of innate and adaptive immune cells into the synovium. This results in the dysregulation of the local joint structure and the promotion of tissue breakdown by local cells including synovial fibroblasts (Buckley et al. 2001;Buckley et al. 2004). Initially, single cell hypotheses were used to explain the contribution of each individual immune component to disease but more recently an integrative approach has been adopted, acknowledging that there are several interconnecting components contributing to the disease process.

Figure 2 - Hand x-rays of patients with RA, JIA and PsA

Figure 2 shows a hand x-ray (L-R) from an RA, JIA and PsA patient (http://images.rheumatology.org/). Joint space narrowing can be seen across all x-rays which is an overlapping feature of these diseases.

1.2.1 Cellular Infiltrate

A key feature of inflamed RA, JIA and PsA synovium is the influx of a variety of inflammatory cells from both the innate and adaptive immune systems. This is mediated by an increase in vascular angiogenesis initiated by the presence of inflammatory mediators and a reduction in lymphangiogenesis (Polzer et al. 2008;Szekanecz and Koch 2008). Establishment of a hypoxic pro-inflammatory

environment in the synovium is facilitated by several innate immune mechanisms including activation of M1 macrophages and dendritic cells (DCs) via pattern recognition receptors such as extracellular toll like receptors (TLRs) and intracellular nod like receptors (NLRs) (Huang and Pope 2009). The process is initiated by the presence of autologous and external antigens, which, to date, have not been identified. These antigen presenting cells (APC) cells are then responsible for the presentation of arthritis-associated antigens to T lymphocytes, activating the adaptive arm of the immune system (Ciechomska et al. 2014;Wilson et al. 2012).

The presence of adaptive immune architecture, such as ectopic germinal centres, indicates a strong effector lymphocyte presence in the inflamed joint (Humby et al. 2009). Both activated B cells (Doorenspleet et al. 2014;Edwards et al. 1999) and CD4/CD8 T cells (Murray et al. 1996;Van Boxel and Paget 1975) are characteristic of the inflamed synovium with the pro-inflammatory environment promoting distinct skewing towards a CD4+ Th1/Th17 phenotype with a particular role for Th17 cells identified across the diseases (Leipe et al. 2010;Omoyinmi et al. 2012). This has been particularly important in PsA, as recent evidence supports a crucial role for the pathway in disease (Kirkham et al. 2014). In addition it has also been shown CD8+IL17+ cells are also present in PsA, indicating that some differences in T lymphocyte profiles exist between the diseases (Menon et al. 2014) . T regulatory cells (Tregs) have also been shown to be present but may be functionally suppressed by the surrounding inflammatory milieu (Morgan et al. 2005). This has been particularly apparent in the JIA as a population of Treg cells have been identified in these joints which exhibit a dysregulated pro-inflammatory profile (Pesenacker et al. 2013), indicating that both activation of effector cell types and suppression of regulatory mechanisms may be contributing to the establishment of autoimmunity.

1.2.2 Autoantibodies

The presence of autoantibodies in RA indicates a central role for B-lymphocytes in disease pathogenesis. Additionally the presence of B cell survival factors such as a

proliferation inducing ligand (APRIL) and B lymphocyte stimulator (BLyS) have been shown to be associated with high autoantibody levels and have been identified as potential therapeutic targets for the treatment of RA (Daridon et al. 2009). Common autoantibodies found in RA include rheumatoid factor (RF), anti-citrullinated peptide antibodies (ACPA) and the most recently identified anti-carbamylated antibodies (anti-CarP) specific (Lee et al. 1992;Shi et al. 2011;Shi et al. 2013;van der Linden et al. 2009).). Rheumatoid factor is an IgG or IgM autoantibody raised against the fragment crystallisable portion (Fc portion) of immunoglobulin G (IgG) antibodies. Interactions between IgG and RF lead to formation of immune complexes, which are subsequently detected by surrounding immune surveillance leading to an inflammatory response. ACPA antibodies target citrulline epitopes which are generated post-translationally from arginine by peptidyl deaminase (PAD) enzymes. Finally, the presence of anti-CarP antibodies has recently been shown to be a predictor of joint destruction in ACPA- RA patients(Shi et al. 2013;Shi et al. 2014). These antibodies, which target homo-citrulline residues converted by carbamylation, have been shown to present in 38% of patients prior to their diagnosis of RA, indicating this may be a marker of the pre-articular non –clinical phase of RA;(Shi et al. 2014a). Very often seropositive RA is considered more progressive and poorer in outcome, with erosions and extra-articular rheumatoid nodules occurring earlier in disease (Mottonen et al. 1998). However, RF is found in a substantial minority of the healthy population, in other autoimmune diseases and transiently after infections suggesting that it may be a bystander product rather than disease causal. (Carson et al. 1978;Carson et al. 1987)

RF is used diagnostically to identify disease but the features mentioned above make it neither particularly sensitive nor specific to RA. ACPA are also used as clinical markers of disease because, although less common than RF, they are more specific (van der Linden et al. 2009). These autoantibodies are generated against specific citrullinated epitopes, several of which have been identified, but are detected with an anti-cyclic citrullinated peptide antibody assay designed to capture a wide range of ACPAs. Although a useful tool, absence of autoantibodies does not exclude RA as a diagnosis with approximately 20% of cases testing seronegative, even in hospital based series, and higher percentages in community

settings. These seronegative patients often have a slightly different clinical presentation and have a better outcome than the seropositive group.

The presence of autoantibodies has also been important in JIA diagnosis, with autoantibodies representing a defining factor in the classification into 7 disease subsets. As with RA, both RF and ACPA are known to be present in a proportion of disease cases and it is thought they are generated by identical mechanisms. Furthermore anti-nuclear antibodies (ANAs) can be found in JIA patients, with some data suggesting that ANA seropositive patients represent an independent JIA disease group, which is currently not defined by the ILAR guidelines (Ravelli et al. 2007).

Overall, PsA is considered a seronegative spondyloarthropathy therefore; seronegativity can be used as a clinical diagnostic to differentiate it from RA. Despite this, emerging evidence suggests that the presence of ACPA may be an indicator of disease severity in a subset of PsA patients, indicating that autoantibodies can contribute to disease (Perez-Alamino et al. 2014).

1.2.3 Cytokine Disequilibrium

The incorporation and activation of inflammatory cells results in the establishment of an inflammatory cytokine milieu and the up-regulation of matrix metalloproteinase (MMP) enzymes, promoting bone matrix degradation. Prominent inflammatory cytokines include the interleukin 1 (IL-1) family members IL-1 (Dayer 2003), interleukin 18 (IL-18;(Maeno et al. 2002;McInnes et al. 2005) and interleukin 33 (IL-33; (Xu et al. 2008) which can drive persistent inflammation through promotion of Th1 differentiation. Furthermore both interleukin 6 (IL-6(De et al. 1994;Hirano et al. 1988) and tumour necrosis factor alpha (TNF-α ;(Brennan et al. 1989) cytokines have key roles in promoting inflammation, further evident by their effectiveness as therapeutic targets in the clinic. This cytokine disequilibrium drives the promotion of joint destruction by site-specific cells such as synovial fibroblasts, osteoclasts and chondrocytes. Under these circumstances fibroblasts of the synovial membrane lose contact inhibition, developing a tumour like phenotype and promote the recruitment of inflammatory

leukocytes to the synovium (Buckley et al. 2004). The presence of TNF-α, IL-6 and IL-17 promotes the production of receptor activator of nuclear factor kappa-B ligand (RANKL), which induces osteoclast differentiation and subsequent cartilage and bone degradation characteristic of clinical disease (Gravallese et al. 2000). It has been shown in vitro that osteoclast inhibition can limit bone degradation but synovial inflammation still persists (Cohen et al. 2008). This chronic degradation process results in the exposure of neo-antigens which in turn can be presented to lymphocytes by synovial APCs, establishing a feedback loop of chronic inflammation. The mechanism described is specific for articular inflammation in disease but it is important to note that several immune mediated extra-articular pathologies are characteristic of inflammatory arthritides including serositis in JIA and psoriasis in PsA.

Figure 3 - Diagram of normal and disease affected joint (Strand et al. 2007).

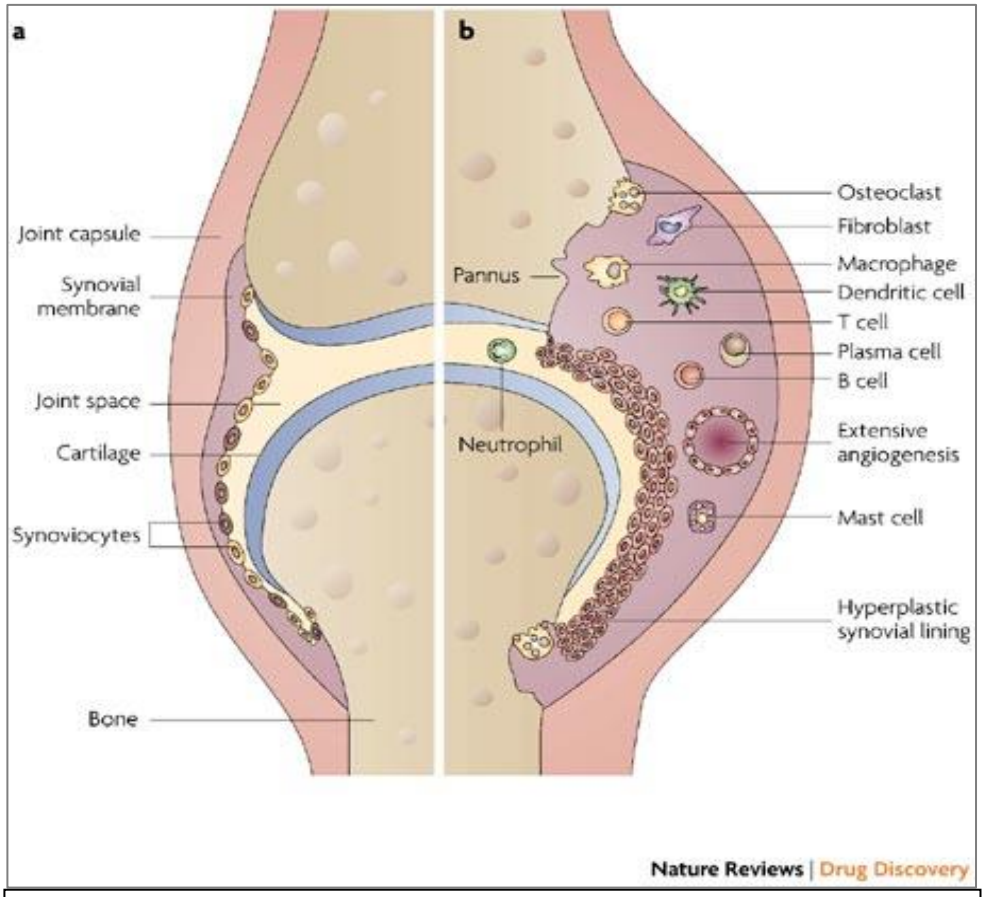


Figure 3 is a representation of the different cell types involved in the pathogenesis of RA at the synovial joint site. Taken from Strand et al 2007.

1.2.4. Animal models and the pathogenesis of IA

Although many of the findings described in section 1.2 were identified using human derived tissue, the availability of animal models has been crucial in the development of hypotheses about the induction of RA pathogenesis. The advantage of animal models is that mechanisms, such as precise response to stimuli and response to potential therapies, can be monitored in specific tissues such as joints or lymph nodes, which are often challenging from human subjects. Studies can also be performed in a regulated environment in a large number of subjects which increases the reliability of the results obtained.

Animal models are deliberately induced pathologies, which mimic the human disease of interest in a selected host animal, with the mouse or rat representing the most widely used species. The most commonly utilised mechanisms for generation of RA animal models are induced animal models such as collagen induced arthritis (CIA), zymosan induced and proteoglycan-induced arthritis (PGIA) or genetically altered spontaneous arthritis models such as TNF-α transgenic and K/B×N mouse model (Courtenay et al. 1980). Of these, the most widely studied are CIA and PGIA models. These involve the inoculation of genetically susceptible mice with type II collagen (CII) or in the case of PGIA, cartilage proteoglycan plus adjuvant. In both cases after 20-30 days, the inoculated mice develop polyarthritis, which can be used as a disease model organism for RA. These models resolve automatically after 30+ days but relapse can be induced, making them a good representation of the human pathology.

Although several similarities between murine models and human disease have been identified, there are still many differential features between the two. Both human and mouse pathologies show an MHC class II restricted T response, with specific T cell responses to citrullinated proteins such as fibrinogen and aggrecan (Cordova et al. 2013;Misjak et al. 2013). Additionally ACPA and anti-class II collagen autoantibodies, which are believed to be a key factor in the seropositive human disease, have been found in CIA mouse models (Goldschmidt et al. 1992;Kidd et al. 2008). Overall, in both mice and humans a RA a strong T cell influence has been identified as a key meditator of pathogenesis. In both species,

this appears to be driven by Th1, Th17 and Treg cells with the notable absence of Th2 helper cells (Wehrens et al. 2013).

Although the similarities are vast, several distinct differences in T cell cytokine expression, polarization and plasticity have been identified between human disease and mouse models. For example although Tregs are found in both human RA and mouse models, the suppression of Treg function observed in humans has not been identified in CIA or PGIA models (Wehrens et al. 2013). Furthermore, when stimulated with TNF-α, Treg populations were shown to have very different effects with Treg expansion observed in mice whilst Treg reduction was observed in humans (Ehrenstein et al. 2004;Valencia et al. 2006). This is further supported by the observation that therapies blocking T cell responses in the mouse are enough to completely resolve inflammation but is ineffective in humans. (Goldschmidt et al. 1992; (Keystone 2003) This has also been observed in differential responses to the B cell inhibitor Rituximab, indicating therapeutic outcomes in mice cannot always be directly translated into humans (Hamel et al. 2011) Despite this, mouse model remains a strong tool for understanding RA pathogenic mechanisms and are a good indication of what is driving human disease.

1.3 Aetiology of Inflammatory Arthritis

Currently it is accepted that induction of IA occurs due to a combination of several risk factors including genetic predisposition, environmental exposure and the presence of random stochastic events such as infection. These factors may not be apparent singularly but in combination may overcome a putative threshold triggering disease induction Figure 4. To date the exact timing and level of this threshold has not been determined but it is believed that a pre-articular phase where the disease mechanism is established occurs well before the clinical manifestations of disease.

Like many immune mediated diseases, RA has a significant genetic component. This was first identified using twin studies, where disease concordance between monozygotic twins (15%) was much higher compared to dizygotic twins (2.3%) (Silman et al. 1993). This has been complemented by several studies estimating

that genetic factors alone contribute ~50% to overall disease risk, which is attained by comparing disease prevalence in multiple populations (MacGregor et al. 2000;van der et al. 2009) This heritability can be described by a moderate sibling recurrence risk ratio (λ_s) of 2-17. λs is a statistical parameter describing the "prevalence of a disease amongst siblings compared to the general population," and hypothetically can be used to estimate the contribution of genetic factors to disease (Seldin et al. 1999). In addition, the contribution of environmental factors both preceding and throughout disease is thought to have a significant impact on the induction and subsequent course of disease.

Several environmental factors have been shown to be associated with induction and progression of RA including socioeconomic status, obesity, smoking and alcohol consumption (Kallberg et al. 2009). Additionally hormones are believed to also have an impact, as RA is more prevalent in females compared to males and oral contraceptives have been associated with the presence of RF antibodies which are present in 70% of RA patients (Cutolo and Straub 2008). Of the environmental factors described, cigarette smoking has been shown to be the strongest risk factor identified as this finding has been replicated in several independent cohorts. (Karlson et al. 1999;Symmons et al. 1997) Although it has been shown that the influence of smoking is strongest in ACPA+ disease, the specific mechanisms underlying the process of local citrullination have yet to be characterized (Klareskog et al. 2006). One potential hypothesis is that the presence of microbes such as Porphyromonas gingivalis (P.gingivalis) in the gums produce enzymes which are responsible for the citrullination of epitopes and triggers the production of ACPA antibodies, which drive disease (Quirke et al. 2014;Scher et al. 2014).

JIA and PsA exhibit a similar risk profile to RA through integration of genetic and environmental risk factors. With a λ_s value of 15, genetics are thought to be highly important for JIA with an estimated 17% of risk coming from risk HLA-DR gene alleles (Glass and Giannini 1999;Prahalad 2004). Furthermore, several JIA environmental risks have been investigated including maternal smoking and stress but no validated results have been obtained (Herrmann et al. 2000;Jaakkola and Gissler 2005) . Interestingly, PsA has the highest λ_s value of 30.8 but identification of environmental impact has been limited so far (Chandran et al. 2009) with only

obesity and smoking shown to be associated with disease (Love et al. 2012) .

Figure 4 – Induction of autoimmunity in rheumatic diseases (Deane and El-Gabalawy 2014).

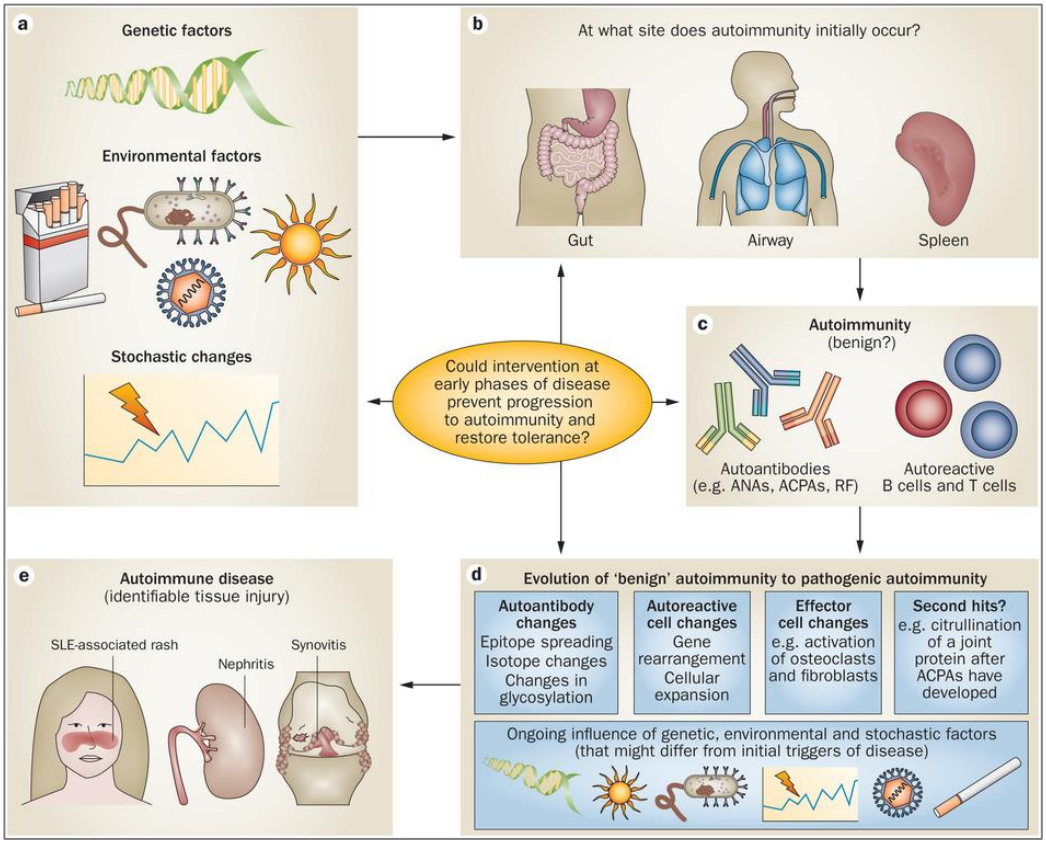


Figure 4 shows a hypothesis for the induction of rheumatic diseases such as RA and SLE. As all 3 types are IA discussed in this section are complex diseases, they are believed to be a contribution of genetic, environmental and stochastic factors. Taken from Deane and El-Gabalawy 2014.

Overall, it is apparent that genetic factors contribute a substantial proportion of disease aetiology for IA, reinforcing the importance of genetic studies in understanding disease causation. Currently several methods are used to identify disease associated genes with consistent technological advancements and greater understanding of the genome driving evolution of this field. The associations can then be further interrogated to identify the functional consequence. This process is summarized in Figure 5 is described in detail in the following sections.

Figure 5 - Summary of stages involved in identifying and characterising a disease-associated locus.

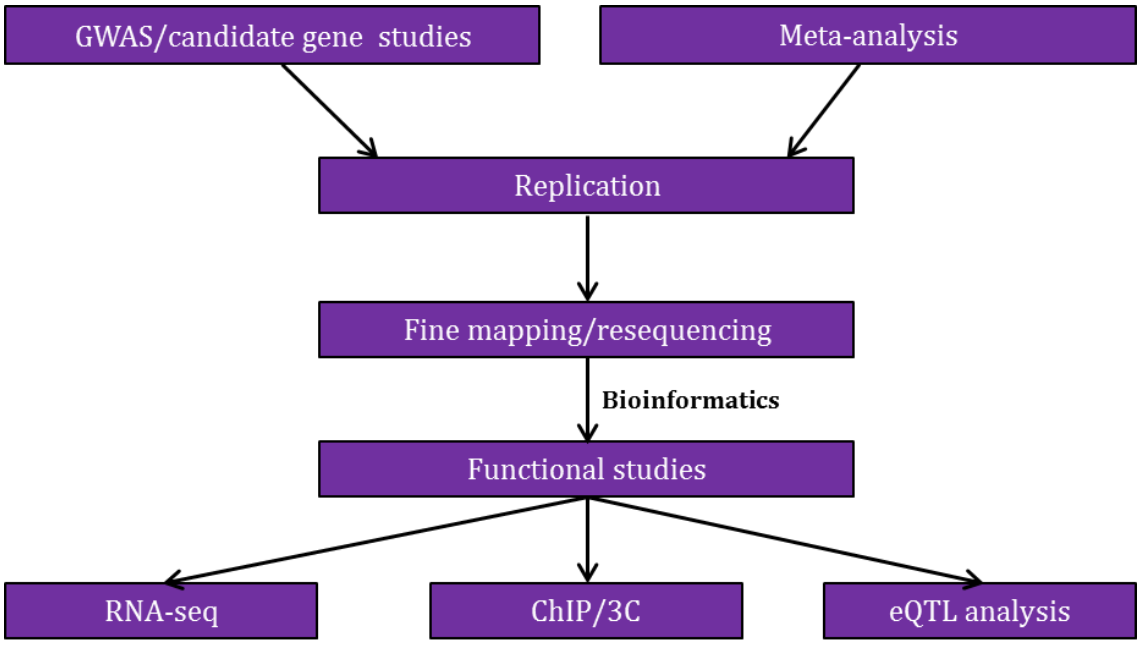


Figure 5 shows a sequential series of stages for following up disease associations from GWA studies through to functional analysis.

1.4 Identification of Disease Susceptibility Genes

Early genetic analysis of disease traits involved the investigation of co-occurrence of genetic markers and disease within affected families, known as linkage analysis. Such studies were limited by small sample sizes and therefore issues of power. Although linkage within the HLA-region was consistently observed, very few other regions showed reproducible findings in independent cohorts. Although a successful tool for identifying genes responsible for monogenic traits such as Huntington's disease and some forms of breast cancer, this did not appear to be as successful in identifying genes associated with complex diseases such as RA (Holloway et al. 1998).

Increasingly, genetic association studies have been adopted as the method of choice for the investigation of complex diseases such as RA, JIA and PsA. Association studies can be performed using a candidate gene case-control model but these have been largely superseded by genome wide association studies (GWA studies), which represent the most widely used technique to date.

1.4.1 Case Control Studies

Case control association studies are underpinned by the hypothesis that differences in genetic variation can be detected between disease cases and healthy controls, which are responsible for contributing to disease. This involves genotyping genetic markers in groups of unrelated cases and controls, with and without disease, respectively.

In the past this was performed using DNA microsatellites but currently, the most commonly used genetic marker is the single nucleotide polymorphism (SNP) Figure 6. These genetic variations occur approximately every 200 base pairs throughout the genome and can be genotyped using a variety of platforms. The frequency and ease of detection makes SNPs ideal markers for investigating genetic variation in populations.

Figure 6 – SNP polymorphism

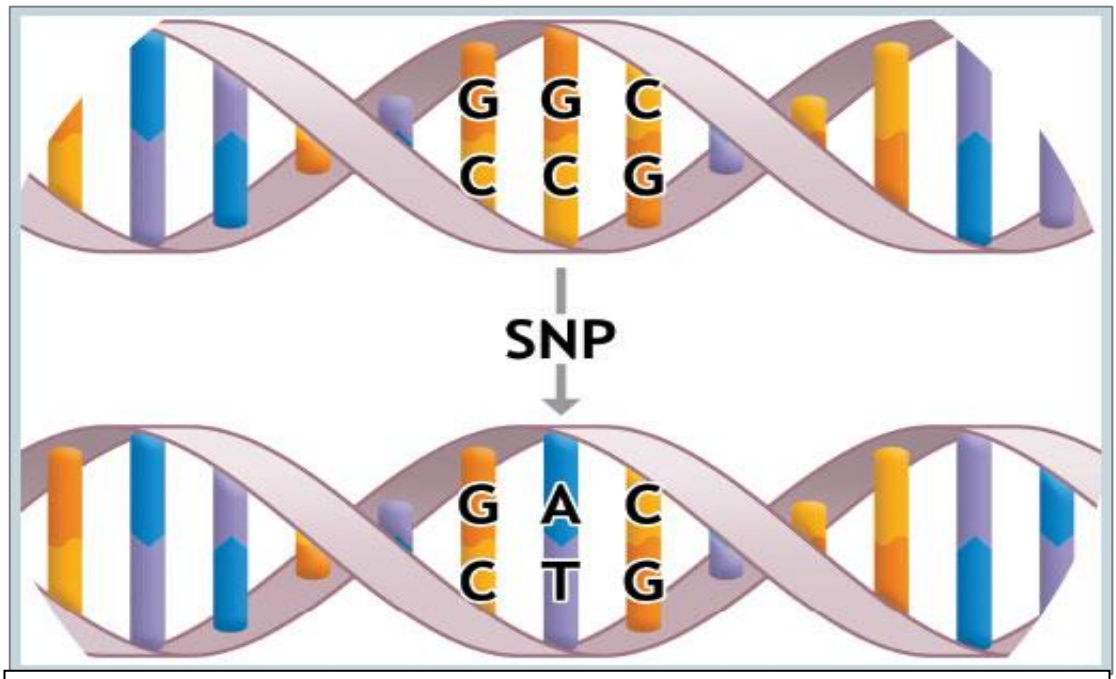


Figure 6 shows a single nucleotide polymorphism change in double stranded DNA. Taken from the integrated biobank of Luxembourg. (www.iblu.lu)

SNPs commonly have 2 alleles, described as major and minor alleles according to their frequency in a population, but at some loci more alleles may be present (Crawford and Nickerson 2005) Figure 6. Case control association analyses use statistical methods to assess the differences in allele/genotype frequency between DNA samples from disease cases and healthy controls. The identification of SNPs whose allele frequencies occur at a statistically different frequency in cases compared to controls is considered a significant genetic association and therefore represents a potential candidate variant for disease causation.

Case control studies can be performed using a candidate gene approach or by a hypothesis free genome wide association (GWA) study. Candidate gene studies involve the selection of a region for association testing based on previous knowledge of disease biology. A good candidate may be a region or variant which has been previously associated with a related disease, a component of a known disease pathway or a candidate identified by previous linkage analysis which requires further investigation. This strategy often involves genotyping of between

one and several hundred SNPs within these selected regions and has been a successful technique for the identification of several SNP associations with JIA and PsA, diseases which did not harness the power of GWA studies until recently.

In contrast, GWA studies involve the genotyping of a much larger number of SNPs across the genome and do not focus on particular genetic regions or hypotheses. This approach is underpinned by the phenomenon of linkage disequilibrium (LD), which allows variation which is not physically genotyped to be captured using a tag-SNP approach, therefore increasing the chance that disease associations will be identified.

1.4.2 Linkage Disequilibrium

LD is a non-random observation that 2 or more alleles are inherited together more frequently than expected during random formation of haplotypes. Haplotypes are a set of closely linked alleles which are inherited together as a complete unit due to lack of recombination events between them. This phenomenon extends across the genome, with blocks of LD ranging from few base pairs to larger sections of chromosomes. This means that across the genome sections of variable LD can be present with regions of high LD considered "LD blocks," with the presence of several correlated alleles. As LD is non-quantitative, there is no natural scale for measuring it; therefore, two ways of measuring this are used. The D' value reflects the recombination events which have occurred between two markers but is often affected by distance and frequency of SNPs, therefore the most informative measurement is the correlation, r^2. The value is the correlation coefficient between two SNP markers squared and takes into account the power which is required to detect LD between markers. r^2 is scored from 0-1, with 1 representing complete LD and therefore complete correlation between two or more SNP markers (Balding 2006;Palmer and Cardon 2005). The value is a more accurate representation of LD as it is not based purely on distance between SNPs and takes into account blocks of LD which stretch across larger distances.

The presence of LD is considered both a great advantage and disadvantage for association studies. LD across a region means that a single tag SNP can be used as a

proxy for all SNPs which lie in high LD with it (often r^2>0.8). It allows information about allele frequencies to be gained for multiple SNPs without genotyping each individual SNP, therefore reducing the volume of genotyping required. In some populations, a selection of as few as 500,000 SNPs are capable of tagging over 10 million variants and this has been the cornerstone of large GWA studies which have identified over 100 associations with RA alone .

Conversely, the presence of LD may potentially complicate association results. As a single SNP may be responsible for several association signals being picked up with SNPs which are in high LD, it makes it challenging to identify which SNP represents the true causal variant. A causal variant represents the SNP in a disease-associated region which results in a functional change or regulation of a gene and which confers disease risk as a result. The lack of identification of true causal variants been a limitation of the GWA studies and has resulted in the utilisation of strategies designed to localise association signals such as those described in section 1.5 (Weiss and Clark 2002).

1.4.3 Genome Wide Association Studies

The field of complex disease genetics have been revolutionised by the development of GWA studies (Balding 2006;Klein et al. 2005) which have made association testing across the whole genome possible. Crucially GWA studies require no prior hypothesis, candidate gene selection or previous knowledge regarding a genetic region .This makes it a good starting point for diseases in which genetic associations have not been detected using other techniques. Using similar statistical techniques to candidate gene studies, GWA studies are used to investigate differences in genotype or allele frequencies between cases and controls but they are performed on a genome wide scale. This is achieved by genotyping a large number of SNPs in a substantial number of unrelated cases and controls using high throughput SNP genotyping technology.

Large sample sizes are used in GWA studies as they increase the power to detect minor differences in allele frequencies and, therefore, identify smaller genetic effects (Evans and Purcell 2012;Sham and Purcell 2014). The technique is aided by

a greater availability of large case/control cohorts as the result of international consortia such as the Wellcome Trust case control consortium (WTCCC), which was the source of the first comprehensive GWA study published in 2007 (Wellcome Trust Case Control Consortium 2007). In addition, the mapping of a larger number of SNPs across the genome has allowed for a greater understanding of LD and provided the potential to tag as many SNPs as possible using high throughput SNP genotyping platforms. This has been greatly aided by the establishment of the International Hap Map (http://hapmap.ncbi.nlm.nih.gov/) (The International HapMap Consortium 2003) and 1000 genomes projects (http://www.1000genomes.org/) (Abecasis et al. 2012;Durbin et al. 2010). These projects have involved the mapping and cataloguing of SNPs across the genome in several different populations. Whilst the Hapmap project was designed to map common variation and LD across the genome, the 1000 genomes project used next generation sequencing to identify low frequency SNPs present in only a small number of individuals. Combined, these datasets provide a high-density map of common and low frequency variation in the human genome therefore increasing the chance of identifying genetic associations. To date thousands of genetic associations have been identified using GWA studies, and the findings identified up to 2013 are shown in Figure 7.

Figure 7 – Associations identified by GWA studies

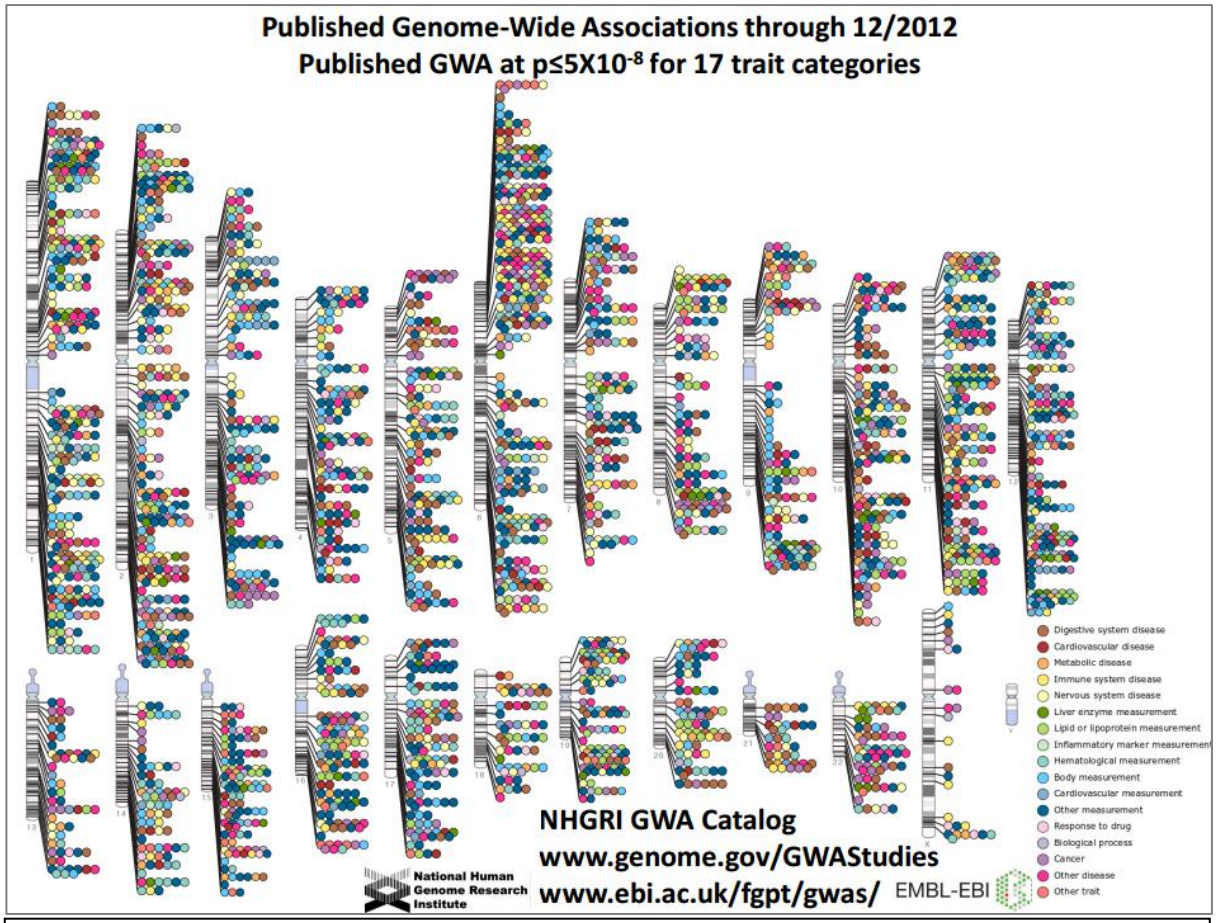


Figure 7 shows the variants associated with complex diseases from the genome GWAS catalogue. Reproduced from the NIH GWA study catalogue. (www.genome.gov/gwastudies/)

Although a great tool for discovering genetic associations across the genome, GWA studies do have their limitations. These include restrictions on throughput of SNP genotyping platforms, the availability of case and control samples to genotype and the risk of identifying spurious associations because of multiple testing. To resolve these, genotyping technologies are continuously evolving to genotype larger number of SNPs in larger sample sizes whilst keeping costs affordable. Illumina have currently developed an Infinium HD assay which can analyse up to 4.5 million SNPs per sample with the option of adding custom content to meet the specific needs of the user. In addition the establishment of an increasing number of international collaborations has allowed for increased sharing of case and control data, maintaining the samples sizes required for identification of novel associations. This is also supplemented by utilisation of meta-analysis techniques

to achieve the largest sample sizes and therefore the greatest power.

As GWA studies involve the testing of a large number of SNPs in an extensive number of samples, comparing allele frequencies at these levels requires a large number of statistical tests to be performed. As the number of statistical tests in an analyses increases, the chances of type I errors (false positive) arising as a consequence of multiple testing problems are increased. To minimise the risk of reporting spurious associations, false discovery rate (FDR) is calculated or Bonferroni corrected p-values are often used, which adjust for the number of statistical tests that have been performed. As genetic studies in particular involve many simultaneous statistical tests, an association is usually only considered significant if it passes the genome wide association threshold, which is currently $p < 5x10^{-8}$. This is particularly important in large well-powered studies such as meta-analyses of individual datasets, which include a particularly large number of statistical tests. The threshold was calculated on the basis that there are estimated to be one million independent SNPs in the genome and application of a Bonferroni correction results in the p-value used for claims of confirmed association (0.05/1,000,000).

1.4.4 Meta-analysis

Subsequent meta-analyses of GWA studies have been an essential technique in identification of additional disease susceptibility loci. Meta-analysis is the combination of results from multiple studies which have addressed similar research questions. In genetic association studies, this is often the combination of case control datasets assessing genetic components of the same disease. By increasing the number of case/control samples included in the statistical analysis, the power of studies is enhanced to detect smaller genetic effects. This is crucial in identifying loci which may be masked in smaller studies but may have a significant contribution to a disease trait. Several novel susceptibility loci have been identified from meta-analysis (Raychaudhuri et al. 2008;Stahl et al. 2010) including a recent example in which two novel RA susceptibility loci were characterized by combining results from large GWA study dataset with a smaller replication study to reach genome wide significance (McAllister et al. 2013).

Traditionally meta-analyses have involved combining datasets from GWA studies conducted on samples with identical ethnicity to minimise population stratification. Recently methods to combine GWA study datasets from different populations have been developed, allowing the combination of even more datasets and therefore a large increase in the power to detect genetic effects (Morris 2011). One particular success has been the identification of 40 novel RA loci in a meta-analysis combining GWA study datasets from Caucasian and Asian populations (Okada et al. 2014b) . This is particularly interesting as analyses were previously restricted to samples from the same population, to avoid detecting spurious associations as a consequence of population stratification.

1.4.5 Population stratification

Population stratification describes the differences in allele frequencies which are detected in genetic studies as a consequence of population structure and not as a result of allele differences between cases and controls. As genetic association studies are underpinned by the hypothesis that differences in allele/genotype frequencies between cases and controls are responsible for disease susceptibility, it is important that this is avoided to prevent identification of spurious false associations.

As differences in allele frequencies across different populations are determined by unavoidable evolutionary change, genetic studies in the past were restricted to analysis between cases and controls from the same population, limiting power. As power to detect genetic effects is a crucial factor in genetic studies, it is important that sample sizes are maximised in order to increase power. One way is to include samples in analyses which are from different populations but are believed to be of similar ancestry. Although it is expected that underlying allele frequencies will be similar, it is important to assess the study population for any underlying substructure, which has been shown to be an issue for genetic studies in the past (Campbell et al. 2005).

Several ways have been used previously to identify and account for substructure

including the genomic control (GC) and structured association approach, but these are subject to limitations. Currently the most widely used method for identifying population stratification is using Principal components analysis (PCA) (Price et al. 2006;Zheng et al. 2005) . PCA involves the identification of principal components which account for the largest differences between allele frequencies in a sample set. These principal components can then be plotted against each other on a PCA plot to provide a visual representation of how similar samples are as shown in Figure 8.

A common method used is to include samples of different ancestries from the Hapmap project in the PCA analysis, to see which population the tested samples share the greatest similarity with. The technique can then be used to identify and exclude population outliers who may skew analysis; furthermore, the principal components can be used as co-variates in analysis to adjust for the underlying variation in a sample set. Further information about this technique can be found in section 2.1.4.

Figure 8– PCA analysis of Hapmap populations (Heath et al. 2008)

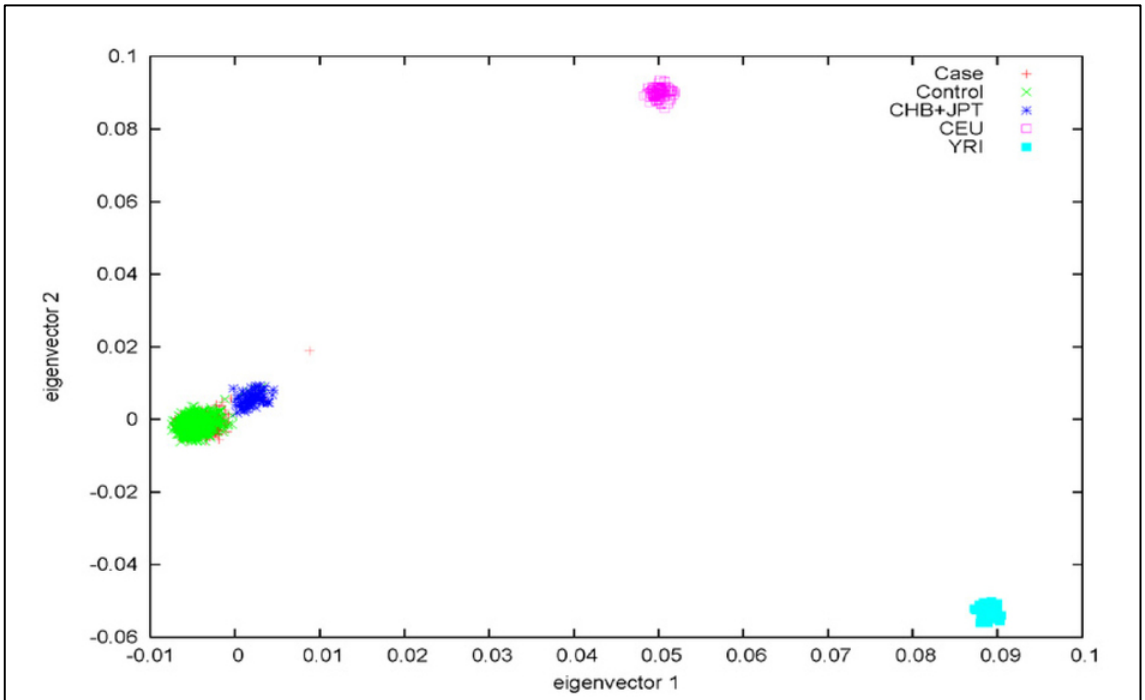


Figure 8 shows a PCA plot showing the clustering of different Hapmap and European populations. The x-axis represents the first principal component whilst the y-axis represents the second principal component generated in the analysis. Each dot represents a sample and the colour represents the population to which the sample belongs. Taken from Hou et al. Nature genetics 2014 (Hou et al. 2014).

1.4.6 Common Disease Common Variant Hypothesis

The strategies used in the search for genetic risk factors make assumptions based on the common disease common variant hypothesis (CD/CV), which was proposed to explain the contribution of common alleles to disease susceptibility (Reich and Lander 2001). The hypothesis proposed that genetic components of common complex diseases are of common frequency and distribution across a population. If this observation is true, then it is expected that GWA studies would represent the ideal method of identifying genetic susceptibility SNPs. This observation has driven the design of many GWA studies which often use a tag-SNP approach to capture as much common variation across the genome as possible using high throughput SNP genotyping arrays (Wellcome Trust Case Control Consortium

2007). Common variation is usually considered to be any SNP which occurs at a frequency of greater than 5% and has moulded the way that SNP genotyping technologies have evolved over the past decade. In a number of diseases a large proportion of common SNPs have been genotyped in large sample cohorts, therefore it is expected that complete heritability of some diseases should have been identified already. Although a large number of disease associations have been identified, no complex disease susceptibility has been completely explained by common variation alone which indicates there are additional factors contributing to disease susceptibility.

The CD/CV hypothesis does not take into account the contribution of low frequency variants, present in less than 5% of the population, to disease susceptibility. To address this many studies have moved towards searching for rarer variation throughout the genome. This has been reflected in novel genotyping technologies with Illumina developing both the Immunochip and Exome arrays, which have both been designed to capture low frequency variation across the genome. Although advantageous, the low frequency of these variants makes robust genotyping calling challenging using standard methods (Nievergelt et al. 2014). Furthermore, the power required to detect association with low frequency variants is much higher than common variants, therefore much larger sample sizes are required. Therefore, re-sequencing and fine mapping in a moderate number of cases and controls remain the most utilised methods for detecting low frequency variation. These are described in more detail in section 1.5. As the majority of low frequency variation has not been characterised it is thought that this variation may be one of the major factors contributing to missing heritability of RA (Bodmer and Bonilla 2008;Reich and Lander 2001).

1.4.7 Missing Heritability

Although many valuable associations with common diseases have been identified to date, it is now becoming clear that these associations account for only a fraction of the total genetic contribution to disease. It is therefore important to identify the remaining missing heritability to gain a complete understanding of the genetic component of disease (Manolio et al. 2009). Several factors could account for this missing heritability including multiple genetic effects within a region, copy number

variations (CNVs) and the presence of low frequency variants. Furthermore, gene-gene interactions and non-sequence epigenetic effects could also be contributing to disease susceptibility and has not been investigated by current studies.

It has been shown in RA that multiple independent effects can occur within a genetic locus, with carriage of multiple variants conferring higher disease risk than any individual single SNP alone (Orozco et al. 2009). This may also be true for the carriage of combinations of susceptibility loci and indicates that analysing data on an individual SNP basis may not generate an accurate picture of disease risk (McClure et al. 2009). A recent study has also shown that including all SNPs from regions identified by GWA studies provides a greater estimate of the heritability at each locus compared to using GWA study index SNPs alone (Gusev et al. 2013). Additionally the developments of polygenic risk scores, which combine multiple genetic markers to generate a risk score for predicting disease allows inclusion of markers which have not reached significance in studies, as hypothetically they may confer an effect when combined with other markers. The first successful application of the technique was performed in schizophrenia but has been shown to be applicable to RA risk and ACPA antibody levels, which are key pathogenic mediators of RA (Cui et al. 2014;Hamshere et al. 2011a;Hamshere et al. 2011b;Stahl et al. 2012) .

It is also thought that a proportion of missing heritability may lie in low frequency variation that has not been captured by previous common variant studies. Low frequency variation represents the proportion of polymorphisms with a minor allele frequency (MAF) of less than 5% in populations. Compared to common variants, there are many more low frequency variants across the genome which may confer disease susceptibility. Capturing low frequency variation is much more challenging than common variation as it requires a much larger sample sizes, which are unattainable for some diseases. Additionally, technologies for detecting low frequency variations are in their infancy. Low frequency variation was not captured by the majority of SNP genotyping arrays until recently, with the advent of the Illumina exome chip. Additionally, although DNA resequencing represents the ideal technique for identifying all variation in a sample, it is both costly and notoriously low throughput, making it unsuitable for large sample sets required to

identify low frequency variation. Once data is generated for low frequency variation, it is also challenging to analyse as algorithms used to call SNP genotypes are often incapable of calling low frequency variants. Furthermore, in order to efficiently analyse data from these studies, methods which collapse rare variants into a single effect, such as burden testing, are used. These methods involve assessing the presence and absence of rare variants per individual sample. Some methods can also take into account the functional prediction of rare variants which can then be prioritized as the most likely causal variants (Liu et al. 2014). A recent study utilised these methods to investigate low frequency variants in 25 RA susceptibility loci and found an accumulation of non-synonymous low frequency coding variants exclusive to RA cases in two loci (Diogo et al. 2013).

Due to the associations observed in other common diseases such as psoriasis, it was thought that CNVs could potentially account for some of the missing heritability of RA (Zhang et al. 2009). Although many genotyping arrays include markers for CNVs, it is often challenging to analyse this data as, like microarray data for transcriptomic analysis, it requires both normalisation and specific association testing to assess whether gain or loss of DNA nucleotides has occurred. In 2010, a large study was set up to analyse CNVs in 8 common diseases including RA. Although several CNVs were shown to be present at much higher frequency in cases compared to controls, these were all within the HLA region and had previously been identified by SNP association studies (Craddock et al. 2010). Although a contribution of CNVs to missing heritability cannot be excluded, it is unlikely as this was a large and well designed to assess the contribution of CNVs to disease.

Several other factors could potentially account for the unexplained heritability of RA which does not lie in the DNA sequence itself. These include gene-gene, gene-environment (epistatic) interactions and epigenetic effects as a consequence of post translational modifications such as methylation. It has been shown that interactions occur between the HLA shared epitope and the *PTPN22* genetic regions which contribute to RA susceptibility. Furthermore, evidence of interactions between environmental cigarette smoke and the shared epitope have also been shown (Kallberg et al. 2007;Morgan et al. 2009) . Another mechanism

which could be affected by the environment is the presence of epigenetic post translational modifications of DNA such as DNA methylation. The presence of these modifications has become a recent topic of interest in relation to disease susceptibility as the presence of these epigenetic factors can contribute to gene expression changes and subsequent alterations to the immune response (Glant et al. 2014) . A recent study in RA has shown differential methylation in the HLA region to be a mediator of RA disease risk (Liu et al. 2013). This has also been shown in synovial fibroblasts, where hypomethylation, which is believed to regulate gene expression was identified in the *CXCL12* promoter (Karouzakis et al. 2011). Although these studies are performed in limited sample sizes and are restricted by the fact that methylation is variable between cell types and responsive to environmental stimuli, they show that post-translational, modifications could potentially be responsible for a proportion of the missing heritability of RA.

Overall, it is likely to be a combination of all the factors mentioned which is responsible for the missing heritability in RA. As technologies and analysis pipelines evolve it is expected that how much each contributes will be identified. Additionally due to LD across the genome, the variants which have been identified as being associated with disease may not always represent the true causal variant, therefore a crucial aim in the future is to utilise emerging techniques to identify the causal gene and causal variant in disease-associated loci.

1.5 Identification of a Causal Variant

Once a statistically significant association is detected, it is desirable first that the association is replicated in an independent population, even if it has reached genome wide significance in the original study. Secondly, as the outcome of an association study is often the identification of a tag-SNP which captures several SNPs in LD, the identification of a causal variant(s) is often challenging. Additionally there may be other SNPs which are in LD with the associated SNP, which have not been captured by previous genotyping projects. Therefore post replication, it is essential that investigation into the genetic architecture of the

associated locus is performed using techniques such as imputation, resequencing and fine mapping.

1.5.1 Imputation

Imputation involves the analysis of SNPs in a region which have not been directly genotyped in previous analyses. This is achieved by using LD patterns in a region to predict the genotype of an individual at a particular SNP, based on their genotype at other SNPs in high LD. This can be performed either regionally or at a genome wide level. The technique has greatly advanced in recent years due to the Hapmap, 1000 genomes and the development of more accurate imputation software. The Hapmap and 1000 genomes projects represent a freely accessible database of genetic variation based on genotyping and sequencing data from a large number of individuals(Abecasis et al. 2012;The International HapMap Consortium 2003) . The great advantage of imputation is that it does not require any additional physical genotyping, although for reliable imputation to be performed, regions are required to be genotyped fairly densely (Huang et al. 2009).

1.5.2 DNA Resequencing

To gain the best chance of identifying a causal variant, it is essential that as many variants within an associated region are captured as possible. Resequencing involves the genotyping of complete sequences of genome using high throughput sequencing platforms. In the past high costs resulted in the testing being performed on a candidate gene basis using a small number of individuals. Recent advances in next generation sequencing technologies such as the development of the Illumina sequence by synthesis (SBS) technology have significantly reduced the cost of the method. This has increased accessibility to both capture and whole genome sequencing methodologies, allowing sequencing data to be generated on a large number of samples. Comparison of sequence data between cases and controls has been a successful technique for a number of disease studies and has

resulted in identification of highly penetrant variants such as the *PLB1* locus which contains a number of variants which are potentially associated with RA risk (Okada et al. 2014a) . Although promising, further studies will indicate how much value this technique holds for the identification of RA susceptibility loci.

1.5.3 Fine Mapping

Identification of a causal variant following an association study can be achieved by performing fine mapping analysis of a confirmed susceptibility region. Fine mapping uses high throughput genotyping technology to test a high density of SNPs within a region of interest. This involves the genotyping of an extensive panel of SNPs in an extensive set of cases and controls. A case control association study is then performed providing a detailed analysis of the genetic region. The advantage of fine mapping is that it provides additional information about a region and refines the peak of association. If this is performed in an independent cohort, this gives the opportunity to replicate existing associations and increases the chance of identifying a true causal variant which can then be carried forward for potential functional validation (section 1.6). Furthermore, it can be supplemented using statistical techniques such as conditional logistic regression, which can be used to determine whether independent effects exist within a genetic locus. For example, 3 independent effects in the RA associated 6q23 *TNFAIP3* genetic region were identified following fine mapping and conditional logistic regression analysis (Orozco et al. 2009).

Recently an array was generated by Illumina and a consortium of investigators studying immune mediated diseases to fine map genetic regions which have previously been associated with immune mediated disease. This array provides dense SNP coverage of a large number of regions and is ideal for identifying both causal variants and novel loci. As the array was used in the current project, it is discussed in more detail in the following sections.

1.6 From Genotype to Phenotype

For many of the susceptibility loci identified for complex diseases, the causal variant within an associated region has not yet been identified but studies are working towards this goal. In addition, as more techniques are developed there is a series of steps emerging to follow up these loci. Due to LD, fine mapping studies may identify a number of variants with equal genetic evidence for association but from which the causal variant cannot be distinguished. Functional studies are, therefore, critical in order to identify the likely causal variant(s) and to explore their biological role in disease causation (section 1.6). Functional studies can be performed using bioinformatics data mining and then performing laboratory techniques to confirm these findings and uncover the biological consequence of disease associated variants.

1.6.1 Bioinformatics

Prior to the initiation of costly laboratory studies, bioinformatics data is often extracted to identify any existing functional information about a variant or region of interest. This is often performed on the GWA study lead SNP and all proxies (r^2>0.9) to make sure that all SNPs correlated strongly by LD are included. A number of tools can be used to mine bioinformatics databases including ASSIMILATOR http://assimilator.mhs.manchester.ac.uk/cgi-bin/assimilator.pl; (Martin et al. 2011) and UCSC genome browser (http://genome.ucsc.edu/; (Karolchik et al. 2004;Karolchik et al. 2014), which are web interfaces that source data from the Encyclopaedia of DNA elements (ENCODE) project (Bernstein et al. 2012). The ENCODE project is a collaborative project with the role of bringing together laboratory data to identify functional elements across the genome. This includes information on evolutionary selection, transcription factor TF binding motifs, RNA/chromatin regulation, enhancer/promoter characterisation and epigenetic regulation. The data is generated using a wide variety of laboratory techniques including whole transcriptome expression, exome arrays, and chromosome confirmation capture (3C/5C), ChIP-seq, DNase-seq and RNA-seq in

an extensive number of tissues. When this data is utilised it can provide a comprehensive functional prediction which can be seen in the large number of studies which have accessed the data.

Although this wealth of information is valuable in predicting the function of associated variants, it is always desirable that a functional prediction is replicated using laboratory techniques. Some of the most commonly used techniques are described in section 1.6.2-1.6.3.

1.6.2 Expression and eQTL analysis

One potential function of a disease susceptibility variant may be to regulate the expression of gene transcripts. This is called an expression quantitative trait locus (eQTL) and can occur between closely positioned variants/genes (cis-eQTL) or at a longer range across the genome (trans-eQTL). To identify eQTLs gene expression levels are quantified and correlated with the carriage of a disease variant. Initially, due to ease of sample collection, the majority of eQTL studies were performed in whole blood. Recently it has been shown that many changes in gene expression occur at a cell specific level therefore investigation using a homogenous tissue of interest is desirable (Fairfax et al. 2012;Westra et al. 2013). Several public databases detailing previously identified eQTLs in a number of tissues are available to access online. These include the gene expression variation database (Genevar; http://www.sanger.ac.uk/resources/software/genevar/; (Yang et al. 2010) which contains data for cis-eQTLs exclusively and the SNP and CNV annotation database (SCAN; http://www.scandb.org/newinterface/about.html; (Gamazon et al. 2013)) which contains date for both cis and trans-eQTLs. Both databases include data for a number of tissues including adipose, skin and lymphoblastoid cell lines allowing identification of eQTLs in specific tissues.

Further advances in array technology and the quality of gene expression data have allowed a large number of independent eQTL studies to take place. Whole transcriptome arrays such as the humanht-12 v4 expression beadchip allow the generation of expression data for up to 47,000 transcripts, which can then be correlated with existing genotyping data to identify eQTLs in a specific tissue of

interest. Combined with advances in cell isolation technologies this can be used to generate large volumes of data in specific cell populations not included in the existing databases.

1.6.3 DNA interactions and chromatin structure

When a variant lies in a region which interacts with a TF or an enhancer/promoter containing DNA sequence, it may have the ability to regulate cell function or gene expression. Due to 3 dimensional structure of the genome, this may occur between close ranges proximally or across a long range distally. Techniques such as chromosome conformation capture (3C) and Chromatin immunoprecipitation (ChIP) allow capturing of such interactions in their natural state within the cell. These can then be used to map relationships between proteins/DNA elements and the sequence containing the disease associated variant. The most common role for these techniques is to identify whether a SNP of interest lies within a region which can potentially bind an enhancer/promoter or a (TF) which can potentially dysregulate gene expression. Data from these techniques is used to populate the ENCODE project (section 1.6.1) but is often performed in a very small number of samples and therefore requires replication.

Chromosome conformation capture allows capture of interactions between DNA elements using formaldehyde cross linking, digestion by restriction enzymes and ligation to form 3C libraries. The libraries can then be quantified by quantitative polymerase chain reaction (qPCR) or resequencing to identify the interaction levels between the DNA elements. Additional digestion and ligation stages can be added to the technique in the form of 4C and 5C. As the structural confirmation of each cell type is different, the technique is performed in specific homogeneous cell populations of interest.

Chromatin immuno precipitation (ChIP) represents a similar technique which allows characterisation of interactions between proteins such as TFs and specific DNA sequences. This is achieved by allowing cross linking of protein-DNA interactions before shearing the DNA into small fragments. Specific antibodies are then used to select the proteins of interest and the attached DNA sequence is

purified. Resequencing can then be to use to identify the sequence of the bound DNA. As with 3C, this has to be performed in specific homogenous cell populations under controlled conditions. When the genotype of the cells is known, it can be used to show whether protein binding is affected by the presence of the disease associated variant.

1.7 Genetics of inflammatory arthritis

To date, multiple candidate gene and GWA studies have been published for IA. These are described in more detail below.

1.7.1 RA Genetics

The search for RA causal genes has spanned decades but the establishment of GWA studies has allowed for a rapid progression in the identification of disease associated genetic regions in recent years. As the power of genetic studies increases, progressively more gene regions have been confirmed as disease susceptibility loci. In many cases, associated SNPs are not located within a gene and lie within intergenic regions. In this case, the closest biologically relevant gene is assigned. Currently 101 confirmed RA loci have been identified, as described in more detail below . These loci have shown statistically significant associations with RA and several have been subsequently replicated in an independent cohort. The majority of these loci have been identified since 2008, indicating the essential role that GWA studies have played in the identification of common disease heritability. As many studies exclusively use RA cases which are seropositive for RF or ACPA many of these disease associations are driven solely by seropositive RA but the collection of larger numbers of samples for the, less common, RF- and ACPA- RA cases will see an increase in the number of seronegative disease associations identified.

1.7.1.1 The Major Histocompatibility Complex (MHC)

The major histocompatibility complex (MHC) has been the subject of genetic studies for over 40 years and was the first confirmed RA locus association to be replicated across multiple populations. To date numerous markers within this region have exhibited very high levels of association in multiple independent studies. Located on chromosome 6p21, the human leukocyte antigen (HLA) locus is a complex region, extending over 3.6 Mb and is the most densely packed mammalian gene region housing over 200 genes (Newton et al. 2004).

Structurally the region is divided into 3 segments, containing the class I, class II and class III gene regions. The class I segment houses the *HLA-A*, *HLA-B* and *HLA-C* genes which encode class I MHC molecules. These molecules are present on all nucleated cells and function to present intracellular proteins and potential antigens on the surface of the cell to CD8 T cells. The process is a critical component of the immune system for detecting intracellular pathogens such as bacteria and viruses. Further downstream of this region is the class II segment which houses the genes encoding the HLA-DR, HLA-DP and HLA-DQ molecules. Functionally these genes encode class α and β polypeptide chains, which assemble to form complete class II MHC molecules. These molecules are capable of presenting processed extracellular proteins on the surface of the cell and are essential in maintaining immune surveillance. Unlike class I molecules, class II MHC are only featured on specialised APCs such as macrophages and B cells, which are essential components of the adaptive immune repertoire. The presence of these molecules is essential to immunity as MHC Class II molecules uniquely interact with CD4 T cells via the T cell receptor. Structurally the class III genes are flanked by class I and class II genes yet have no antigen presentation role. Instead, they encode several cytokine and complement system molecules essential for chemotaxis and interactions between immune cells. Due to the crucial functional role of the MHC region, high variability between MHC molecules is critical and therefore a high degree of polymorphism is characteristic of genes in this region.

Subsequent to the primary *HLA-DR4* association reported by serology in 1987 several alleles within the *HLA-DRB1* gene have been associated with RA (Stastny 1976). The shared epitope hypothesis proposes that associated alleles encode for specific amino acid sequences which form specific protein structures in the third hyper-variable region of the antigen binding site (Gregersen et al. 1987). These structures have the ability to modify the antigen presentation properties of the class II molecule and therefore alter the immune response. The alleles associated with RA risk in various populations include DRB1*0401, DRB1*0404, DRB1*0405, DRB1*0408, DRB1*0101, DRB1*0102, DRB1*1001 and DRB1*1401 (Orozco et al. 2006). It has been shown experimentally that DRB1*0405 is most prevalent in Asian populations, DRB1*0401 and DRB1*0404 are predominantly associated with Caucasians whilst the DRB1*0101 allele is associated with RA risk in Israeli Jews (Newton et al. 2004). In addition, several disease protective alleles have been identified in Caucasian populations including DRB1*0103, DRB1*0402, DRB1*0802 and DRB1*1302 (Milicic et al. 2002). Most recently analysis of imputed HLA data in a large dataset (5018 cases and 14974 controls) of sero-positive RA cases and controls has shown 3 amino acid positions (11, 71 and 74) in *HLA-DRB1* combined with a single position at *HLA-B* (position 9) and *HLA-DPB1* (position 9) to be responsible for the majority of RA susceptibility within the HLA region in Caucasian Europeans (Raychaudhuri et al. 2012). All these positions are located in peptide binding grooves which indicate that these RA risk alleles affect the ability of the MHC molecules to bind peptide epitopes.

Overall, the genetic association at the HLA region is very complex, exhibiting multiple risk and protective effects. This region alone is estimated to contribute to approximately 30% of the genetic component of RA (Deighton et al. 1989), indicating that a significant portion of RA heritability occurs outside the HLA locus (Orozco et al. 2006).

1.7.1.2 Non-MHC RA Loci

The first comprehensive RA GWA study was published in 2007 by the Wellcome Trust Case Control Consortium (WTCCC) as part of a combined genetic analysis of 7 common diseases (Wellcome Trust Case Control Consortium 2007) . Genotyping

of ~500,000 SNPs in 1860 RA cases and 2938 shared controls was undertaken resulting in detection of 3 independent associations at stringent significance thresholds in the *MHC*, *PTPN22* and 7q32 regions. Replication of data obtained from this study has led to the confirmation of several associated genetic loci including *TNFAIP3*, *KIF5A* and *TRAF1/C5* (Barton et al. 2008b;Plenge et al. 2007;Thomson et al. 2007).

Subsequent meta-analysis of GWA study data has resulted in the discovery of a large number of RA susceptibility loci. This has been aided by increasing accuracy of imputation software, allowing analysis of variants which have not been directly genotyped. In 2010 a meta-analysis was performed on 5,539 RA cases and 20,169 controls (plus 6,768 RA cases and 8,806 controls for replication), which identified 7 novel RA disease associations in samples of European ancestry including variants near the *IL6ST* and *SPRED2* genes (Stahl et al. 2010). More recently 8 novel associations have been identified by meta-analysis of 4047 RA cases and 16,891 controls (plus 5277 RA cases and 21,468 controls for replication) in samples of Japanese ancestry (Okada et al. 2012). These included the *PTPN2* region which has also been associated in European populations, indicating a small amount of overlap exists. As analysis methods for meta-analysis grow more sophisticated, samples from different ancestries can be analysed together, greatly increasing the power of studies to detect genetic effects. These methods have been crucial in the recent identification of 42 novel RA susceptibility loci, which have been identified through a large trans-ethnic meta-analysis of 29,801 RA cases and 73,758 controls from both European and Asian ancestry (Okada et al. 2014b). These studies show that increasing power through sample size is still an important method for identifying novel susceptibility loci. Once novel susceptibility loci are identified the region is often investigated further to identify the causal variant which is contributing to disease susceptibility. In RA this has been achieved using the Immunochip fine mapping array and further functional analysis. All studies mentioned in sections 1.7.1.3-1.7.1.4 were performed after or in parallel to the studies described in this thesis, therefore they were not available to inform the current study.

1.7.1.3 The Immunochip study

Although GWA studies and meta-analyses are crucial tools in the identification of RA susceptibility loci, the limitations of genotyping arrays often mean that a region is covered by a small number of tag variants. LD patterns therefore make it challenging to identify the true causal variant within a region, so techniques such as resequencing and fine mapping are utilised. In 2009, the generation of the Immunochip fine mapping array allowed dense genotyping of previously identified RA susceptibility regions at a low cost per sample, allowing genotyping of a very large group of RA cases and healthy controls. This dense genotyping allowed haplotypic and conditional analysis to be performed with the aim of localising disease signals. In addition the genotyping of a large number of regions previously associated with other immune mediated diseases provided the opportunity to identify novel susceptibility loci and perform analysis of genetic overlap between different immune mediated diseases.

Analysis of fine mapping data for 11,475 RA cases and 15,870 controls resulted in the localisation of association signals in 19 disease regions, allowing the identification of 19 potential causal variants (Eyre et al. 2012). Furthermore using conditional analysis, 6 regions were identified as harbouring multiple genetic effects which may contribute to disease. This included the *TNFAIP3* region where 2 independent risk effects were identified and the *REL* region where an independent risk and protective genetic effect have been identified. In addition 14 novel RA susceptibility regions were identified, with many strong biological candidates identified for future analysis, including several variants in exonic regions which are known to affect protein function. Since the initial Immunochip study, subsequent studies have used the Immunochip array to perform their own analysis, including a recent study which identified 8 new RA susceptibility loci in a trans-ethnic analysis of samples from European and Korean ancestry (Kim et al. 2014). This indicates that the Immunochip is an excellent study design for identification of RA genetic susceptibility loci.

1.7.1.4 Identifying the functional role of RA associated variants

Although a large number of RA susceptibility loci have been identified, the function of many of these associations has yet to be elucidated. It has been shown that many of the RA susceptibility genes are specifically expressed in primary CD4+ T lymphocytes indicating that this may be a key cell type in the pathogenesis of the disease (Trynka et al. 2013). Furthermore analysis of the identified associations can lead to the identification of several specific pathways underpinning disease. This includes crucial immune factors such as the IL2 and TNF pathways, whose components include several known RA susceptibility loci. Additionally this is supported by the successful use of biological treatments such as anti-TNF and rituximab (anti-CD20) in the treatment of RA, which, although has not explained the disease mechanism, indicates that these immune components contribute to the disease susceptibility and pathogenesis.
As functional data becomes more accessible through bioinformatics data and advancement in techniques the picture will become clearer as to how these associations contribute to disease susceptibility and pathogenesis.

1.7.2 JIA genetics

To date several JIA loci have been identified using association studies but the heterogeneous nature of JIA and the fact that it is a relatively rare disease provides challenges in obtaining the large and thus appropriately powered, cohorts for analysis. As a result genetic studies are usually performed on the most common and more homogeneous subtypes: oligo articular and RF- polyarticular JIA, with rarer subtypes included in stratified analyses. Until 2009, all genetic association studies in JIA were performed purely on a candidate gene basis with only 3 regions consistently associated with JIA at genome-wide significance (*HLA*, *PTPN22* and *PTPN2*). More recently multiple small GWA studies have been performed resulting in the identification of 6 novel associations (Hinks et al. 2009a;Thompson et al. 2012). This has been supported by the establishment of international collaborations, providing the opportunity to increase sample sizes and therefore power to detect genetic effects. As sample sizes increase, we can expect more JIA

associations to be identified, giving a greater understanding of disease susceptibility.

1.7.2.1 The Major Histocompatibility Complex (MHC)

JIA is similar to RA in that Human leukocyte antigen (HLA) region provides the largest contribution to the genetic component of disease, with ~8-13% of heritability estimated to be explained by the HLA region (Hinks et al. 2013) . Many HLA associations have been characterized for different subtypes in Caucasian populations. The oligoarticular subset of JIA has the most HLA-associations with both the class I allele HLA-A2 and the class II alleles HLA-DRB1*08, DRB1*11, DPB1*0201, DQA1*04 and DQB1*04 associated with disease susceptibility. Furthermore DRB1*04 and DRB1*07 alleles confer disease protection in this subtype. In the polyarticular subtypes HLA-DRB1*08, DQA1*04 and DPB1*03 alleles are associated with RF- susceptibility whilst DRB1*04 and DRB1*07 is associated with protection. In addition HLA-DRB1*04 and DQA1*03 are associated with susceptibility in RF+ cases but DQA1*02 is associated with protection (Fernandez-Vina et al. 1994;Prahalad 2004;Thomson et al. 2002). Most recently, the Immunochip study, which investigated oligoarticular and RF- polyarticular disease, found that a SNP, which tagged the HLA-DRB1*0801–HLA-DQA1*0401–HLA-DQB1*0402 haplotype, represented the strongest signal, conferring disease protection. In addition conditional analysis identified 14 independent SNPs showing disease associations in the region (Hinks et al. 2013)

1.7.2.2 Non-MHC JIA Loci

1.7.2.2.1 Candidate gene studies

Until recently, the number of non-MHC regions associated with JIA was limited, with regions often selected for candidate gene investigation based on associations in other immune-mediated diseases. This method was successful in identifying variants within regions such as *STAT4*, *TRAF1/C5* and *IL2RA* but, although associated with disease in multiple cohorts, they did not reach genome wide significance level (Albers et al. 2008;Behrens et al. 2008;Hinks et al. 2009b;Prahalad et al. 2009). In addition, a polymorphism within the chemokine (c-c motif) receptor (*CCR5*) gene (32 base pair deletion mutation) has also been associated with JIA in 3 independent cohorts (Hinks et al. 2010c;Prahalad 2006) . The deletion leads to a shift in the reading frame and production of a non-functional receptor. Although this inability to respond to chemo-attractants would normally be detrimental, the association has been shown to be protective against JIA and other autoimmune diseases. This may be a result of suppression of inflammatory mediators which are responsible for disease processes.

As with RA, the *PTPN22* SNP rs2476601 was been shown to be associated with multiple subtypes of JIA, with the strongest associations detected in the polyarticular and oligoarticular subtypes (Cinek et al. 2007;Hinks et al. 2005;Viken et al. 2005). This still represents the largest non-MHC genetic effect identified to date.

1.7.2.2.2 GWA studies

To date, a limited number hypothesis free GWA studies have been performed in JIA, resulting in the identification of 6 novel JIA regions since 2009 (Hinks et al. 2009a;Thompson et al. 2012). The most recent and largest, testing 814 cases and 3058 controls, identified 2 regions (*CD80-KTELC1* and *JMJD1C*) to be significantly associated. The *JMJD1C* gene has been shown to remove methyl marks on histones in T and B lymphocytes, indicating a potential role in gene regulation (Thompson et al. 2012). As with RA, the region often identified by GWA studies is a tag SNP

variant and therefore further fine mapping is required to identify the causal variant in a region. In JIA this has been achieved using the Immunochip fine mapping array and further functional analysis. Therefore, the study mentioned 1.7.2.2.3 was performed after or in parallel to the studies described in this thesis; therefore they were not available to inform the current study.

1.7.2.2.3 The Immunochip study

The Immunochip study represents the largest and most comprehensive study of JIA genetics to date, with fine mapping data generated for 2816 oligo articular and RF- polyarticular JIA cases and 13056 controls (Hinks et al. 2013;Liu et al. 2008). The study successfully confirmed association with all previously known JIA loci at genome wide significance (*HLA, PTPN22* and *PTPN2*) whilst identifying 14 novel JIA loci. Several of these novel associations (*IL2RA* and *STAT4*) had previously shown association with JIA, but not at accepted genome wide significance thresholds. Therefore the Immunochip study represented the first well powered analysis of genetic susceptibility in oligo articular and RF- polyarticular JIA. Combined with the HLA findings described in section 1.7.2.1 the genetic associations in this study are believed to explain 19% of the heritability of JIA.

As a result of fine mapping and bioinformatics analysis, the disease association signal was localised in 8 regions to a single gene. In both the *IL2RA* and *IL2RB* regions the signal was localised to these genes. Combined with the association signal at *IL2/IL21*, this indicates that the IL2 pathway is important in pathogenesis of JIA and will require further investigation through functional studies to identify the exact role in contributing to disease susceptibility.

In total 17 genetic loci have now been associated with JIA at genome wide significance, with more showing suggestive significance ($<1x10^{-6}$). As sample sizes expand it is expected that more associations will be identified and confirmed, thus explaining more of the heritability of this disease. Although the functional consequences of these associations have not been investigated, the availability of bioinformatics databases and functional techniques provides a promising future for uncovering the pathogenesis of this disease.

1.7.3 PsA Genetics

To date a limited number of susceptibility regions have been identified for PsA. Although several GWAS have been performed using PsA cases, these have normally been in combination with PsV. As a result confounding of results is expected due to the genetic overlap between PsA and PsV. Currently 10 confirmed PsA loci have been identified. These loci have shown genome wide statistically significant associations with PsA and in many instances have been replicated in independent PsA cohorts.

1.7.3.1 The Major Histocompatibility Complex (MHC)

As with RA and JIA, The HLA region represents the locus with the greatest contribution to the genetic component of PsA. By contrast to RA and JIA, however, the strongest PsA association is with the class I gene segment HLA-C and not a class II segment. The allelic association which has been identified is with HLA-Cw*06, which was initially identified in cases with type 1 psoriasis (age of onset of psoriasis < 40 years) and type 1 psoriasis with arthritis (Gladman et al. 2005). Since then this association has been replicated in multiple PsV and PsA cohorts (Huffmeier et al. 2010;Liu et al. 2008;Nair et al. 1997). Although significant associations were obtained for PsA in the HLA-Cw*06, region, both p values and odds ratios indicate a greater effect in type 1 psoriasis cases than psoriasis with arthritis. This indicates that the HLA-Cw*06 allele may be associated with the skin rather than articular manifestations of PsA (Ho et al. 2008;Winchester et al. 2012).

In addition a SNP association in the *HLA-B27* region has been identified as a risk association for PsA but not PsV. The region has been previously associated with other seronegative spondyloarthropathies such as ankylosing spondylitis. The SNP rs116488202 has been shown to confer disease risk in 2 independent populations. The fact that this region is not associated with PsV indicates that, unlike *HLA-CW6*, it is involved specifically in the musculoskeletal manifestations of disease (Winchester et al. 2012).

1.7.3.2 Non-MHC PsA loci

In addition to the associations identified in the MHC region, multiple regions have been association with PsA.

Several SNPs in the interleukin 12B (*IL12B*) and interleukin 23 receptor (*IL23R*) regions have been associated with PsA and replicated in a UK, German and Canadian cohorts (Filer et al. 2008). As with *HLA-C*, both genes have shown increased association in type 1 psoriasis cases than PsA cases. A similar odds ratio for *IL12B* and *IL23R* in PsV and PsA cohorts indicates that the presence of articular disease does not have any significant effect on allele frequencies in cases and controls. This may also indicate the similar pathogenicity shared between skin lesions in PsV and PsA Despite this, it has been observed clinically that patients treated with Ustekinemab, a monoclonal antibody which inhibits IL2B and IL23R through the shared p40 subunit show improvements in both joint and skin disease (Gottlieb et al. 2009). This indicates that although genetic components appeared to be shared between PsV and PsA, several factors are still to be explored to identify the pathways of articular pathogenesis in PsA.

Additionally, a number of PsA associations represent interesting biological candidates which give some insight into the pathways responsible for disease. For example associations in the TRAF3 interacting protein 2 (*TRAF3IP2*) and interleukin 13 (*IL13*) regions encode proteins essential for B cell responses. *TRAF3IP2* in particular encodes NFKB activator (ACT1), an immune signalling adaptor which has a role in negative regulation of B cell signalling via the NFKB pathway. Furthermore *IL13* is a cytokine which is essential for B cell maturation and development and therefore risk variants in this region may result in an up regulation of the B lymphocyte response.

Additionally several PsA susceptibility regions encode genes involved in the activation and differentiation of T lymphocytes. In particular associations in the signal transducer and activator of transcription 2 (*STAT2*) and tyrosine kinase 2 (*TYK2*) regions are involved in the Janus kinase-signal transducer and activator of

transcription (JAK-STAT) pathway, acting as signal transducers to rely the message to activate the cell. In addition a risk association has been identified in the Runt related transcription factor 3 region (*RUNX3*), which encodes a transcription factor essential for activation and suppression of T lymphocyte enhancers and promoters. This has been shown to be particularly important in CD8+ T lymphocytes, which indicate that this cell type is important in PsA pathogenesis (Apel et al. 2013).

Furthermore strong evidence of association has been detected at regions involved in the NFKB signalling cascade, essential for cytokine production and response. This includes the TNFAIP3 interacting protein 1 (*TNIP1*) locus, which interacts with the A20 product of *TNFAIP3* to inhibit TNF induces NFKB expression (Huffmeier et al. 2009). In addition the association of the *REL* gene, a subunit of the NFKB pathway shows that this pathway looks to be crucial for the development of disease.

The increase of sample sizes in PsA studies over time has shown that with increased power, more susceptibility loci can be identified for this disease. In the past, genetic studies in PsA have also been a subset of larger PsV studies, but a recent Immunochip study represents the first PsA exclusive GWA study in the largest PsA sample size investigated, to date. The Immunochip study was performed after or in parallel to the studies described in this thesis, therefore they were not available to inform the current study.

1.7.3.3 The Immunochip study

The PsA Immunochip study was performed in a large cohort for this disease, with analysis of 1962 cases and 8,923 healthy controls performed (Bowes et al. Manuscript in preparation). In this study 8 regions which had been previously associated with PsV or PsA reached genome wide significance. Although the SNP identified was not always identical to that identified in PsV, the finding further confirms a significant genetic overlap between both diseases.

Interestingly a further 2 novel regions were identified in the *CSF2/P4HA2* and

DENND1B regions, at a suggestive level of $p<1x10^{-6}$. These SNPS were then taken forward for replication in an independent cohort of 572 PsA cases and 888 controls, the results of which were meta-analysed with the Immunochip results. Although the SNP in the *CSF2/P4HA2* replicated and reached genome wide significance in the meta-analysis at $p= 4.83x10^{-13}$, the SNP in the *DENND1B* did not reach significance. When examined in several independent PsV datasets, this SNP was found to reach suggestive significance at $p=2.4x10^{-7}$. This indicates that this association in the Immunochip analysis was likely to be due to the large number of disease cases which will have psoriasis, and therefore genetic susceptibility to PsV.

In order to localise the signal in the *CSF2/P4HA2* region, imputation and bioinformatics mining of databases was performed to elucidate the most likely causal variant(s). Three SNPs in high LD with the associated SNP were identified as eQTLs with *P4HA2*, *SLC22A4* and *SLC22A5* genes. This was examined in a cell specific eQTL dataset generated from CD4+ and CD8+ T lymphocytes from 23 healthy volunteers. Confirmation of the eQTL with *SLC22A5* was achieved in the CD8+ T lymphocyte data at $p=1.41x10^{-4}$. This finding provides further evidence of the crucial role of CD8+ lymphocytes in conferring susceptibility to PsA (Bowes et al. Manuscript in preparation).

1.7.4 Overlap of Susceptibility Loci

1.7.4.1 The Concept of Shared Loci

It has been previously observed that autoimmune and immune mediated diseases share characteristics of disease presentation and pathogenesis (Parkes et al. 2013). This may include predominantly female presentation and the presence of autoantibodies. As more disease associated loci have been identified using GWA studies it has become apparent that different diseases to some extent, share the same genetic susceptibility factors. Findings suggests that significant genetic overlap exists between immune mediated diseases such as RA, JIA, PsA, Type 1 diabetes (T1D), Crohn's disease (CrD), Celiac Disease (CD), Systemic Lupus Erythematosus (SLE) and Multiple Sclerosis (MS). It remains to be determined whether disease susceptibility is a result of the same alleles, disease specific alleles

or a combination. The overlap indicates that processes mediated by the products of these pan-autoimmune genes occur in several related diseases, highlighting the possibility of shared autoimmune pathways and networks but functional investigation is required for confirmation.

In many cases the associated polymorphism in a region may be identical across each disease but often disparity exists and different polymorphisms in the same regions are identified. This could be the result of truly different associations or a result of LD masking the true causal SNP within a region. Furthermore associations in different diseases may be differential in risk, conferring susceptibility in one disease and protection in another. As these regions have mainly been identified by GWA studies, which mean that often the SNP identified is a tag SNP which captures several variants, further analysis is required to identify causal variants and generate a true correlation between diseases.

1.7.4.2 Methods to identify genetic overlap

Recently, the development of various statistical techniques to determine overlap between diseases has led to the identification of a number of shared genetic factors. As these were developed and first published when the analysis in the current study was already underway, they were not used but are something to be aware of. Often these techniques can be applied to previously generated genetic data and can be applied to large sample cohorts, therefore increasing power. One example is pan-meta- GWA studies, which involves meta-analysis of existing GWA study datasets from different diseases, which share some overlapping features and are therefore expected to share genetic factors. This technique has been successful in the identification of three novel susceptibility loci for systemic sclerosis (Ssc) and SLE in a study which included 6835 combined Ssc and SLE cases versus healthy controls (Martin et al. 2013)

In 2011 Cotsapas et al. reported a method to assess genetic sharing across multiple diseases. This cross phenotype meta-analysis (CPMA) technique was applied to 107 SNPs previously associated with 1 or more of seven immune mediated diseases including RA, T1D and MS (Cotsapas et al. 2011). It was estimated using this technique that 44% of the 107 SNPs tested were shared across more than 1 disease, with a single region *SH2B3*, associated across the 7 diseases. Furthermore 4 distinct gene clusters were identified, with different immune mediated disease showing specific relatedness to each other. The method assesses the expected distribution of the SNP p values in each disease, using a likelihood ratio test and is believed to be a more powerful way to identify overlap than normal meta-analysis.

Although these techniques have been developed to assess the overlap between different diseases, candidate genes are often selected based on previous disease associations. This method has been pivotal in driving identification of numerous loci which are associated with multiple autoimmune and immune mediated diseases. This has included and to date has been the most popular method for identifying overlap between different types of IA.

1.7.4.3 Overlap between different types of inflammatory arthritis

As more associated polymorphisms for IA are identified, an overlap of susceptibility loci between the diseases has been emerging. As the current study involved the investigation of overlap using data generated in the Immunochip project, the following section will described the genetic associations/overlapping regions which were identified prior to the start of the current work.

Overall prior to the Immunochip study, 7 genetic regions showed association with RA and JIA, with evidence for a single shared genetic association across RA, JIA and PsA Figure 9. These are described in more detail below.

Figure 9 - Overlapping regions prior to the Immunochip study

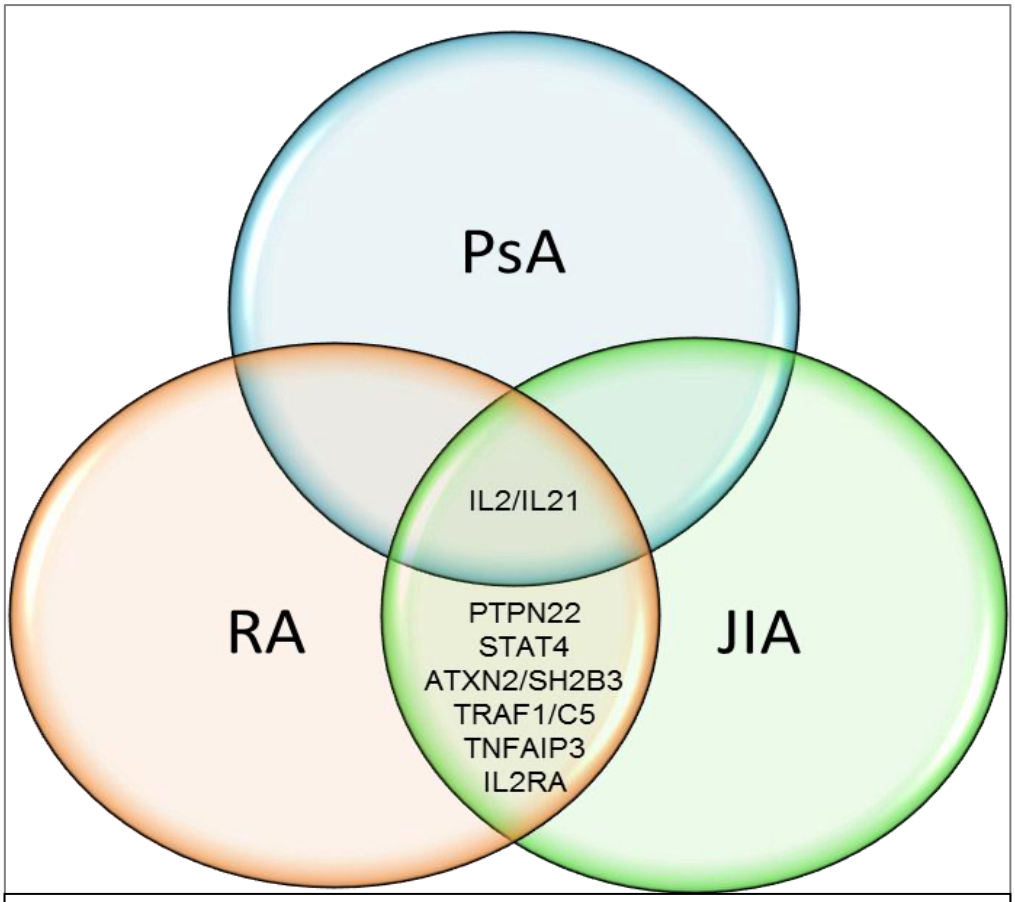


Figure 9 shows the number of regions associated with multiple types of IA, prior to this study (pre-2010). In total 7 regions were associated with more than 1 type of IA, with 6 regions associated across RA/JIA and 1 region associated across the 3 inclusively.

1.8 Inflammatory arthritis overlapping regions

1.8.1 PTPN22

Since the original report of association with T1D in 2004 , the protein tyrosine phosphatase 22 (*PTPN22*) gene has been identified as the second largest contribution to RA and JIA heritability, secondary to MHC. The casual variant rs2476601 has been shown to be associated with both RA and JIA across multiple Caucasian populations (Hinks et al. 2005;Pierer et al. 2006;Plenge et al. 2005;Seldin et al. 2005;Viken et al. 2005) . In each study, allele frequencies indicate that this polymorphism is a risk variant and confers susceptibility to disease. *PTPN22* represents a true pan-autoimmune locus as associations with the SNP have been replicated in a wealth of autoimmune disorders including CrD, MS, CD and SLE (Barrett et al. 2008;Todd et al. 2007;Wellcome Trust Case Control Consortium 2007).

Located on chromosome 1q13, the variant represents a T to C base change resulting in an arginine to tryptophan amino acid change at position 620 in the polypeptide chain, called LYP-W620. Consequently this induces a change in motif of the tyrosine phosphatase LYP, an 110kDa intracellular protein which is a key component in maintenance of T cell activation (Vang et al. 2005).

There have been several hypothesis generated to explain how this polymorphism alters induction of the immune response. One hypothesis is that this represents a gain of function mutation which results in a higher threshold for T cell stimulation and T cell hypo-responsiveness (Bottini et al. 2004). Crucially a key factor of immune tolerance is deletion of self-reactive T cells in the thymus. When the activation threshold is raised, this process will not work as efficiently, allowing self-reactive T cells to escape into the periphery. However, it has been recently shown in mice that this LYP-W620 does not cause any change in thymic output..

A more recent hypothesis proposed to explain the effect of LYP-W620 in RA is its role in neutrophils. In a study of RA patients and healthy controls, neutrophils from individuals with LYP-W620 were increased in migration markers, Ca2+ release and reactive nitric oxide synthesis in both cases and controls. Neutrophils are known to be characteristically increased in the RA joint, indicating that it may be hyper activation of these cells which is responsible for maintaining pathogenesis (Bayley et al. 2014).

1.8.2 STAT4

The signal transducer and activator of transcription 4 (*STAT4*) regions was initially identified as susceptibility region for RA using linkage analysis (Remmers et al. 2007). To date, in both RA and JIA, several associated SNPs which lie within the *STAT4* gene have been identified by candidate gene and GWA studies. Initially the SNPs rs7574865, rs8179673 and rs11889341 were shown to be associated with RA (Barton et al. 2008a). These three polymorphisms have also shown association with JIA indicating that in both diseases, there appear to be multiple effects in the region (Prahalad et al. 2009). In particular rs7574865 has been shown to be associated with poor outcome in early RA patients indicating it may have an important role in the early stages of IA pathogenesis (Lamana et al. 2012). It is not known yet whether these diseases will confer the same or different effects with regards to pathogenesis.

The functional product of *STAT4* is a vital component of the JAK-STAT signalling pathway and therefore is crucial to haematopoietic cytokine signalling. The potential of this pathway as a target for therapeutics is already apparent with the development of therapies which target tyrosine kinases within the pathway currently underway (Weinblatt et al. 2013).

1.8.3 ATXN2/SH2B3

The ataxin 2/SH2B adaptor protein 3 (*ATXN2/SH2B3*) region was first associated with RA in a combined investigation of RA and CD susceptibility loci. Although it reached genome wide significance in this combined analysis, it only reached suggestive significance in a subsequent RA specific meta-analysis but continued to be a region of interest (Coenen et al. 2009;Stahl et al. 2010). Furthermore this region has shown suggestive association with JIA across US and UK cohorts, indicating this is potentially an overlapping region for IA (Hinks et al. 2012).

The interesting features of these RA and JIA associations are that they are either with, or highly correlated by LD with the SNP rs3184504. This SNP is located in exon 3 of the *SH2B3* gene and is responsible for an amino acid change at position R262W. As *SH2B3* encodes the T cell adaptor protein Lnk this change could potentially cause dysregulation of T lymphocytes, which are known to be important in IA pathogenesis. Lnk has also been shown to regulate endothelial cell signalling in response to (Fitau et al. 2006). Carriers of the R262W polymorphism have also been shown to have strong NOD2 signalling responses, indicating that Lnk may be involved in defence against bacterial pathogens (Zhernakova et al. 2009).

1.8.4 TNFAIP3

It has been shown that two variants (rs6920220 and rs13207033) within close proximity to the tumour necrosis factor, alpha-induced protein 3 (*TNFAIP3*) region have been associated with both RA and JIA (Prahalad et al. 2009;Wellcome Trust Case Control Consortium 2007).One SNP rs6920220 confers increased risk of developing RA and JIA whilst the SNP rs13207033 is protective, indicating differential effects may occur across different diseases in this region. In RA a third independently associated SNP, rs5029937, within an intron of *TNFAIP3* has been identified. A study investigating the nature of these polymorphisms in RA has

shown that these three SNPs exhibit independent genetic effects. Furthermore it has been shown that the carriage of risk alleles rs6920220 and rs5029937 combined with the absence of the protective polymorphism rs13207033 confers a greater risk of developing disease (Orozco et al. 2009).

The *TNFAIP3* gene encodes the enzyme A20, which is involved in the ubiquitin editing and negative feedback regulation of immune signalling cascades of the NFKB pathway. The pathway is essential for the transcription of pro-inflammatory cytokines and survival of effector cells during an immune response, therefore A20 is an important feature in preventing autoimmunity (Baltimore 2011). Furthermore, expression of A20 by dendritic cells has been shown to regulate immune homeostasis and prevents spondyloarthritis in *TNFAIP3* knockout mice (Hammer et al. 2011). This has also been found in myeloid cells as mice that do not have functioning A20 develop spontaneous polyarthritis. This is accompanied with an increase in inflammatory mediators such as IL-6 and increased osteoclast activation, which are characteristic of RA. Altered *TNFAIP3* expression has been correlated with response to etanercept in RA patients. But the mechanism by which this occurs is not yet understood (Koczan et al. 2008).

These findings indicate that *TNFAIP3* represents a very interesting region, which due to its crucial role in immune homeostasis, should be investigated further to identify the mechanism through which it confers disease susceptibility.

1.8.5 TRAF1/C5

Several SNPs which are located between the tumour necrosis factor associated factor 1 (*TRAF1*) and complement component 5 (*C5*) genes have been shown to be associated with RA and JIA. Currently a single identical SNP rs3761847 has been associated in both RA and JIA cohorts (Albers et al. 2008;Behrens et al. 2008;Plenge et al. 2007). This SNP lies in a region of high LD and therefore further analysis is required to identify the causal variant. The *TRAF1* gene encodes a member of the TNF Receptor Associated Family (TRAF) which mediates signalling

transduction from a number of TNF receptors. The *C5* gene encodes the fifth member of the inflammatory complement cascade and in its active form is a potent inflammatory mediator. Risk polymorphisms in this region could result in uncontrolled stimulation of TNFR and the complement cascade, generating inflammatory pathways such as that seen in both RA and JIA.

1.8.6 IL2RA

The interleukin 2 receptor alpha (*IL2RA*) region was primarily reported as a T1D association using a family based study (Vella et al. 2005). This was subsequently replicated and the association pinpointed to this locus using fine mapping analysis of the region (Lowe et al. 2007).

A SNP in the *IL2RA* region has been identified as a shared locus in RA and JIA. This SNP rs2104286 appears to confer disease protection in both RA and JIA (Hinks et al. 2009b;Thomson et al. 2007). The same SNP has also been shown to confer protection in MS but risk in T1D (Hafler et al. 2007;Maier et al. 2009). Furthermore this region has also been associated with joint destruction and poorer disease outcomes in RA (Knevel et al. 2013).

This region represents a strong functional candidate for analysis as the product of the gene *IL2RA/CD25* is a subunit of the IL-2 receptor, which is essential for Treg induction. Alteration of the subunit could result in a non-functional receptor and therefore a reduction in T-regs, which are essential in maintaining peripheral tolerance.

1.8.7 IL2/IL21

The interleukin 2/interleukin 21 *IL2/IL21* gene region represents a region which contains SNPs which have been shown to be associated with RA, JIA and PsA inclusively. In each study, a different SNP has been identified to be associated with each disease. This may be due to true differences in disease association or masking of true associations by LD within the region and limited capture on SNP

genotyping arrays.

The SNP rs13119723 has been shown to confer disease susceptibility in RA (Zhernakova et al. 2007) and has previously shown risk in both ulcerative colitis and celiac disease (Glas et al. 2009;van Heel et al. 2007). In addition the SNP rs6822844 has been shown to be a protective allele in JIA cases (Hinks et al. 2010a). In PsA cases, the SNP rs13151961 has been shown to be associated with disease protection (Liu et al. 2008). The SNPs do not appear to be correlated by LD which indicates they may represent different effects but this will require further analysis as the region is located in a block of strong LD so identification of effects is challenging.

As the *IL2/IL21* region also represents a genetic susceptibility region in SLE, a fine mapping study was performed in that disease to localise the signal. 45 tag-SNPs were directly genotyped and imputation performed in 4248 SLE cases and 3813 healthy controls. The study was successful in localising the signal to 2 independent SNPs in the *IL21* gene (Hughes et al. 2011). In order to confirm this for RA, JIA and PsA, testing of the SNPs would ideally be performed to determine whether the effects are identical or different between these diseases.

The *IL-2/IL21* region is an ideal candidate gene for investigation as it contains 2 biologically interesting genes. The *IL2* gene encodes a cytokine which is a survival factor for T cells and NK cells and is especially important for the generation of T regulatory cells, which are essential for maintenance of immune regulation. The *IL21* gene encodes a cytokine which promotes the differentiation of Th17 T cells and B cells.

1.9 Summary

Until the start of the current study, candidate gene and GWA studies provided a wealth of information regarding genetic susceptibility to immune mediated diseases, including over 50 suggestive genetic associations with IA. Some of these associations were shared across more than one disease indicating the existence of

common pathways which underlie disease pathogenesis. Although this is an interesting finding, it was believed that greater overlap could be present and that to identify the true common pathways between these diseases, causal variants would have to be identified.

As all the associations mentioned the section 1.8 were identified by candidate gene and GWA studies the majority have represented an association with a tag-SNP. Due to LD throughout the genome, it was believed that in the majority of cases this would not represent the true causal variant in the region. To identify causal variants, fine mapping or resequencing of a region would have to be performed to increase the chance of localising the disease signals. As resequencing is costly and relatively low throughput this drove the establishment of the Immunochip consortium and therefore the generation of the Illumina Immunochip array (Cortes and Brown 2011).

1.10 The Immunochip

The Immunochip array was designed to address many of the constraints of previous genotyping arrays. Firstly the array is specifically designed to perform dense fine mapping of regions which have been previously associated with disease, with regions being selected by a consortium of 12 immune mediated diseases shown in Figure 10.

For each disease associated region, all 1000 genomes variants within 0.1cM recombination blocks around the associated SNP were included on the array. This was supplemented by inclusion of a number potentially interesting candidate regions from each disease, providing dense SNP coverage across a large number of regions. Furthermore genome wide tag SNP coverage allowed identification of novel loci which had not been previously investigated in each disease. The multi-disease nature of the study also allowed identification of loci which are associated across more than one disease, providing a greater understanding of the shared genetics of immune mediated disease. (Cortes et al. 2011). It is this feature of the array which makes it ideal for identifying overlapping genetic susceptibility loci for

different types of IA.

Figure 10 – Immune mediated diseases which contributed to the Immunochip

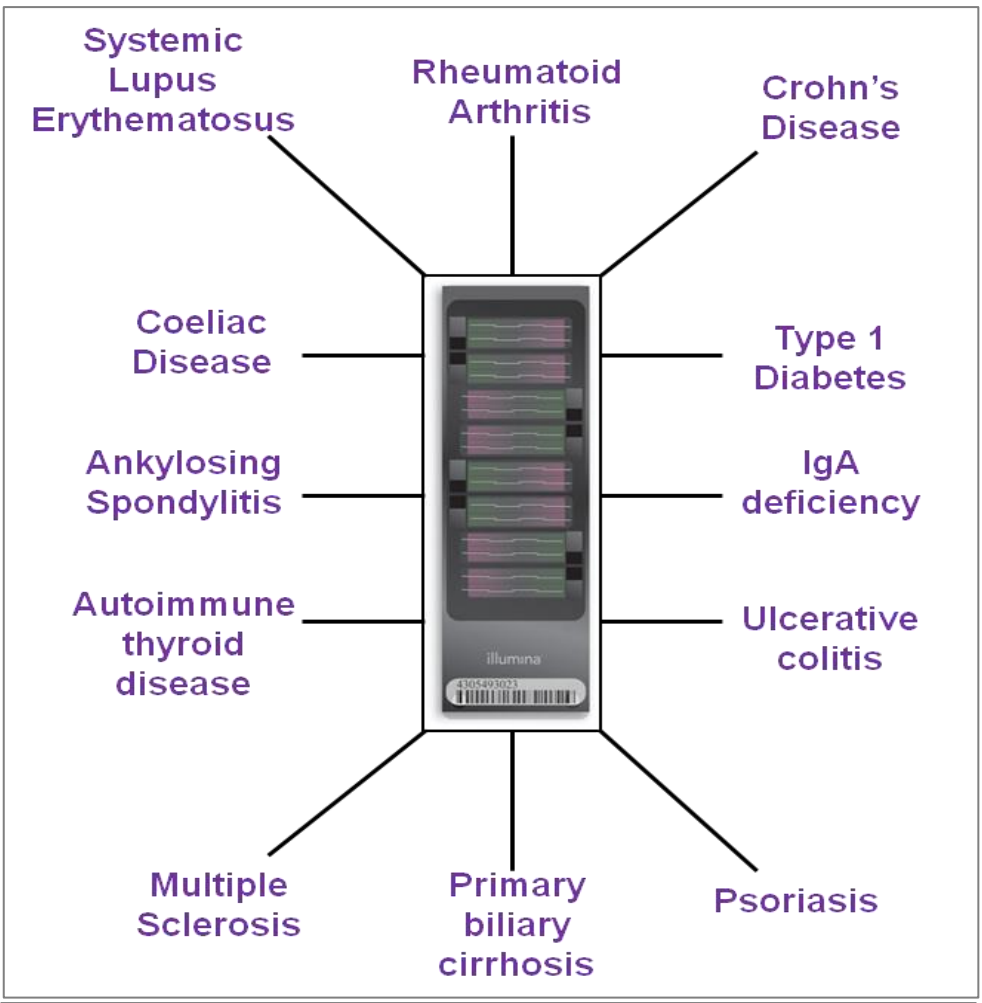


Figure 10 details the 12 immune mediated diseases (including RA), which initially contributed to the Immunochip array. A photo of the array used is shown in the centre.

1.11 Aims of study

The aim of this project is to identify the genetic overlap between the three different types of IA using data generated using the Immunochip array. A biologically promising region will then be selected for further analysis. Further genetic analysis will be performed to identify the likely causal variant in the region, followed by functional investigation using gene expression studies in relevant tissues. This will be used to determine the functional role of this region and therefore its contribution to the genetic susceptibility of IA.

1.12 Objectives

- To perform overlap analysis of IA using data generated on the Immunochip array.
- To select a genetic region associated with more than one disease, which will be subjected to further analysis.
- To perform further analysis of the selected region with the aim of the identifying true causal variants.
- To identify a biological role for selected region using bioinformatics databases and laboratory techniques.

2.0 Methods

2.0 Methods

2.1.1 Inflammatory arthritis overlap

Inflammatory arthritis describes a group of diseases, which, although clinically distinct, share common clinical characteristics such as synovial joint inflammation and response to treatment.

In this study identification of common genetic susceptibility factors for IA was performed by genotyping a large number of SNPs across the 3 diseases. A direct comparison of associated regions was then performed to identify overlapping regions. This was greatly facilitated by the establishment of the multi-disease Immunochip consortium, which resulted in the generation of a custom genotyping array ideal for identifying overlap between multiple diseases.

2.1.2 Subjects

Samples from 3 types of IA (RA, polyarticular/oligoarticular JIA, and PsA) and healthy controls were genotyped using the Immunochip Illumina Infinium custom array in accordance with Illumina protocols (http://support.illumina.com/downloads/immunochip_product_files.ilmn). The majority of samples were genotyped at several sites, however I genotyped 500 IA samples at the Arthritis Research UK Centre for Genetics and Genomic.

Table 2 shows details of the number and source of samples genotyped. A number of common controls were used by all diseases. All cohorts were comprised of individuals of European descent. Quality control and case control analysis were performed separately for each disease as described previously (Eyre et al. 2010;Hinks et al. 2013). .

Table 2 – Total number of samples included in Immunochip analysis

Disease	Cohort	Cases	Controls (Shared)
RA	UK	3870	8430
	Swedish EIRA	2762	1940
	US	2536	2134
	Dutch	648	2004
	Swedish UMEA	852	963
	Spanish	807	399
	TOTAL	11475	15870 (8430 shared)
JIA	UK	772	8530
	US	1596	4048
	German	448	478
	TOTAL	2816	13056 (8530 shared)
PsA	UK	929	4537
	TOTAL	929	4537 all shared

Table 2 shows the number of samples used in the Immunochip analysis. RA = Rheumatoid arthritis, JIA = Juvenile idiopathic arthritis, PsA = Psoriatic arthritis.

2.1.3 Illumina Infinium HD assay genotyping

The Illumina Infinium HD assay allows high throughput analysis of genetic variation in a population. This makes it ideal for large case control studies to discover genetic associations with disease. Each assay involves 3-phase treatment of whole genome DNA before hybridisation to a bead chip array

In the first phase, genomic DNA is denatured, neutralised and amplified overnight. It is then enzymatically fragmented. The use of end point fragmentation allows consistent fragmentation to occur whilst maintaining sample integrity.

Samples are then precipitated using iso-propanol and re-suspended before hybridisation to the beadchip array. Sample hybridisation to the array is achieved by annealing of fragmented sample to specific 50-mers attached to locus specific beads on the array. Post hybridisation, arrays are cleaned to remove unhybridised DNA and sample loaded chips are stained using the x stain HD process. This method uses labelled nucleotides to extend the DNA by a single base extension before harnessing avidin and biotin technology to amplify the signal exponentially. The process incorporates detectable labels on the array for the genotype calls to be made accurately.

Finally beadchip arrays are scanned and analysed using the iScan reader system. The system uses a laser to excite the fluorescence of the single base extension product which is visualised in high resolution to determine the presence of different alleles in samples at positions across the genome.

Figure 11 – Illumina workflow

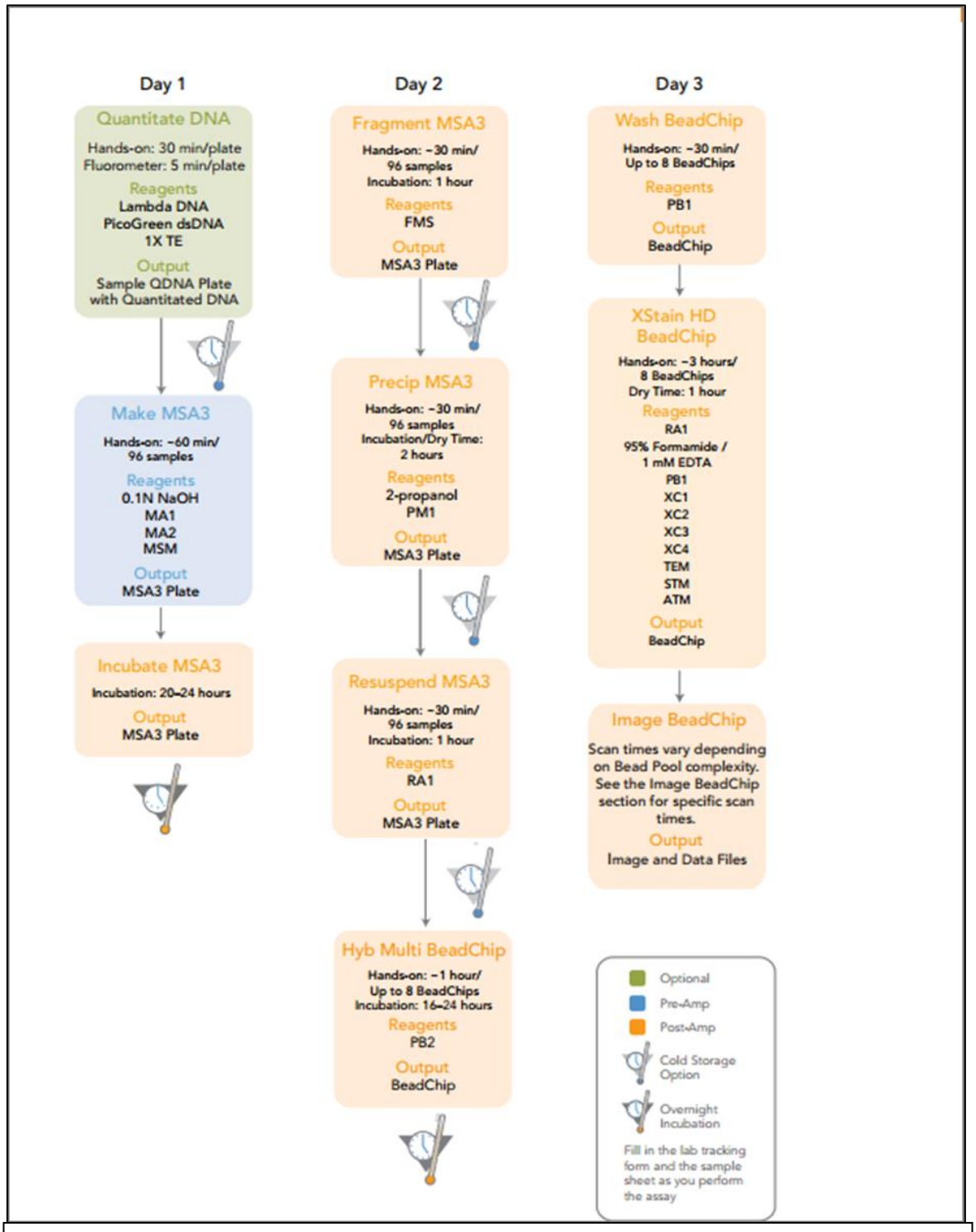


Figure 11 shows the workflow for the Illumina HD assay featuring the amplification, fragmentation, hybridisation and staining stages. Adapted from Infinium HD Assay Ultra Manual Workflow Rev. B (http://support.illumina.com/downloads/immunochip_product_support_files.ilmn).

2.1.3.1 DNA quality control

Whole blood DNA was extracted from 10ml blood using phenol chloroform extraction by technical teams at a number of sites described in Table 2.

Post extraction DNA samples were stored at -20°C. Prior to genotyping, DNA samples were assessed for quantity and purity using the Nandrop-N1000 spectrophotometer (Thermo Scientific). This spectrophotometer uses UV light absorbance ranging from 200-350nm to quantify nucleic acids present in a sample and determine the presence of any contaminants. As maximum DNA absorption takes place at A260, absorption at this wavelength is analysed and converted to a quantifiable concentration of ng/µl using the Beer-Lambert Law. Absorption values are also obtained at A260 and A280, which are used to generate the 260/280 ratio. As pure double stranded DNA has a 260/280 ratio of 1.8, the closer to this figure, the more likely it is that the sample is pure and free of contaminants. Deviations from this number are often caused by the absorbance of proteins at 280nm, phenol at 260nm and thiocyanate salts at 230nm. The formulas used to determine quantity and purity are shown below in Figure 12. Samples with an adequate concentration and a 260/280 ratio of ~1.8 were considered suitable for genotyping. Suitable samples were diluted to working dilutions of 50ng/µl for genotyping. The following protocol is based on the Infinium HD Assay ultra manual protocol (Illumina 2009) with minor optimisations recorded. All reagents were provided by Illumina (San Diego, United States) unless otherwise stated.

Figure 12 – Formula to determine DNA sample concentration and purity

$$\text{DNA quality} = \frac{\boldsymbol{A260}}{\boldsymbol{A280}}$$

$$\text{DNA concentration} = \frac{\mathbf{40}}{\boldsymbol{A260}}$$

2.1.3.2 Amplification of DNA

To prepare for genotyping, 4µl of 50ng/µl DNA was added to a 96 well semi deep well plate to give 200ng genomic DNA in total. To denature and neutralize samples, 20µl MA1 reagent and 4µl 0.1M sodium hydroxide (Sigma Aldrich) was added to each sample well. After shaking at 1600rpm for 1 minute using the High Speed Microplate Shaker (WVR) samples were incubated for 10 minutes at 27°C in the hybridisation oven (Illumina). Post incubation, 34µl MA2 reagent and 34µl MSM reagent were added. Samples were shaken at 1600rpm for 1 minute and incubated overnight at 37°C for 20 hours in the hybridisation oven.

2.1.3.3 Fragmentation and precipitation of DNA

To enzymatically fragment DNA, 25µl FMS reagent was added to each sample well; plates were shaken at 1600 RPM for 1 minute and incubated at 37°C for 1 hour on a Hybex microsample incubator (Scigene). The use of end point fragmentation prevents over-fragmentation of DNA and maintains sample integrity. DNA precipitation was then performed by adding 50µl of PM1 reagent, shaking at 1600rpm for 1 minute and incubating at 37°C for 5 minutes on a Hybex microsample incubator (Scigene). Post-incubation, 155µl of 2-isopropanol (Sigma Aldrich) was added and contents mixed thoroughly by a 10 times inversion. Samples were incubated at 4°C for 30 minutes before being centrifuged at 3000xg for 20 minutes at 4°C. Post centrifugation, the supernatant was decanted thoroughly and the remaining DNA pellet left to air dry at 27°C for 1 hour.

2.1.3.4 Resuspension of DNA and hybridisation to bead chip array

To resuspend the DNA pellet, 23µl RA1 reagent was added and samples incubated at 48°C for 1 hour in the hybridisation oven. Post incubation, samples were shaken at 1800rpm on the sample shaker to completely re-suspend the DNA pellet. Samples were then denatured for hybridisation by heating at 95°C for 20 minutes

on a Hybex microsample incubator (Scigene) followed by cooling at 27°C for 30 minutes. 20µl of each sample was then loaded onto the corresponding bead chip array and placed in humidified hybridisation chambers at 48°C for 16 hours in the hybridisation oven.

2.1.3.5 Washing of bead chip array

To remove unhybridised DNA from the surface of the chip, each chip was washed in PB1 reagent twice for 1 minute per wash. Each bead chip flow through assembly chamber was then assembled using the bead chip alignment fixture, plastic spacers and glass slides in preparation for single base extension and bead chip staining.

2.1.3.6 Single base extension and bead chip staining

The flow through assembly chambers were placed in the TE-flow rack chamber (Tecan) at 44°C for the single base extension stages described in Table 3. The TE-flow rack chamber temperature was then adjusted to 37°C for the staining stages described in Table 3. The flow through assembly chambers were then dismantled and bead arrays washed in PB1 reagent for 5 minutes before coating in XC4 reagent for 5 minutes.

Table 3 – Beadchip xStain stages

Reagent added	Incubation Time (Minutes)	Repeat	Staining Rack Temperature (°C)
150µl RA1	0.5 (30 seconds)	x5	44
450µl XC1	10	NA	44
450µl XC2	10	NA	44
200µl TEM	15	NA	44
450µl formamide/1mM EDTA	1	x1	44
None	5	NA	37
450µl XC3	1	x1	37
250µl STM	10		37
450µl XC3	1	x1	37
None	5		37
250µl ATM	10		37
450µl XC3	1	x1	37
None	5		37
250µl STM	10		37
450µl XC3	1	x1	37
None	5		37
250µl ATM	10		37
450µl XC3	1	x1	37
None	5		37
250µl STM	10		37
450µl XC3	1	x1	37
None	5		37

Table 3 shows the series of stains used in the BeadChip xstain. The volume added, incubation time, number of repeats and incubation temperature is shown.

2.1.3.7 Imaging of bead chip array on iScan system

Bead chip arrays were loaded into the iScan system using the iScan control software v.1.0. Decode content files (DMAP files) were downloaded using Illumina decode file client v.3.0.2 for each individual array and were combined with corresponding fluorescence data to perform genotype calls. All preliminary analysis was performed in Illumina GenomeStudio v.1.0 before being exported in PLINK v.1.07 format for further analysis (Purcell et al. 2007).

2.1.4 Immunochip SNP and sample QC

All QC was performed in house by Dr John Bowes (RA and PsA QC) and Dr Anne Hinks (JIA QC). Visual inspection of a number of genotype clusters determined the cluster separation score cut-off which was used. As a result different thresholds of 98-99% were defined for each disease.

For sample QC a call frequency threshold was set and samples with a lower success rate than this value were excluded from further analysis.
Autosomal heterozygosity is used as an additional quality control to identify samples with an over- or under-abundance of autosomal heterozygous SNPs. Mean genotype heterozygosity (excluding X chromosome markers) across all samples was assessed and any samples showing higher or lower levels of sample heterozygosity may be the result of genotyping error or sample contamination and were removed from further analysis. Mean heterozygosity across a study will depend on the population and SNP genotyping panel so needs to be calculated for each dataset and appropriate thresholds determined. Generally only a small number of individuals are removed with this QC step.

As relatedness and population sub-structure are also major sources of confounding in case control studies, methods were adopted to remove inappropriate samples. These are described in the next section.

2.1.4.1 Identity by descent analysis

To avoid over-representation of particular genotypes, all samples included in a case control analysis should be unrelated. Identity by descent (IBD) analysis measures how many alleles are shared at a genotyped position between two individuals. If two individuals share more DNA than is expected by chance, they are considered to be related. IBD analysis was performed using the --genome command in PLINK v.1.07 to remove duplicates and first/second degree relatives. Both identity by descent (IBD) and principal component analysis (PCA) (section 1.4.5) require the presence of independent genetic markers. To enable this analysis, the MHC and 17 additional high LD regions were excluded from these analyses. In addition SNPs were pruned for LD between markers using a sliding window approach based on r^2=0.2. Values for shared IBD are obtained by performing genome wide pair-wise comparisons of genetic markers between samples. Duplicate samples and monozygotic twins have a shared IBD of 1. Siblings have a shared IBD of 0.5, half siblings a value of 0.25 and third degree relatives a value of up to 0.175. It is also possible to have false positive relatedness as a consequence of sample duplication or contamination between samples. Duplicates and related individuals were excluded from analysis to minimise the risk of confounding.

2.1.4.2 Principal Components Analysis

Confounding by population stratification is often the result of case and control mismatching in a genetic study. It can result in false positive genetic association with a disease. In a genetic association study you are looking for differences in genotype frequencies between cases and controls. The genotype frequencies for SNPs vary across different populations, so unless you carefully match your case and control populations you may get artificial genotype differences due to different populations in the cases and controls. Principal components analysis (PCA) is a statistical model used to detect ancestry of samples using genetic markers which indicate variation between populations. The analysis calculates continuous axes of genetic variation (eigenvectors) or PCs that reduce the data to a small number of

dimensions, whilst describing as much of the variability between individuals as possible. The model is built using an LD pruned genome-wide dataset of independent SNPs, often using data from Hapmap samples of known ancestry. The model can then be applied to novel samples and individuals can be removed who appear to have an ancestry different from the rest of the dataset. In addition, the PCs can be used as covariates in the association analysis to account for more subtle gradients of ancestry in the dataset,

PCA analysis was performed using EIGENSOFT V.4.2.(Patterson et al. 2006;Price et al. 2006) To maximize homogeneity between samples, 5 PCA iterations were performed, outlying individuals were removed after each iteration, and the principal components from the 5th iteration were then used as covariates in the logistic regression analysis described in section 2.1.4.4.

2.1.4.3 Hardy-Weinberg Equilibrium

To identify any genotyping error that was not detected during previous QC, statistics were generated for each SNP by calculating the Hardy Weinberg Equilibrium (HWE). The principle can be used to predict frequencies of expected genotypes in a population and is based on the equation: $(p^2) + (2pq) + (q^2) = 1$.

In the equation, if p represents allele A and q represents allele a then it is expected that p + q = 1 in a population. Therefore the frequency of a homozygous genotype AA is p^2, aa is q^2 and 2pq is heterozygous Aa genotype. Under expected HWE these values will add up to 1. Deviations from these expected values in observed genotypes, so for example if you see more heterozygous genotypes than you would expect, can indicate genotyping error or association with disease; therefore this technique is usually performed in healthy controls exclusively. If a SNP deviates from HWE in controls then it should be excluded from further analysis.

The dataset was tested for HWE using the --hwe command in PLINK v.1.07. This was performed for cases and controls separately but only the results in controls

were used to exclude SNPs. The command performs a Chi^2 genotypic test for each SNP, generating a p value indicating conformation or deviation from the expected HWE value. Once all association tests were performed, SNPs with a HWE p value of less than 0.001 in controls were considered deviant from HWE and excluded from further analysis.

Following all the QC stages, a final dataset was built excluding all SNPs and samples that failed QC before analysis.

2.1.4.4 Association analysis

All association testing was performed in house by Dr John Bowes (RA and PsA) and Dr Anne Hinks (JIA). Association testing was performed separately from each disease using the logistic regression model. This model allows assessment of allele frequency differences between cases and controls, with option of using values generated in the PCA analysis as covariates. This allows the analysis to account for subtle differences in allele frequencies generated as a consequence of population differences, which may be picked interpreted as a spurious disease association otherwise.

Logistic regression analysis was performed using the –-logistic command in PLINK V.1.0.7, using the principal components generated in section 2.1.4.2 as covariates. Results were generated under the additive genetic model and output of MAF in cases/controls, logistic regression p values, odds ratios and 95% CI were used to perform the overlap analysis in section 2.1.

2.1.5 Power of each disease to detect genetic effects

Power to detect common genetic effects for the sample sizes available for each disease were calculated using the unmatched case-control model in QUANTO v.1.2.4. Results are expressed as the percentage power to detect an effect at a standard odds ratio (1.2) in both common minor allele frequency (MAF>0.05) and low frequency SNPs (MAF>0.01).

2.1.6 Calculating the number of inflammatory arthritis overlapping regions

All SNPs which passed QC and had a MAF of greater than 0.01 were included in the analysis. Initially, additive model p values for these SNPs were extracted from the case control analysis for each individual disease and SNPs which reached $p<10^{-3}$ in any individual disease selected for further analysis. Association results for these particular SNPs from all 3 diseases were combined. Genetic regions were then analysed individually to determine regions that contained a variant(s) associated with more than 1 disease. Regions were defined by the proximity to genes from the NCBI36 gene build on a genome wide basis. In many cases this was defined by the fine mapping regions covered on the Immunochip, defined as covering all variation from the 1000 Genomes project (September 2009 release) (Abecasis et al. 2012) in 0.1cM recombination blocks around the previous GWAS region lead marker. The level of significance of overlapping regions were designated as either genome wide ($p<5x10^{-8}$) or suggestive ($p<1x10^{-3}$), with the index SNP referring to the most associated SNP in each region for each disease. Association plots were generated regionally for each disease using LocusZoom (http://csg.sph.umich.edu/locuszoom/).

2.1.7 Identifying correlation between SNPs in overlapping regions

Although To identify correlation between associated SNPs in overlapping regions, the LD between index SNPs in each disease was calculated using the --ld function in PLINK v1.07. Highly correlated SNPs were defined as having an $r^2 > 0.8$, moderately correlated had an $r^2 >0.4<0.8$ and $r^2 <0.2$ were considered weak/not correlated.

2.1.8 Selecting a functionally promising region for further analysis

Although association studies are a powerful tool for identifying genetic regions associated with IA, ideally the functional consequence of a genetic association should be determined using additional functional techniques. Therefore, to complement the Immunochip overlap analysis further functional investigation of a

biological promising overlapping region was performed. To decide which region was most suitable for further investigation a number of factors were considered. These included how many of the different types of IA the region was associated with, the size of the association p value for each disease and whether it is the same/highly correlated SNP that was associated with each disease.

Once a promising region was selected, intense bioinformatics mining was performed to predict the function of the associated SNP and its proxies (r^2>0.9) All proxies were obtained using the SNP annotation and proxy search database (SNAP (http://www.broadinstitute.org/mpg/snap/ldsearch.php)). Bioinformatics analysis performed included functional annotation, cis- and trans- expression quantitative trait loci eQTL analysis and transcription factor binding analysis. In addition a literature search was performed to identify any previous disease associations or any previously characterized biological role for the selected region.

2.1.9 Functional annotation

Functional annotation was performed using the relative location track on ASSIMILATOR (http://assimilator.mhs.manchester.ac.uk/cgi-bin/assimilator_new.pl). This program uses data generated by the Encyclopaedia of DNA element s (Encode https://genome.ucsc.edu/ENCODE/) to inform functional prediction of disease associated SNPs.

2.1.9.1 eQTL analysis

To identify whether associated SNPs in regions of overlap between the different forms of IA correlate with gene expression levels, bioinformatics mining of eQTL databases was performed. As the presence of eQTLs is extremely variable between different tissues, data from a range of tissues were examined. Cis-eQTL analysis to identify gene regulation at close proximity was performed using Genevar gene expression variation database whilst trans-eQTL analysis to identify long range interactions was performed using the SNP and CNV database Scan DB

(http://www.scandb.org/newinterface/about.html) (Gamazon et al. 2010;Yang et al. 2010). In each case a p value threshold of $1x10^{-3}$ was adopted as the criteria to define an eQTL of interest.

2.1.9.2 Transcription factor binding analysis

To identify whether associated SNPs in regions of overlap lie in regions which alter transcription factor binding, analysis of the UCSC genome browser data was performed (http://genome.ucsc.edu/) (Karolchik et al. 2014) . As with eQTL analysis, the presence of transcription factor binding sites (TFBS) is variable between different tissue types and between different transcription factors, therefore evidence of transcription factor binding was analysed across a large number of cell types/transcription factors. Any SNP with a score of greater than 500 was considered a strong interaction whilst scores of <500 were considered to represent weak evidence for TF binding.

2.1.9.3 Literature search

To gain an insight into the potential biological role of an overlapping region, the gene names were entered as keyword search queries into NCBI PubMed (http://www.ncbi.nlm.nih.gov/pubmed). Publications were then mined for information about potential biological roles of the genes.

2.2 Replication of overlapping associations

As a number of the overlapping IA loci identified have not been previously associated with disease, replication in an independent cohort provides additional evidence that these are genuine associations and not due to type 1 error. The Sequenom MassARRAY system is the ideal genotyping platform for replication studies. In contrast to the Illumina platform which allows genotyping of a large number of SNPs, this platform allows genotyping of a smaller number of SNPs in

large sample cohorts.

Overlapping regions, which are not included in the individual disease replication projects, were considered candidates for replication. In some cases SNPs were already being genotyped on other platforms or the signal had already been localised on Immunochip so did not require replication. The SNPs were genotyped in an independent RA cohort, selected due to sample availability, as RA is the most prevalent of the 3 types of IA.

2.2.1 SNP Assay Design

Forward/Reverse amplification primers and extension primers were designed using the Sequenom MASS array iPlex assay design suite v.1 (https://www.mysequenom.com/assaydesign).

In some cases SNPs assays could not be designed, due to high risk of non-specific primer binding or the formation of "primer dimers" which reduce the ability of the primers to bind the genetic region of interest, in the assay design. In that situation proxies, SNPs which are highly correlated with the index SNP, were used (r^2=0.9 and above). Proxies were identified using the SNP annotation and proxy search database (SNAP: http://www.broadinstitute.org/mpg/snap/ldsearch.php)

2.2.2 Subjects

DNA samples from 3879 RA cases and 2561 healthy controls from the United Kingdom Rheumatoid Arthritis Genetics Consortium (UKRAG) were included in the replication study. All samples were collected with ethical committee approval (MREC 99/8/84) and all individuals provided informed consent.

2.2.3 Genotyping using the Sequenom MassARRAY Platform

Genotyping using the Sequenom MassARRAY iPlex assays is a 3-stage process to detect SNP polymorphisms in genomic DNA (Figure 13). Primarily, it uses specifically designed primers to amplify DNA surrounding variants of interest using a polymerase chain reaction (PCR). This is followed by an extension reaction which involves insertion of a single base at the site of the polymorphism. Following conditioning to remove contaminants, this can be used for allelic discrimination using MALDI-TOF mass spectrometry analysis. This mass spectrometry method also allows identification of unbound primers and contaminants which may reduce effectiveness of the assay.

Figure 13– Sequenom assay workflow

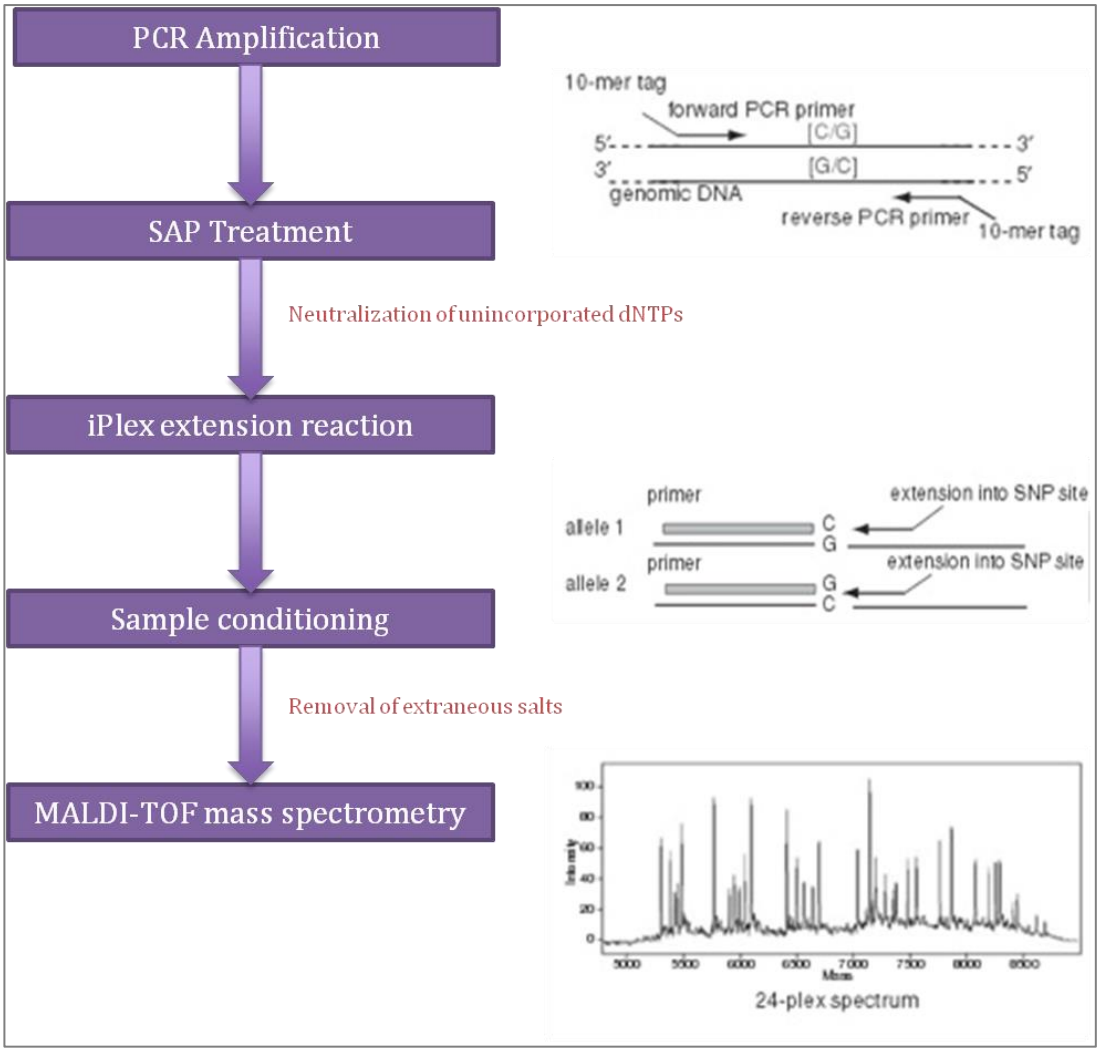


Figure 13 details the workflow for the Sequenom genotyping system. Adapted from Sequenom iPlex gold: Current Protocols in Human Genetics 2.12.1-2.12.18, January 2009.

2.2.3.1 Amplifying DNA for genotyping

Prior to genotyping, 1.2µl of 20ng/µl genomic DNA and several water negative controls were plated into 384-well plates (Applied Biosystems) using the CyBio liquid robot to give ~20ng of DNA in total for each reaction (CyBio, Goeschwitzer Strasse, and Jena, Germany). Firstly DNA around the specific SNPs of interest was amplified by PCR. PCR is a 3 stage-reaction (denaturing, annealing and extension) which allows amplification of specific sequences of DNA. Forward and reverse primer pairs designed as described in section 2.2.1 were used to target specific

sequences of DNA (Metabion, Martinsried, Germany). Each primer was designed with a 10-mer tag attached to the 5' end to allow determination of incorporated primers in the final analysis.

PCR master mixes were generated using volumes shown in Table 4 and 4ul added to wells containing the DNA sample using the Beckman multimeck robot (Beckman Coulter, Fullerton, California, USA). PCR cycling was performed using the 8700 thermo cycler (Applied Biosystems, CA, United States). Details of PCR amplification program are shown in Table 5. As small volumes of reagents were used, an overhang was calculated to prevent loss of reagents due to small pipetting errors.

Table 4– PCR reaction mastermixes

Reagent	Final concentration	Volume (1rxn)	Volume (384 rxn*)
H_20	NA	1.580µl	888µl
PCR buffer (10x) with $MgCl_2$	1.25x	0.625µl	300µl
$MgCl_2$ (25mM)	1.625mM	0.325µl	156µl
dNTP mix (25mM)	500µM	0.100µl	48µl
Primer mix (500nM)	100nM each	1.00µl each	480µl each
Genomic DNA (20ng/µl)	~20ng per reaction	1.00µl	480µl
Taq polymerase enzyme (5U/µl)	0.5U/rxn	0.100µl	48µl
	TOTAL	5.000µl	2400µl

Table 4 details the constituents of the PCR mastermix used in the Sequenom iplex amplification stage. The concentration of each reagents and volume used is shown. *An overhang of 25% was included to account for sampling error. dNTP = deoxyribonucleotide, PCR = polymerase chain reaction.

Table 5 – PCR reaction cycles

Stage	Temperature (ºC)	Time (minutes)
1	94	15
2	94	20
3	56	30
4	72	1
Repeat stage 2 to 4 for 35 cycles		
5	72	3
6	4	Forever

Table 5 shows the PCR reaction cycles used during the Sequenom iplex amplification stage. The temperature and timing of each cycle is shown.

2.2.3.2 Agarose gel electrophoresis

To confirm successful amplification of the DNA, PCR products were run on a 2% agarose gel containing ethidium bromide. 1µl of each sample and negative controls were added and a DNA ladder was run alongside the samples to determine if the product was of the expected size. As ethidium bromide fluoresces under UV light, this was used to visualize gels and determine if any contamination had occurred in the negative controls.

2.2.3.3 SAP Treatment

To dephosphorylate any dNTPs not used up in the PCR reaction, a shrimp alkaline phosphatase (SAP) treatment was performed. This prevents non-specific extension by free dNTPs in the extension reaction and therefore false peaks arising during the mass spectrometry analysis. SAP enzyme solution (Sequenom, Hamburg, Germany) was prepared as described in Table 6, before 2µl was added to all samples. Exact temperatures were required for the SAP reaction; therefore, samples were incubated in the 8700 thermo cycler (Applied Biosystems, CA, United States). Details of incubation are shown in Table 7.

Table 6 – SAP enzyme mastermixes

Reagent	Volume (1rxn)	Volume (*384 rxn)
H_20	1.330µl	638.4µl
10xSAP buffer	0.170µl	81.6µl
SAP enzyme (10U/µl)	0.500µl	240.0µl
Total	2.000µl	960.0µl

Table 6 details the constituents of the SAP mastermix used in the Sequenom SAP treatment. The concentration of each reagents and volume used is shown. *An overhang of 25% was included to account for sampling error. SAP = shrimp alkaline phosphatase.

Table 7 – SAP reaction cycles

Stage	Temperature (ºC)	Time (minutes)
1	37	40
2	85	5
3	4	Forever

Table 7 shows the reaction stages used in the SAP treatment stage. The temperature and timing of each stage is shown.

2.2.3.4 IPlex Reaction

The iPlex reaction involves a single base extension adjacent to the site of the SNP by an iPlex extension primer and a chain terminating dideoxynucleotide triphosphate (ddNTP). This allows discrimination of the mass between alleles which can be detected by mass spectrometry.

It has been shown that mass of a primer is inversely related to a reduction in signal to noise ratios during mass spectrometry. Low signal to noise ratios can make it challenging to call an allele based on mass and can result in inaccurate results. To

account for this, extension primers are split into groups depending on their mass and higher concentrations of primer are added to the high mass group; low mass primers were added at a concentration of 7µm whilst the high mass primers are added at a concentration of 14µm. This increase in concentration of high mass primers allows an increase in the signal to noise ratios and increases the accuracy of the genotyping.

IPlex extension sample master mixes were combined as shown in Table 8 with the extension primers divided into high mass and low mass primers. 2µl of extension master mix was added to each sample well using the CyBio liquid dispenser (CyBio, Goeschwitzer Strasse, and Jena, Germany) and the iPlex reaction performed on the 8700 thermocycler (Applied Biosystems, CA, United States). Details of the iPlex reaction cycles are shown in Table 8.

Table 8– iPlex reaction mastermixes

Reagent	Final concentration	Volume (1rxn)	Volume (384 rxn)
H_20	NA	0.755µl	362.40µl
iPlex Buffer Plus 10x	0.222X	0.200µl	96.00µl
iPlex Termination Mix	1X	0.200µl	96.00µl
Primer mix (7µM low;14µM high)	0.625µM;1.25µM	0.804µl	385.92µl
iPlex Enzyme	1X	0.041µl	19.68µl
	TOTAL	2.000µl	960.00µl

Table 8 details the constituents of the iPlex mastermix used in the iPlex reaction stage. The final concentration of each reagent and volume used is shown.. *An overhang of 25% was included to account for sampling error.

Table 9 – iPlex reaction cycles

Step	Temperature (ºC)	Time (minutes)
1	94	30 seconds
2	94	5 seconds
3	52	5 seconds
Repeat steps 3 to 4 for 4 cycles		
Repeat steps 2 to 4 for 39 cycles		
4	80	5 seconds
5	72	3
6	4	Forever

Table 9 shows the reaction cycles used during the Sequenom iplex reaction stage. The temperature and timing of each cycle is shown.

2.2.3.5 Conditioning the iPlex reaction products — clean resin

In order to optimise mass spectrometry analysis, the reaction products must be desalted using water and clean resin. This works by removing leftover magnesium salts from the iPlex reaction which could potentially interfere with the mass spectrometry reads giving inaccurate genotyping calls.

To desalt the samples, 20µl of water was added to each sample well with 6µg of clean resin (Sequenom). Sample plates were sealed and placed on a rocking platform (VWR) for 30 minutes. Sample plates were then centrifuged at 3200xg for 5 minutes to firmly collect resin in the bottom of the wells.

2.2.3.6 Dispensing sample onto the SpectroChip arrays

Samples were dispensed onto the 384-element SpectroChip bioarray using the MassARRAY Nanodispenser (Sequenom) using the 384 well dispensing program. Multiple washes with 100% ethanol and 0.1M sodium hydroxide (Sigma Aldrich) were performed between chip spotting to condition dispenser pins. Once completely spotted, chips were air dried in the Nanodispenser for 60 seconds and transferred to the MALDI-TOF mass spectrometer for analysis.

2.2.4 Calling SNP genotypes

Assays, plates and acquired spectra were linked using the Plate editor and Assay design suite 1.0 (Sequenom). Once assay and samples were linked, genotyping calls were made using Typer 4.0 genotyping software. Each assay cluster was examined individually to identify any discrepancies in extension rate, peak area and call rate. Assays with high rates of failure (>90% samples failed) were removed from further analysis. Negative control samples were inspected to identify if any sample contamination or non-specific signals were being detected. All genotyping was uploaded to an in-house genotype database (GDB) for export in PLINK v.1.07 format for further analysis.

2.2.5 Sample and SNP QC

SNPs and samples which reached >90% genotyping call rate were included in the analysis. Allele frequencies in controls were assessed to ensure they conformed to Hardy-Weinberg equilibrium. SNPs which deviated from this ($p<0.001$) were removed from further analysis.

2.2.6 Association testing

Allelic association testing was performed using PLINK v.1.07 and association plots visualized using LocusZoom (http://csg.sph.umich.edu/locuszoom/) (Pruim et al. 2010).

2.3 *RUNX1* replication and fine mapping

Unlike many loci on the Immunochip array, variation in the region selected for follow up was not well covered on the chip, therefore further fine mapping was required. Fine mapping is a genetic approach, which involves dense SNP genotyping of a region in order to capture as much variation as possible. This approach can be used to both localize an association signal in a region to a causal variant and identify whether multiple genetic effects exist in the region. Initially the percentage of variants captured in the region by Immunochip alone was calculated before dense SNP genotyping of the region was performed.

2.3.1 Defining the region for fine mapping

The region for fine mapping was defined using recombination rates obtained using Utah residents with Northern and Western European ancestry (CEU) 1000 genomes (July 2010 release) in Haploview v.4.2.

2.3.2. Calculation of coverage for the selected region on the Immunochip array

Coverage of the region on the Immunochip array was assessed using the Tagger function in Haploview v.4.2. All SNPs from the 1000 genomes (July 2010 release) located between recombination hotspots were included in analysis and coverage was based on all common (MAF>0.05) and low frequency (MAF>0.01) SNPs within this region tagged by LD of r^2>0.8 and r^2>0.9.

2.3.3 Subjects

3491 RA cases and 2359 healthy controls from the United Kingdom Rheumatoid Arthritis Genetics Consortium (UKRAG) were included in the study. All samples were collected with ethical committee approval (MREC 99/8/84) and all individuals provided informed consent.

2.3.4 Tag SNP selection and assay design

Tag SNPs from the 1000 genomes Utah residents with Northern and Western European ancestry (CEU/ceph) July 2010 release were selected using the Tagger function in Haploview v.4.2. The coverage of these tag SNPs was then calculated for common (MAF>0.05) and low frequency (MAF>0.01) variants using LD thresholds of r^2=0.8 and r^2=0.9.

Forward/Reverse amplification primers and extension primers were designed using the Sequenom MASS array iPlex assay design suite v.1 (https://www.mysequenom.com/assaydesign) as described in section 2.2.1.

2.3.5 Genotyping using the Sequenom MassARRAY system

Genotyping was performed using the Sequenom MassARRAY system as described in section 2.2.3

2.3.6 Calling SNP genotypes

Assays, plates and acquired spectra were linked using Plate editor and Assay design suite 1.0 as described in section 2.2.4.

2.3.7 Sample and SNP QC

SNPs and samples which reached >90% genotyping call rate were included in the analysis. Allele frequencies in controls were assessed to ensure they conformed to Hardy-Weinberg equilibrium. SNPs which deviated from this ($p<0.001$) were removed from further analysis.

2.3.8 Association testing

Allelic association testing was performed using PLINK v.1.07 and association plots visualized using LocusZoom (http://csg.sph.umich.edu/locuszoom/).

2.3.9 Identification of multiple effects in the selected region

To identify if the region contained multiple genetic effects stepwise logistic regression was performed. All SNPs from the fine mapping were included in this analysis. Conditioning on the most significant, index SNP was performed using the --condition function in PLINK v.1.07. SNPs which remained significant after conditioning ($P<0.0001$) were considered independent. Association plots were visualized using LocusZoom (http://csg.sph.umich.edu/locuszoom/).

2.4 Functional Analysis of the selected region

Once genetic analysis to identify potential causal variants and multiple effects has been performed, it is desirable that investigations are performed to identify the biological function of the gene and its product.

2.4.1 eQTL analysis of the selected region in whole blood

Functional investigations include gene expression and expression quantitative trait loci (eQTL) analysis. An eQTL is single base change in DNA which results in differential expression of a gene. This may occur between SNPs and genes which are proximal to each other (cis-eQTL) or may occur over much larger distances (trans-eQTL).

eQTL analysis involves the genotyping of DNA and gene expression analysis of RNA from identical subjects. The values obtained can then be correlated to indicate whether the change in DNA affects the gene expression. This can be performed in tissue such as whole blood or at a cell specific level. In both cases, these analyses were performed in healthy controls to allow investigation of gene regulation in study cohorts that are not affected by disease status or treatment, which may be the case for IA patients and which could confound results.

Analysis of eQTLs involves generation of both genotyping and gene expression data for the region in the healthy control cohort. Whole genome data was generated using the Human core exome genotyping array (Illumina) and a Taqman allelic discrimination assay (Applied Biosystems) whilst gene expression was quantified using a Taqman gene expression assay (Applied Biosystems).

2.4.1.1 Subjects

75 healthy volunteers from the national repository healthy volunteers (NRHV) study were included in the study. All samples were collected with ethical committee approval (MREC 99/8/84) and all individuals provided informed consent. 10ml of whole blood was collected in Vacutainer Plus tubes containing EDTA (BD) to stabilize the sample at individual sites. Extracted DNA was stored in Manchester at -20°C prior to genotyping using a Taqman allelic discrimination assay (Applied Biosystems). In addition ~5ml of whole blood was collected in Tempus (Applied Biosystems) RNA collection tubes for RNA extraction.

2.4.1.2 SNP genotyping using Taqman allelic discrimination assays

Taqman allelic discrimination assays use primer/probe pairs with specificity to different sequences to determine which allele is present at a particular position in the genome. As the probes are bound to specific fluorescent dyes, this can be detected by the presence of a fluorescent signalling. Matches and mismatches of probes and alleles are shown in Figure 14.

Figure 14 – Taqman allelic discrimination workflow

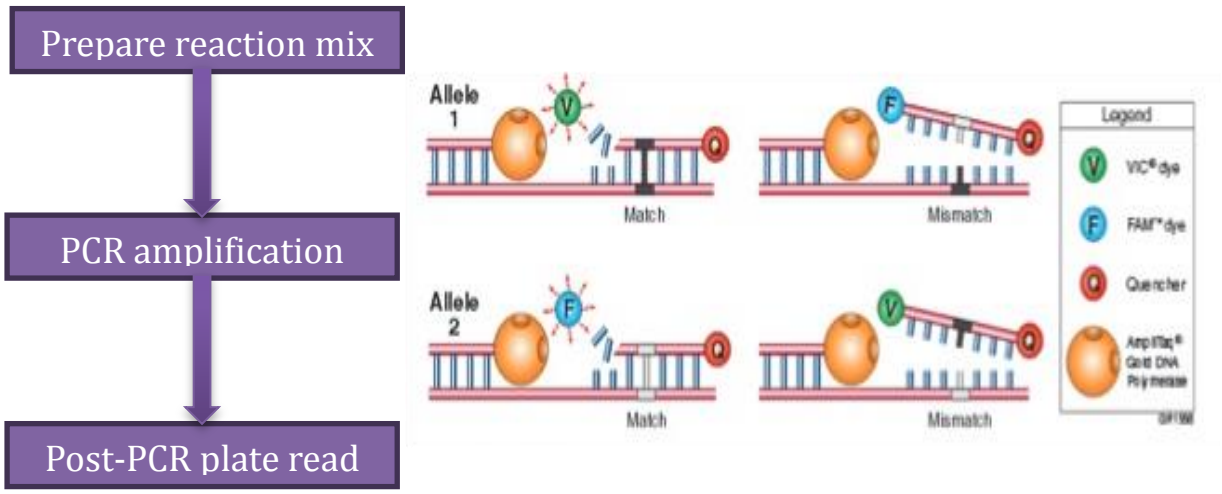


Figure 14 shows an overview of the Taqman genotyping protocol. This process has 3 stages. Adapted from Taqman gene expression assays protocol (Life technologies) (http://tools.lifetechnologies.com/content/sfs/manuals/cms_041280.pdf)

2.4.1.2.1 Extraction of DNA for genotyping

Whole blood DNA was extracted from 10ml blood by phenol chloroform extraction by Joanne Barnes at the Arthritis Research UK Centre for Genetics and Genomics. DNA quantity and quality was not assessed prior to genotyping. Samples were diluted to working dilutions of 20ng/µl for genotyping.

2.4.1.2.2 Plating out of DNA for genotyping

2.25µl of 5ng/µl DNA sample and negative controls (water only) were plated into Micro-amp optical 384-well plates (Invitrogen) using the CyBio liquid robot (CyBio, Goeschwitzer Strasse, and Jena, Germany).

2.4.1.2.3 Preparing the reaction mastermix

Firstly a SNP genotyping mastermix was generated using the reagents listed in Table 10 and thoroughly mixed. 2.75µl of mastermix was added to the sample containing wells and plates were vortexed at 1600rpm for 1 minute using the plate shaker (Illumina). Allelic discrimination was performed using the Quant Studio 12K Flex real time PCR system (Applied Biosystems).

Table 10 – Taqman genotyping reagents

Reagent	Volume (1rxn)	Volume (96 rxn*)
Taqman universal PCR mastermix (2x)	2.5µl	264µl
20x SNP genotyping assay	0.25µl	26.4µl
TOTAL	2.75µl	290.4µl

Table 10 details the constituents of the PCR mastermix used in the Taqman SNP genotyping. The concentration of each reagents and volume used is shown. *An overhang of 25% was included to account for sampling error.

2.4.1.2.4 Allelic discrimination analysis using Quant Studio 12K Flex real time PCR system

Thermal cycling conditions for the 5µl allelic discrimination reaction are listed in Table 11. Genotype calls were performed using the Quantstudio 12k flex real time

PCR software (Applied Biosystems).

Table 11 - Allelic discrimination assay reaction times

Step	Temperature (ºC)	Time
1	95	10 minutes
2	92	15 seconds
3	60	1 minute
Repeat steps 2 to 4 for 39 cycles		
4	4	Forever

Table 11 shows the reaction cycles used during the Sequenom iplex reaction stage. The temperature and timing of each cycle is shown.

2.4.1.3. Calling of genotypes using the Quant studio RT-PCR software

Genotype calls were performed using the QuantStudio 12K Flex Real-Time PCR Software automatic calling algorithm. Negative controls were examined for signal and removed from further analysis.

2.3.1.4 Whole blood gene expression analysis

Gene expression was performed for a selected gene and 2 endogenous controls (GAPDH and ACTNB) using Taqman custom gene expression assays (Applied Biosystems). Endogenous controls are genes which are known to be expressed ubiquitously in cells and therefore can be used to normalize differences in RNA content between samples.

Taqman gene expression analysis involve a 5' nuclease reaction and are composed of specifically designed forward and reverse primers and a probe which is composed of a fluorescent reporter dye, a non-fluorescent quencher and a minor groove binder. Figure 15 shows the chemistry behind this process. The presence of

the quencher means that no fluorescence is observed when probes are free in the reaction. During the PCR reaction, the probe binds at a position between the forward and reverse primer which are bound to a specific target. When this occurs the quencher is cleaved from the reporter dye, resulting in fluorescence. This fluorescence can be quantified with each PCR cycle to determine the levels of gene expression within a sample. To ensure that accurate gene expression readings were obtained, assays for each sample and negative controls were performed in triplicate.

Figure 15 – Taqman gene expression chemistry

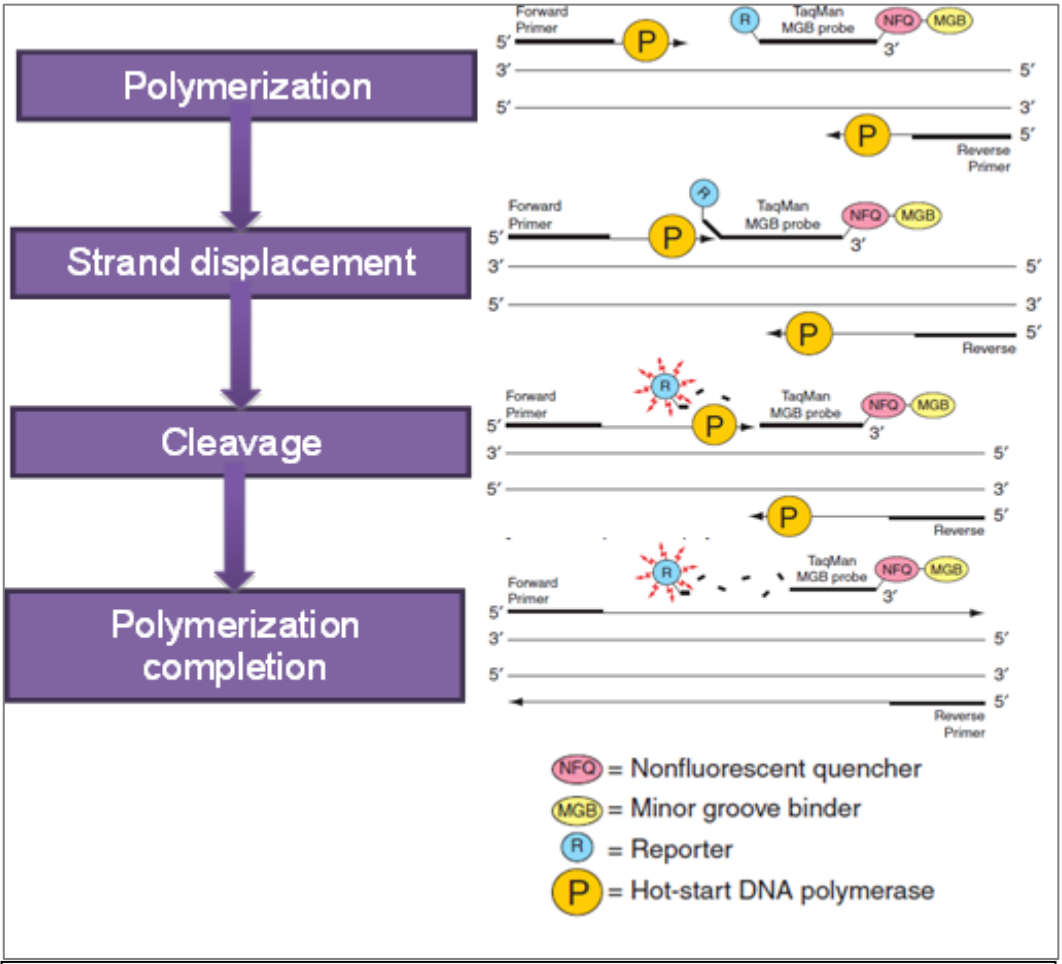


Figure 15 details the chemistry of Taqman gene expression assays. Adapted from Taqman gene expression assays protocol (Life technologies); (http://tools.lifetechnologies.com/content/sfs/manuals/cms_041280.pdf)

2.3.1.5 Design of selected gene and endogenous controls gene expression assays

Information about splice variants were obtained from the ENSEMBL genome browser (ENSEMBL: http://www.ensembl.org/index.html) (Flicek et al. 2011). Taqman gene expression assays were designed using the Taqman assay search tool (Applied Biosystems; http://www.lifetechnologies.com/uk/en/home/life-science/pcr/real-time-pcr/real-time-pcr-assays/taqman-gene-expression.html).

2.3.1.6 Subjects for gene expression analysis

Total RNA was extracted from the 75 subjects using the Tempus spin RNA isolation kit (Life technologies). Prior to extraction; samples were stored at -80°c using Tempus Blood RNA tubes (Applied Biosystems). Details of RNA extraction protocols used can be found in Section 5, Appendix A.

2.3.1.7 Total RNA quality control

Total RNA quantification was performed using the Nanodrop N-100 as described in Section 5, Appendix A. In addition each RNA sample was assessed using the Bioanalyzer 2100 (Agilent), which is an electrophoresis based system used to assess the quality and quantity of RNA. To assess the quality, the RNA is run on a capillary gel and compared to a ladder of standard RNA sizes. Combined with quantification of 18s and 28s ribosomal subunits, this also allows the generation of an RNA integrity number (RIN) which indicates to what extent a sample has been subject to degradation. All reagents were sourced from Agilent unless otherwise stated.

2.3.1.8 RNA quality control using the Agilent Bioanalyzer 2100

Bioanalyzer 2100 (Agilent) uses microfluidics technology to assess RNA quantity and quality inclusively. This is achieved by mixing RNA samples with a sieved polymer and fluorescent dye on a Nanochip which is then placed into the Bioanalyzer instrument. When the chip is run on the Bioanalyzer an electrical current is generated and charged compounds such as RNA migrate, in a similar process to gel electrophoresis. The fluorescent dye intercalates with nucleic acids and therefore can be used to detect the presence of RNA in a sample. This is then compared to a ladder reference, allowing quantification of RNA in a sample. In addition the presence of 28S and 18S ribosomal RNA (rRNA) peaks indicates how intact a sample is and allows calculation of the RNA integrity score (RIN score). The RIN score is a measure of how intact an RNA structure and indicates whether a sample has undergone degradation. Prior to gene expression studies, the RIN score is used as a quality control assessment to ensure all samples are of sufficient quality. Usually a RIN score of greater than 9 is desirable but lower thresholds may be adopted for techniques less sensitive to RNA quality. In addition a strong peak at 28S and 18S rRNA with no peaks in between indicates a good quality RNA sample. All samples were assessed using these parameters.

2.3.1.8.1 Preparing the gel and gel dye mix

After equilibration to room temperature, 550µl of RNA 6000 Nano gel matrix (Agilent) was transferred into a spin filter and centrifuged at 1500xg for 10 minutes. The gel was then separated into 65µl aliquots for use with individual Nanochips. 1µl of RNA 6000 Nano dye concentrate was then added to 65µl of filtered gel and vortexed thoroughly using the Bioanalyzer chip vortexer (IKA) to ensure complete mixing. Gel-dye mix was then centrifuged at 13000xg for 10 minutes.

2.3.1.8.2 Loading the gel dye mix

The RNA Nanochip was placed in the chip priming station (Agilent) and 9µl gel-dye mix added to corresponding reservoir in chip. The chip priming plunger was then pressed and held for 30 seconds until the gel was dispersed across the chip. 9µl gel-dye mix was then added to an additional 2 wells.

2.3.1.8.3 Loading the Nanomarker, RNA ladder and samples onto the chip

5µl of RNA 6000 Nano marker was added to the corresponding ladder well and each individual sample well. 1µl RNA ladder (Agilent) was then added to the corresponding well whilst samples were heat denatured at 70^{o}C for 2 minutes using a 1.5ml Thermomixer R heat block (Sigma Aldrich). 1µl samples were then loaded into corresponding wells. The chip was vortexed at 2000rpm for 60 seconds.

2.3.1.8.4 Running the Nanochip

The chip was inserted into the bioanalyzer instrument and analysis performed using 2100 expert software (Agilent). Both gel and electropherogram results were visually inspected to ensure RNA samples were pure and concentrated enough for further analysis. In the gel section it is expected that 2 clear bands are visible with no sign of degradation. In the electropherogram, two peaks are expected at 18s and 28s. Presence of any additional peaks may indicate sample contamination by DNA or proteins.

2.3.1.9 cDNA conversion using High-capacity cDNA Reverse Transcription Kit

Whole blood total RNA (RIN value >5) was normalized to 50ng/µl for cDNA conversion by the High-capacity cDNA Reverse Transcription Kit (Applied Biosystems). 10µl normalized RNA was added to a 96 well PCR plate (Starlab). 2X reverse transcription (RT) mastermix was prepared on ice using the volumes in

Table 11 and 10µl RT mastermix was added to each well containing sample. No template (H_20 only) and no RT controls (all reagents except RT enzyme) were added as negative controls. Plates were subjected to thermo cycling conditions listed in Table 13.

Table 12– cDNA conversion volumes

Reagent	Volume (1rxn)	Volume (48rxn*)
10xRT buffer	2µl	110.4µl
25xdNTP Mix (100 mM)	0.8µl	44.16µl
10x Random Primers	2µl	110.4µl
Multiscribe Reverse Transcription	1µl	55.2µl
Nuclease Free H_20	4.2µl	231.84µl
TOTAL	10µl	552µl

Table 12 details the constituents of the mastermix used in the cDNA conversion. The concentration of each reagents and volume used is shown. *An overhang of 25% was included to account for sampling error.

Table 13– cDNA conversion thermo-cycling

Step	Temperature (ºC)	Time
1	25	10 minutes
2	37	120 minutes
3	85	5 minutes
4	4	Forever

Table 13 shows the reaction cycles used during cDNA conversion. The temperature and timing of each cycle is shown.

2.4.1.10 Gene expression analysis of selected gene and endogenous controls

4µl of 40ng/µl cDNA was added in triplicate to the wells of a 384 well optical PCR plate (Applied Biosystems). PCR reaction master mix was prepared on ice

according the volumes in Table 14 and 16µl mastermix was added to each well containing sample. No template (H_20 only) and no RT controls (all reagents except RT enzyme) from the cDNA reaction were added as negative controls whilst 4µl genomic DNA was added to ensure that the assay did not bind non-specifically to genomic DNA. Plates were placed in the Quant studio 12k flex real time PCR system and were subjected to thermocycling conditions listed in Table 15. Once the adequate number of reaction cycles was complete, ΔCT values were obtained using the Quantstudio 12k flex real time PCR software (Applied Biosystems).

Table 14–Gene expression reaction mastermix

Reagent	Volume (1rxn)	Volume (144*rxns)
20x Taqman gene expression assay	1µl	170.2µl
2x Taqman gene expression mastermix	10µl	1702µl
cDNA template (40ng/ul)	4µl	680.8µl
Nuclease free H_20	5µl	851µl
TOTAL	20µl	3404µl

Table 14 details the constituents of the mastermix used in the gene expression reaction stage. The concentration of each reagents and volume used is shown. *An overhang of 25% was included to account for sampling error.

Table 15- Gene expression reaction thermo cycling conditions

Step	Temperature (ºC)	Time
1	50	2 minutes
2	95	10 minutes
3	95	15 seconds
4	60	1 minute
Repeat steps 3 to 4 for 39 cycles		
5	4	Forever

Table 15 shows the reaction cycles used during gene expression reaction stage. The temperature and timing of each cycle is shown.

2.4.1.11 Whole blood eQTL analysis

To identify if associated variants represented an eQTL with a gene in whole blood, linear regression was performed in STATA v.11.2 using genotype at the associated SNP as a covariate in correlation with selected gene expression normalised to *GAPDH* and *ACTNB* endogenous controls.

2.4.2 eQTL analysis of the selected gene in T lymphocytes

Although whole blood gene expression analysis is a good tool for identifying eQTLs across the genome, it represents an average of the differential gene expression of the many heterogeneous types of cells which make up peripheral blood. In some cases this may result in cell specific eQTLs being masked by signals from stronger and more abundant cell types in a sample. In the case of RUNX1 eQTLs, these may only be present in a particular subset of cells. Cell specific eQTL analysis involves using whole genome wide genotyping and transcription generated from a single homogenous cell population to identify cell specific regulation of genes across the genome.

T lymphocytes were chosen as the cell of interest in this project as they have been shown to be important cells in the pathogenesis of IA and therefore represent a good candidate cell type for cell specific eQTL analysis (Cope 2008). Furthermore *RUNX1* is a crucial mediator of T cell lineage commitment in CD8+ cells (Taniuchi et al. 2002) and CD4+ skewing towards Th1 and Th17 characteristics (Komine et al. 2003;Lazarevic et al. 2011) making this a good candidate cell to find the consequence of a polymorphism in the *RUNX1* region.

By using a variety of genetic and immunological techniques, genome wide genotyping and cell specific whole transcription data were generated from the major T lymphocyte subsets, CD4+ and CD8+. This was performed in healthy controls, giving a representation of gene regulation across the genome in these highly important cell types. During this project whole genome eQTL data was generated for CD4+ and CD8+ lymphocytes but only data from the *RUNX1* region will be presented here.

2.4.2.1 Subjects

23 healthy volunteers from the national repository healthy volunteers (NRHV) study were included in the study. All samples were collected with ethical committee approval (MREC 99/8/84) and all individuals provided informed consent. 20mls peripheral whole blood was collected in Vacutainer Plus tubes containing EDTA (BD) to stabilize the sample by colleagues within the Arthritis Research UK Centre for Genetics and Genomics. A breakdown of the sample cohort information is given in Table 41.

2.4.2.2 Genotyping of samples

As the subjects used for this project were a subset of the healthy controls used previously, genotyping data for the associated SNP was obtained from the dataset generated in section 2.4.1.

2.4.2.3 Sample collection for PBMC extraction

Within 2 hours of collection 20ml peripheral whole blood samples were diluted 1:1.5 with 30mls ambient MACS running buffer (Miltenyi Biotec). MACS running buffer is a combination of phosphate buffer saline (PBS), bovine serum albumin (BSA), EDTA and 0.09% sodium azide giving a 7.2pH balanced solution. 25mls diluted peripheral blood was then layered on 15mls ambient Ficoll paque plus (GE healthcare) in sterile 50 ml Falcon tubes (Fisher scientific). Samples were centrifuged at 20°C at 400xg for 30 minutes with no brake and the peripheral blood mononuclear cell (PBMC) layer shown in Figure 16 was extracted using a sterile Pasteur pipette (Fisher scientific). MACS running buffer was added to the cell suspension to bring the sample volume to 40mls and centrifuged for 10 minutes at 300xg to wash the cell pellet. The supernatant was discarded and the cell pellet re-suspended in 10mls MACS running buffer. 10µl sample was removed and diluted 1:10 using MACS running buffer for cell counting. In addition 5µl was removed for a viability check. 30mls MACS running buffer was added and a final wash performed by centrifuging at 250xg for 15 minutes. A cell count was then performed using the Casy cell counter (Roche) whilst cell viability was then assessed using trypan blue (Life technologies).

Figure 16– Ficoll separation layers

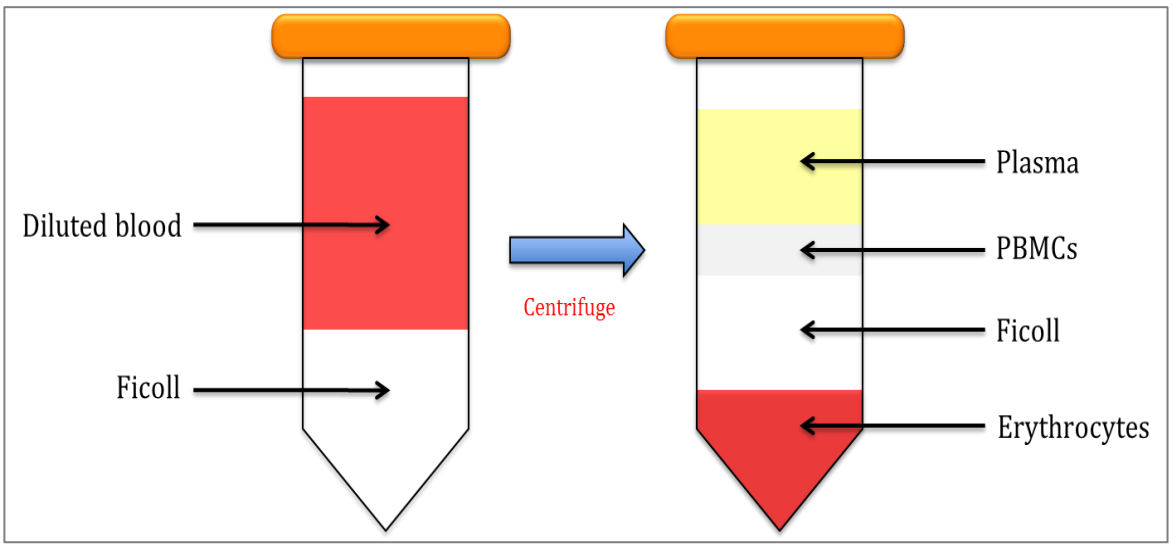


Figure 16 shows the separation of whole blood layers during density grade centrifugation. In this study the layer labelled PBMCs was extracted for further analysis.

2.4.2.4 Cell counts and viability checks

To determine the characteristics of the PBMC population obtained by Ficoll extraction, a full cell count and viability check was performed for each sample. This allows estimation of the total yield of cells obtained and whether they are suitably viable for further analysis. A high number of dead cells may be indicative of inadequate sample collection or poor sample handling which must be identified before further analysis. Both Casy cell counting and trypan blue exclusion were used to determine cell numbers and viability in each sample.

2.4.2.4.1 The CASY Model TT cell counter

Casy cell counting technology uses electrical current exclusion to determine the number and size of cells in a sample. This is achieved by suspending cells in isotonic buffer and passing each sample through a measuring pore which houses a low voltage electrical field. As the electrical current is passed through each sample, the integrity of the cell membrane determines the strength of the electrical signal which passes through to the other side. This is read and translated by the Casy cell counter to determine the number and size of cells present in the sample. Figure 17 shows an overview of the current exclusion process.

2.4.2.4.1.1 Determining cell counts using the CASY cell counter

50µl of 1:10 cell suspension was added to 10ml Casy Ton (Gibco) in a Casy Cup (Gibco) for cell counting. Parameters were set at 6.5µm– 50µmfor counting of mammalian PBMCs. 180µl was sampled for each cell count and an average taken over 3 counts. Counts were given in cells/ml which was multiplied x100 to give the final volume in each sample.

Figure 17– CASY cell counter current exclusion

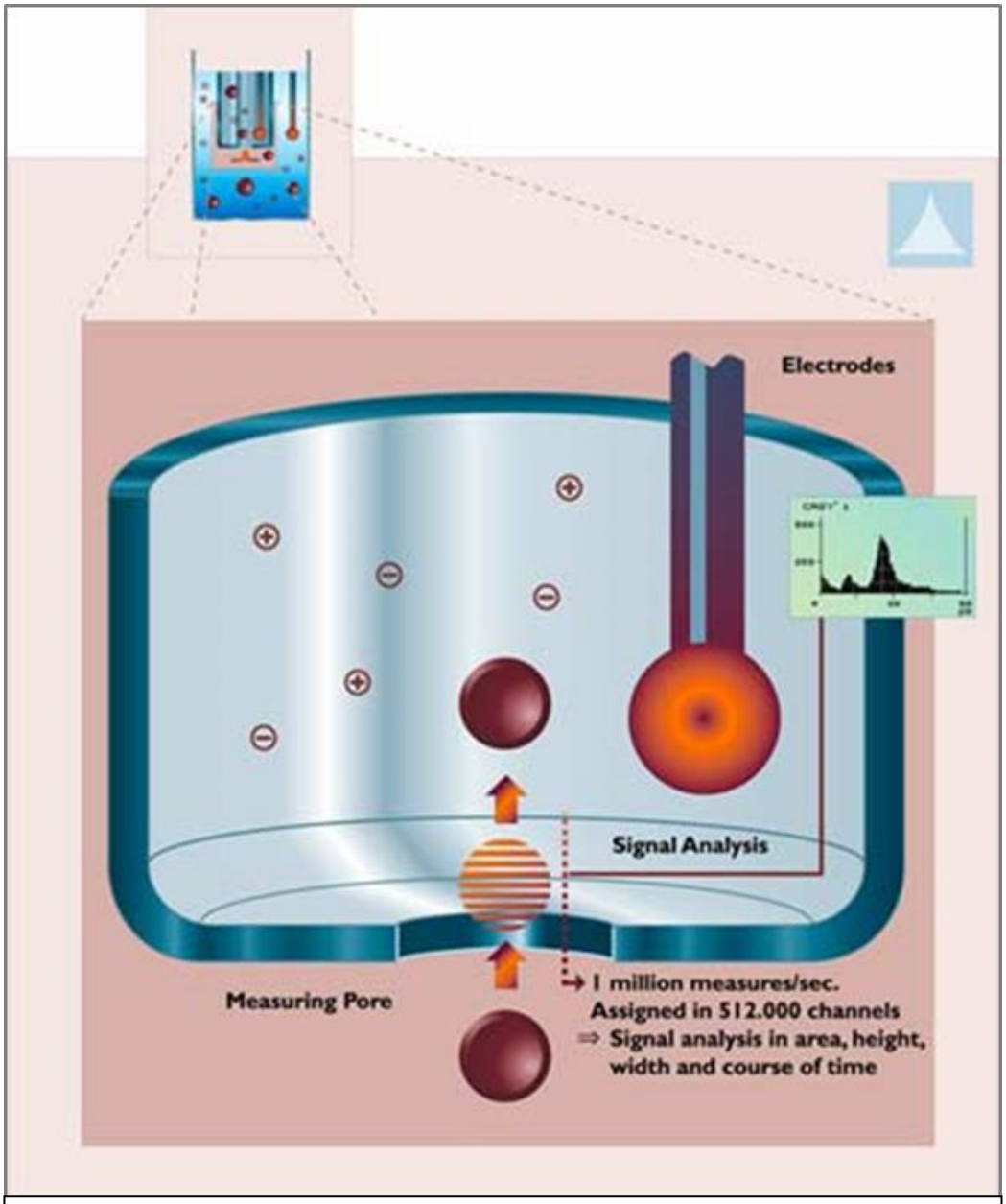


Figure 17 shows how the CASY cell counter uses electrical current exclusion to assess the number and integrity of cells in a sample. This is performed by passing an electrical current (shown as positive and negative charges) through a sample and assessing which signals pass through the cell. Adapted from CASY cell counter protocol (Roche): http://lifescience.roche.com/shop/products/cell-counter-and-analyzer.

2.4.2.4.2 Trypan blue exclusion

Trypan blue exclusion involves staining of cells with a diazo dye and examining the cells under a phase contrast microscope. If a cell membrane is intact the dye is excluded and a cell is considered viable. If the dye is taken up and the cell is visualized as blue then a disruption of the cell membrane has occurred which indicates that the cell is not viable.

2.4.2.4.2.1 Assessing viability using Trypan Blue

For the viability check, 5µl cell suspension was diluted 1:1 with trypan blue (Life technologies) and visualized using a haemocytometer (Burker) under a phase contrast microscope (Olympus). Sample viability was determined using the equation below (Figure 18). If samples exhibited less than 95% viability they were excluded from further analysis.

Figure 18– Cell viability equation

$$\textbf{Percentage of viable cells} = \frac{\textbf{Number of dead cells}}{\textbf{Total number of cells}} \times 100$$

Figure 18 shows the equation used to calculate the viability of each sample.

3.4.2.5 Cryopreservation and thawing of PBMCs

To allow feasible collection of the large number of samples included in this study, extracted PBMCs were cryopreserved. This involves storage of cells at sub-zero temperatures which prevents enzymatic and chemical activity that may cause

damage to cells stored at higher temperatures. The cells can then be thawed when required for cell separation.

3.4.2.5.1 Freezing of PBMC samples

After assessment of yield and viability, samples were re-suspended in 1ml Recovery cell culture freezing medium (Gibco) per $5x10^6$ cells and transferred to 2ml Nunc Cryovials (Sigma Aldrich). Recovery cell culture freezing medium contains PSB, dimethyl sulfoxide (DMSO) and Foetal bovine serum (FBS) which are important for maintaining cell viability during cryopreservation. DMSO is a solvent which dissolves water and therefore prevents the formation of water crystals, which may damage cell membranes during cryopreservation. FBS is a culture supplement which contains growth factors important for cell survival and growth. Samples were then transferred to a Mr. Frosty freezing pot (Nalgene) for gradual freezing at -80°c for 24 hours. The Mr. Frosty freezer pot contains isopropyl alcohol and therefore allows gradual freezing at a cooling rate of -1^0C/min. This rate of cooling minimizes the risk of inducing cell death by drastic temperature change. Samples were then transferred to liquid nitrogen vapour phase (~-150°C). All samples were stored in liquid nitrogen for a maximum of 3 weeks before removal for thawing.

3.4.2.5.2 Thawing of PBMC samples

Thawing was performed by warming vials rapidly at 37^oC in a Hybex multisample incubation system (Scigene) and gently transferring the cell suspension using a sterile pastette to a 50ml Falcon tube containing 30ml pre-warmed (37^oC) MACS running buffer. Pre-warmed (37^oC) MACS buffer was gradually added until the volume reached 40mls. Samples were centrifuged at 1600xg for 7 minutes and the supernatant aspirated to remove dead cells and DMSO. 20ml MACS buffer was then added and both a cell count and viability check performed to determine post-cryopreservation cell numbers and viability.

2.4.2.6 Separation of PBMCs into T lymphocyte subsets

To obtain pure T lymphocyte subsets from each PBMC sample, a series of Miltenyi MACS cell separations were performed for each of the 23 samples. MACS technology involves using magnetic microbeads, conjugated to antibodies targeting cell markers, to isolate individual cell populations from mixed cell populations. This can be achieved by performing either positive or negative selection separation strategies.

2.4.2.6.1 Positive selection

Positive selection involves using magnetic microbeads, conjugated to an antibody targeting a marker expressed on the cell of interest. When incubated with a sample, these antibodies specifically bind each cell expressing that marker and magnetically extract the cells. Cells which do not express this marker remain unlabelled. The sample can then be passed through a separation column containing a magnetic field which retains the treated cells whilst all unlabelled cells pass through. Bound cells can then be eluted from the column as a positive fraction, whilst the unlabelled cells are collected as a negative fraction.

This strategy allows isolation of a highly pure population of cells as the technique directly targets the cell marker of interest. It also allows multiple separations from the same sample as the negative fraction obtained can be extracted with beads targeting a different marker.

2.4.2.6.2 Negative selection

Negative selection involves using magnetic beads, conjugated with a cocktail of antibodies against a number of cell markers not expressed on the cell of interest. When incubated with samples, the antibodies magnetically label all other cells than

that which is being targeted. When passed through a column, these cells are retained whilst the cells of interest pass through, providing a pure cell population.

This strategy allows isolation of an individual cell population, without binding the cell markers of interest directly. This is important if the marker is used in activation of the cell, which may result in aberrant stimulation of the cells.

2.4.2.6.3 Selecting a strategy for T lymphocyte isolation

As the aim of this study was to isolate highly pure populations of cells for gene expression analysis; a positive selection strategy was employed using CD4 Human microbeads and CD8 Human microbeads (Miltenyi). In addition a CD14+ selection was performed before the CD4+ selection using CD14 Human microbeads, to remove CD14+ monocytes which may express low levels of CD4 marker. This ensures that a pure CD4+ population was obtained. As the negative fractions obtained were not extracted with microbeads, this technique also allowed isolation of both CD4+ T helper and CD8+ cytotoxic T cells from each individual PBMC sample.

2.4.2.6.4 Separation of PBMCs using Miltenyi MACS

To prepare for microbead treatment, samples were centrifuged at 300xg for 10 minutes at 4°C and the supernatant aspirated. The cell pellet was then resuspended in 160µl of chilled MACS running buffer, transferred to a 5ml test tube (BD) and placed in a MACS chill rack which had been previously chilled at 4°C overnight (Miltenyi). The chill rack was then placed on the Automacs pro separator (Miltenyi) and the "autolabelling," program selected. CD8+ Human microbeads (Miltenyi) were then scanned using the 2D code reader, the "CD8+ posseld2" program selected and cell separation run started. Autolabelling involves complete automation addition of reagents, incubation and separation of cell subsets. Posseld2 is a double column program which passes the cell suspension through the magnetic field twice and is used when highly pure cell populations are required. Once cell separation was complete 50µl of the positive fraction (containing CD8+ cells) was removed for cell counting and the negative fraction (containing

remaining PBMCs) centrifuged at 300xg for 10 minutes at 4°C and the supernatant aspirated. This process was then repeated for CD14+ and CD4+ positive selections, allowing harvesting of multiple cell types from an individual PBMC sample.

2.4.2.7 Assessment of cell viability and purity using flow cytometry

Although MACS separation is a reliable method for isolating pure cell populations, flow cytometry was used to determine the purity of the isolated CD4+ and CD8+ cells. Flow cytometry is a laser-based technique that involves passing a sample containing cells in a fluid stream through an electronic detection system. This detection system directs a light laser through the stream, which can then be directly analysed by a number of detectors sensitive to different types of scattered and fluorescent light (Fluorescent channels; FL). The scattered light is generated by the diffraction caused by the presence of cells in the sample whilst the fluorescence is generated by specific flourochromes, which are used to detect particular markers on cells. This is achieved by treating samples with antibodies targeting specific cell markers conjugated to flourochromes. The use of different flourochromes, emitting different wavelengths of fluorescent light allows staining of cells with multiple antibodies targeting different markers. This allows extensive profiling of the composition of cells in a sample. In addition the data collected from the light scatter allows the size and granularity of cells in a sample to be determined which can indicate the type and physical state of cells present in the sample.

2.4.2.7.1 Using flow cytometry to analyse T lymphocyte subsets

In order to determine the purity of each CD4+ and CD8+ sample, a number of anti-human antibody-flourochrome conjugates were used to label each sample for flow cytometry analysis (Table 16). Initially for the CD8+ samples an anti-CD8-Allophycocyanin (anti-CD8-APC) antibody was used for single staining whilst the CD4+ samples were double stained with an anti-CD3-APC and anti-CD4-R-Phycoerythrin (anti-CD4-PE) antibody. For each antibody, a matched anti-mouse isotype control antibody was also included to compensate for any non-specific antibody binding, which may occur (Table 16). As an additional quality control

measure, a fixation and dead cell discrimination kit (Miltenyi) was used to determine the viability of the cells in each sample. This could then be used to include only live cells in the purity analysis, as dead cells may non-specifically bind the antibody and therefore generate false positive results. As with trypan blue exclusion (section 2.4.2.4.2.1), the dead cell stain works on the principal that viable cells with intact cell membranes will exclude the dye whilst cells with a compromised cell membrane will take up the dye. As this particular dye emits fluorescence at 488nm-625nm, cells that take up the dye can be detected using flow cytometry.

2.4.2.7.2 Staining of cells for flow cytometry

100µl of $1x10^6$/ml of each of the 46 separated samples was added to a 96 well round bottom plate (Corning) for flow cytometry analysis. 2µl dead cell discriminator (Miltenyi) was added to each sample and incubated on ice under a 60w light source for 10 minutes. 10µl antibody solution (Miltenyi) was added to each sample as described in Table 16 and samples incubated at 4^oC in the dark for 15 minutes. Samples were then washed with 1ml MACS running buffer (Miltenyi) and centrifuged at 300xg for 5 minutes. This was repeated to ensure complete removal of unbound antibodies and the cell pellet resuspended in 300µl running buffer, 150µl fix solution and 5µl discriminator stop reagent (Miltenyi). Fixed samples were stored at 4°C overnight in the dark and transferred to the 5ml test tubes (BD) for flow cytometry analysis.

Table 16 - Antibody cocktails added to first group of samples (n=6)

Antibody	Flourochrome conjugate	Clone	Dilution	Fluorescence channel
Anti-human CD8	APC	BW-135/80 Mouse IgG2a	1:11	FL8
Anti-human CD3	APC	BW-264/56 Mouse IgG2a	1:11	FL8
Anti-human CD4	PE	M-T466 Mouse IgG1	1:11	FL2
Mouse IgG2a isotype	APC	S43.10 Mouse IgG2a	1:11	FL8
Mouse IgG1 isotype	PE	IS5-21F5 Mouse IgG1	1:11	FL2

Table 16 includes details of the antibody staining cocktails used for the initial analysis of 6 samples including Antibody target; Flourochrome conjugate used; clone of cells used to produce antibody; the dilution used to stain cells; Fluorescence channel corresponding to the flourochrome used. Abbreviations: APC = Allophycocyanin, PE = R-Phycoerythrin FL = fluorescence channel.

2.4.2.7.3 Flow cytometry analysis

Flow cytometry was performed using the Cyan ADP flow cytometer (Beckman Coulter), which has 9 fluorescence detection channels (FL). All flow cytometry capture and gating was performed by Michael Jackson at the University of Manchester FLS core flow facility. Samples were analysed using the Cyan Hypercyte data collection program (Beckman Coulter) and plots exported.

Viability of CD8+ and CD4+ samples was firstly assessed using un-gated data (Figure 50; R1 = CD8+; R4 = CD4). Viability was expressed as the percentage of cells, which did not take up the dead cell exclusion dye, compared to the total number of cells analysed. To assess purity in the separated CD4+ and CD8+ cell

populations, gating was performed on live cells that did not take up the dead cell exclusion dye (R1; Figure 19). Forward vs. side scatter was then examined and gating performed to include intact cells and not cell debris, which exhibits a distinctively high forward scatter (R3; Figure 19). Pulse width vs. forward scatter was then examined and gating performed on single cells, which did not appear as cell clumps with high pulse width (R2; Figure 19). This increased the likelihood that the cells being analysed were viable and results accurate. The purity of each CD8+ (R4; Figure 55) and CD3+CD4+ (R7; Figure 56 A-B) lymphocyte population was then examined. Purity was determined by the percentage of positive cells compared to the total cells analysed. All remaining sample was then suspended in 1ml Trizol reagent (Ambion) and stored at -80°C for Trizol-chloroform total RNA extraction.

Figure 19– Gating for analysis of purity of CD4+ and CD8+ samples

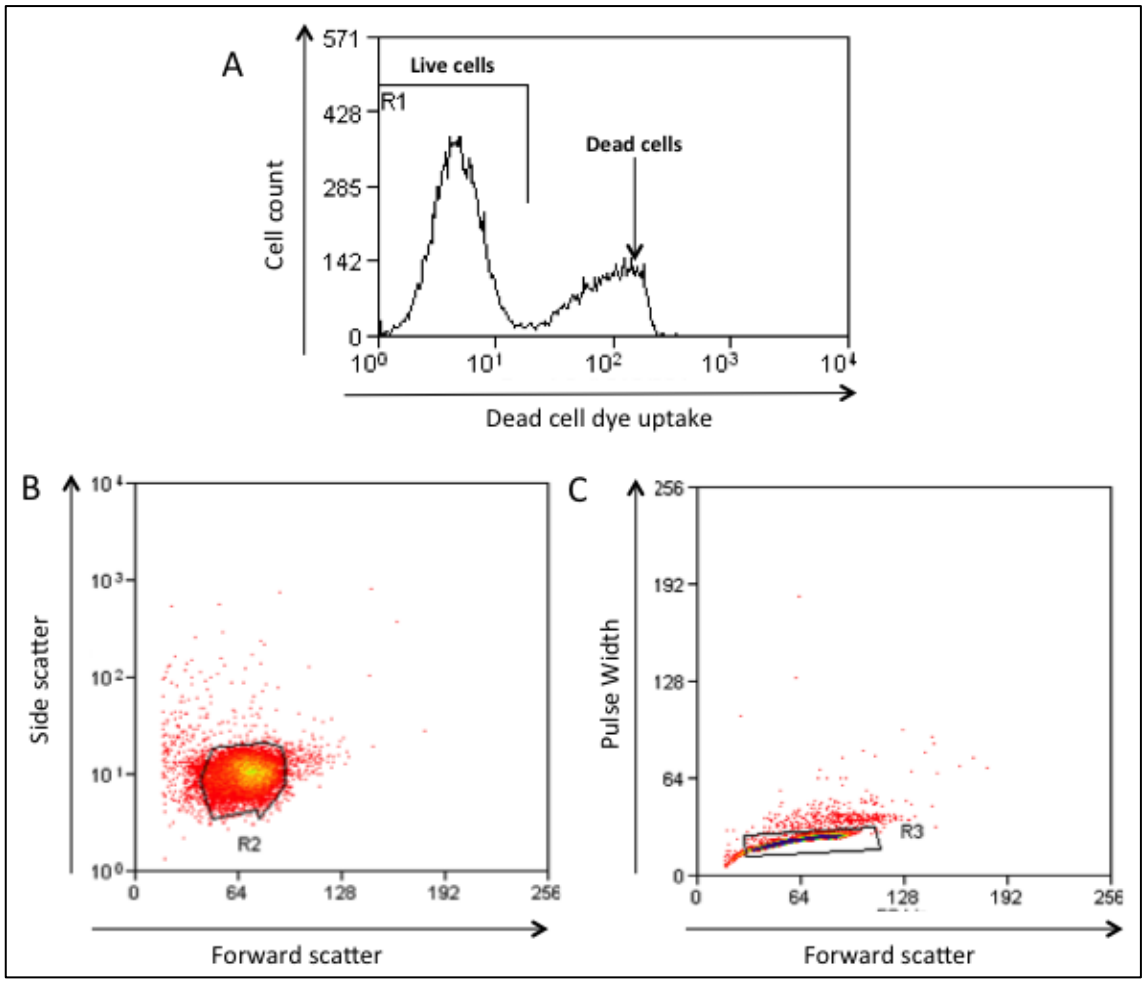


Figure 19 shows the 3 phase gating strategy used in the analysis of all CD8+ and CD4+ samples prior to analysis of purity A) Histogram of dead cell discrimination dye uptake showing cell count and uptake of dead cell dye. Live cells which did not take up the dead cell stain (negative left peak) were gated on (R1) to exclude dead cells (right peak). B) Dot plot of the side scatter and forward scatter (SS and FS) representing size and granularity of cells. Gating was performed to exclude cells with noticeably high values of SS or FS, which indicate cell debris (R2) C) Dot plot of the pulse width against forward scatter. Gating was performed to exclude cells with high pulse width, which indicate clumps of cells (R3). Further analysis was performed on cells within these gates only.

2.4.2.7.3.1 Issues with fluorescence overlap

Once the analysis of the first set of 6 CD4+ samples was performed it became apparent that the dead cell discrimination dye, which uses channel FL3 channel on the flow cytometer was leaking into the nearby FL2 channel, which is used to detect the PE flourochrome. As CD4-PE was used to stain the CD4+ cells, compensation had to be performed to discriminate the PE signal from the dead cell stain signal. For the dead cell dye overlap in to the PE detector the compensation was 93.3% and for the PE overlap in to the dead cell marker detector the compensation was 22.6%. Once this issue was identified, a new anti-CD4 antibody conjugated to the flourochrome Vioblue (Miltenyi) was used to stain the 17 subsequent CD4+ samples. All other antibody-flourochrome conjugates were not changed (Table 17). The Vioblue flourochrome uses the FL6 channel on this flow cytometer meaning the dead cell discrimination dye could not interfere and therefore no compensation was required for the remaining samples. Samples were analysed, gating performed and purity examined as described in the section 2.4.2.7.3 (R11; Figure 19 C-D). All remaining sample was then suspended in 1ml Trizol reagent (Ambion) and stored at -80°C for Trizol-chloroform total RNA extraction.

Table 17 - Antibody cocktails added to remaining group of samples (n=17)

Antibody	Flourochrome conjugate	Clone	Dilution	Fluorescence channel
Anti-human CD8	APC	BW-135/80 Mouse IgG2a	1:11	FL8
Anti-human CD3	APC	BW-264/56 Mouse IgG2a	1:11	FL8
Anti-human CD4	Vioblue	VIT4 Mouse IgG2a	1:11	FL6
Mouse IgG2a isotype	APC	S43.10 Mouse IgG2a	1:11	FL8
Mouse IgG2a isotype	Vioblue	S43.10 Mouse IgG2a	1:11	FL6

Table 17 details of the antibody staining cocktails used for the initial analysis of 6 samples including Antibody target; Flourochrome conjugate used; clone of cells used to produce antibody; the dilution used to stain cells; Fluorescence channel corresponding to the flourochrome used. Abbreviations: APC = Allophycocyanin, PE = R-Phycoerythrin FL = fluorescence channel.

2.4.2.8 Extracting total RNA from cell subset suspensions

To extract total RNA from each cell suspension, cell lysates stored in Trizol at -80^{o}C. (Ambion) were defrosted, transferred to 2ml heavy phase lock gel tubes (5 prime) and incubated at room temperature for 5 minutes. 200µl 1-Bromo-3-chloropropane (BCP; Sigma Aldrich) was added and samples shaken vigorously for 15 seconds before centrifuging at 12,000xg for 10 minutes at 4^{o}C to separate sample into distinct organic and aqueous layers. The aqueous phase was removed carefully and transferred to a clean 1.5ml eppendorf (Starlab). RNA was

precipitated by adding 500µl isopropyl alcohol (Sigma Aldrich), 2µl glycoblue (Life Technologies) and incubating at room temperature for 30 minutes. During this time, the sample was carefully inverted to aid precipitation before centrifuging at 12,000xg for 10 minutes at 4°C to pellet precipitated RNA. The supernatant was aspirated and the cell pellet washed twice in 1ml ice-cold ethanol. This was achieved by adding 1ml 75% ethanol, vortexing briefly and centrifuging at 7,500xg for 5 minutes at 4°C. All supernatant was removed carefully and the pellet allowed to dry at room temperature for 10 minutes. Each pellet was then dissolved in 40µl RNase free water and RNA concentration determined using the Nanodrop N-1000.

2.4.2.9 RNA quality control

2.4.2.9.1 RNA quality control using the Agilent bioanalyzer 2100

Assessment of RNA quality and concentration was performed using the Nanodrop-1000 (Fisher scientific) as described in Section 5, Appendix A.

Total RNA concentration and RNA integrity number (RIN) values for each sample were determined using the Bioanalyzer 2100 gel system (Agilent) as described in section 2.3.18. All samples were normalized to 36.4ng/µl using RNase free water (Gibco) and 11µl aliquoted into 0.2ml Eppendorf's (Starlabs) for Illumina total prep RNA amplification.

2.4.2.10 DNase treatment of total RNA

RNA samples were normalized to a 43µl volume by adding 3µl RNase free water (Gibco) and vortexing briefly. 5µl DNase I reaction buffer and 2µl DNase enzyme (Life technologies) were added and each sample incubated at 37°C for 30 minutes using an Eppendorf 1.5ml Thermomixer (Sigma Aldrich). 50µl RNase free water was added, each sample transferred to a 2ml phase lock gel tube (5 prime) and tubes incubated for 5 minutes at room temperature. To inactivate the DNase treatment 100µl acid phenol: chloroform with Isoamyl alcohol (pH 4.5; 25:24:1;

Sigma Aldrich) was added and shaken vigorously for 15 seconds to form an emulsion. Samples were centrifuged at 12,000xg for 10 minutes at 4°C and the aqueous phase transferred to a clean 1.5ml treatment (Starlabs). RNA was precipitated by adding 330µl 100% ethanol (Sigma Aldrich), 10µl 3M-ammonium acetate (Ambion), 1µl glycoblue (Life technologies) and incubating on ice for 30 minutes. Samples were centrifuged at 12,000xg for 20 minutes at 4°C and the supernatant aspirated. The RNA pellet was then washed twice by adding 1ml ice-cold 75% ethanol (Sigma Aldrich/Gibco) and centrifuging at 12,000xg for 5 minutes. All supernatant was then removed carefully and pellets allowed to air dry for 10 minutes before resuspending in 25µl RNase free water (Gibco).

2.4.2.11 RNA amplification using Illumina TotalPrep Amplification Kit

In order to have sufficiently large volumes of RNA for gene expression studies, RNA amplification is performed. The Illumina TotalPrep Amplification Kit (Ambion) is a series of reactions to synthesise large volumes of cRNA which are suitable for hybridization with Illumina whole transcription arrays (Figure 20). Initially the isolated RNA is reverse transcribed using a high yield reverse transcription enzyme and a T7 oligo (dT) primer to synthesise complementary DNA (cDNA) containing a T7 promoter sequence. This cDNA is then transcribed to a double stranded DNA (dsDNA) template using DNA polymerase and RNase H enzymes. This reaction allows the simultaneous generation of complementary second strand cDNA and degradation of template RNA. The cDNA template is then enzymatically cleaned to remove RNA, primers, salt and enzyme carryover from the previous reactions, which may inhibit transcription. Finally using the cDNA as a template, multiple copies of biotinylated complementary RNA (cRNA) are synthesized. The cRNA is then purified to remove unincorporated deoxynucleotide triphosphates (dNTPs), enzymes and salts from the synthesis reaction. The purified cRNA can then be hybridized to the Illumina HT-12v4 expression chip. All reagents were sourced from Ambion unless otherwise stated.

Figure 20– TotalPrep amplification workflow

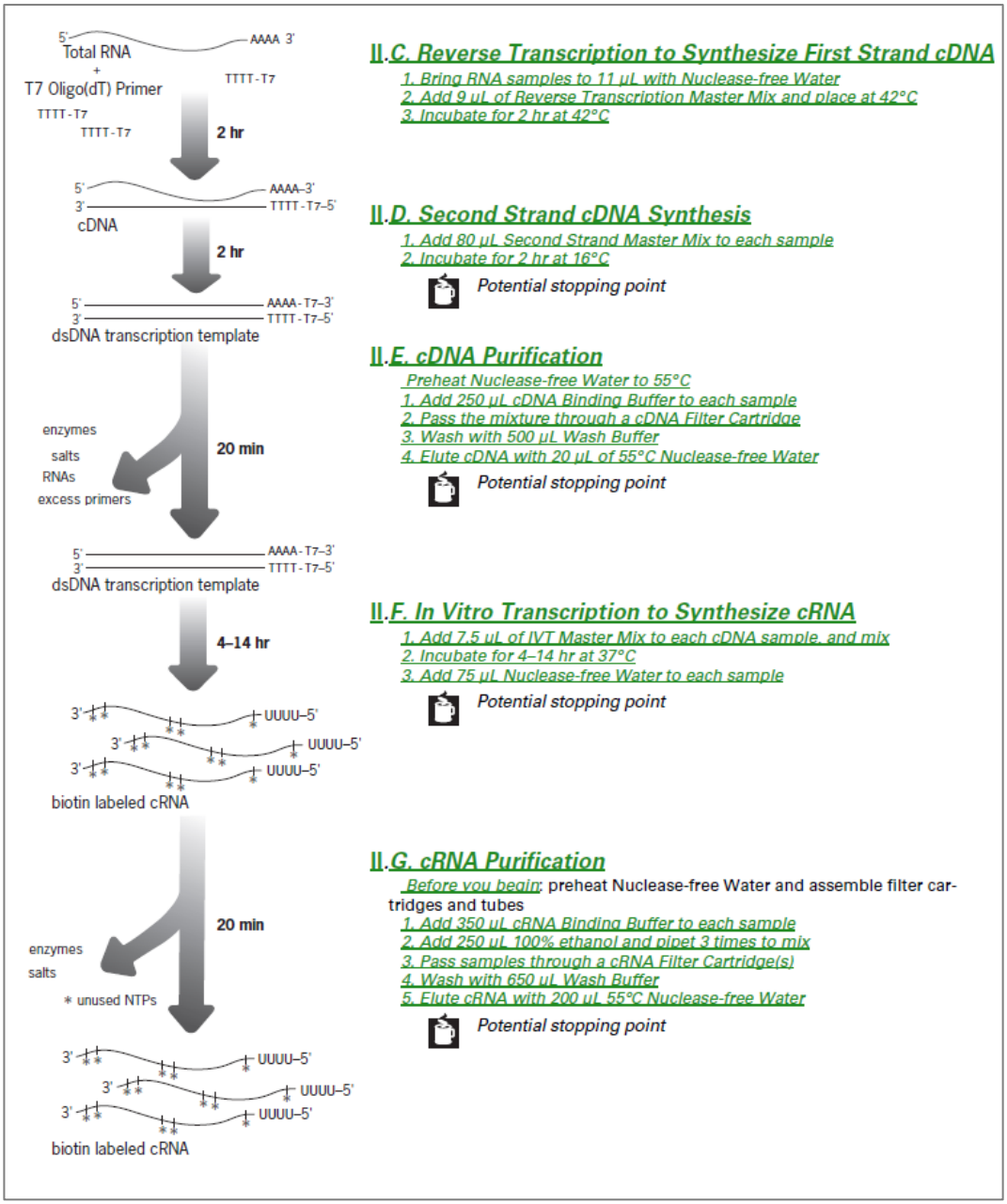


Figure 20 details the workflow for the Illumina total prep amplification kit. This technique involves 5 stages which prepare the RNA for hybridisation with the whole transcriptome array. Adapted from Illumina Total prep amplification protocol (http://tools.lifetechnologies.com/content/sfs/manuals/IL1791ME.pdf),

2.4.2.11.1 Reverse transcription to synthesize First Strand cDNA

The RT mastermix was prepared at room temperature according to Table 18 and 9µl of RT mastermix was added to each sterile 0.2ml eppendorf (Life Technologies) containing 11µl (400ng) of RNA sample prepared in section 3.4.2.8. Samples were then incubated according to Table 19 using a PTC-220 DNA Engine Dyad thermocycler (MJ Research).

Table 18– Reverse transcription mastermix

Reagent	Volume (1 rxn)	Volume (24*rxns)
T7 Oligo(dT) Primer	1µl	26.4µl
10x First Strand Buffer	2µl	52.8µl
dNTP mix	4µl	105.6µl
RNase	1µl	26.4µl
Arrayscript	1µl	26.4µl
TOTAL	9µl	237.6µl

Table 18 details the constituents of the reverse transcription mastermix. The volume required of each reagent is shown. *An overhang of 10% was included to account for sampling error.

Table 19– Reverse transcription reaction times

cDNA reverse transcription thermal cycling		
Step	Temperature (ºC)	Time
1	42 (50 heated lid)	2 hrs.
2	4	Forever

Table 19 shows the reaction cycles used during the cDNA reverse transcription reaction stage. The temperature and timing of each cycle is shown.

2.4.2.11.2 Second strand cDNA synthesis

Second strand mastermix was prepared on ice according to Table 20. 80µl of second strand mastermix was added to each sample and incubated for 2hrs according to Table 21 using a PTC-220 DNA Engine Dyad (MJ Research)

Table 20– Second strand transcription mastermix

Reagent	Volume (1 rxn)	Volume (24*rxns)
Nuclease free H_20	63µl	1663.2µl
10x Second Strand Buffer	10µl	264µl
dNTP mix	4µl	105.6µl
DNA polymerase	2µl	52.8µl
RNase H	1µl	26.4µl
TOTAL	9µl	2112µl

Table 20 details the constituents of the second strand transcription mastermix. The volume used of each reagent is shown. *An overhang of 10% was included to account for sampling error.

Table 21– Second strand transcription reaction times

cDNA second strand transcription thermal cycling		
Step	Temperature (ºC)	Time
1	16 (heated lid disabled)	2 hrs.
2	4	Forever

Table 21 shows the reaction cycles used during gene expression reaction stage. The temperature and timing of each cycle is shown.

2.4.2.11.3 cDNA purification

250µl cDNA binding buffer was added to each sample and transferred to the centre of a 1.5ml collection tube containing a cDNA filter cartridge. Each sample was centrifuged at 10,000xg for 1 minute at 27°C and the flow through discarded. 500µl wash buffer (containing 100% ethanol; Sigma Aldrich) was added to each cDNA filter cartridge and centrifuged at 10,000xg for 1 minute at room temperature. The flow through was discarded and the cDNA filter cartridge centrifuged at 10,000xg for 1 minute at room temperature to remove trace amounts of wash buffer. The cDNA filter cartridge was then transferred to a new cDNA elution tube and 20µl of nuclease free water (preheated to 55°C) was added. The cDNA filter cartridge was incubated at room temperature for 2 minutes and centrifuged at 10,000xg for 1 minute at room temperature.

2.4.2.11.4 In vitro transcription to synthesise cRNA

In vitro transcription mastermix was prepared at room temperature according to Table 20. 7.5µl of IVT mastermix was added to each sample and incubated for 14hrs according to Table 22 using a PTC-220 DNA Engine Dyad (MJ Research).

Table 22– IVT transcription mastermix

Reagent	Volume (1 rxn)	Volume (24*rxns)
T7 10x reaction buffer	2.5µl	75µl
T7 enzyme mix	2.5µl	75µl
Biotin-NTP mix	2.5µl	75µl
TOTAL	7.5µl	225µl

Table 22 details the constituents of the IVT transcription mastermix. The volume used of each reagent is shown. *An overhang of 10% was included to account for sampling error.

Table 23 – IVT reaction times

cRNA IVT thermal cycling		
Step	Temperature (ºC)	Time
1	37 (105 heated lid)	14 hrs.
2	4	Forever

Table 23 shows the reaction cycles used during the IVT transcription stage. The temperature and timing of each cycle is shown.

2.4.2.11.5 cRNA purification

350µl cRNA binding buffer and 250µl 100% ethanol (Sigma Aldrich) was added and each sample pipetted 3 times to mix. Each sample was then transferred to the centre of a 1.5ml collection tube containing a cRNA filter cartridge. Each sample was centrifuged at 10,000xg for 1 minute at 27°C and the flow through discarded. 650µl wash buffer (containing 100% ethanol; Sigma Aldrich) was added to each cRNA filter cartridge and centrifuged at 10,000xg for 1 minute at room temperature. The flow through was discarded and the cRNA filter cartridge centrifuged at 10,000xg for 1 minute at room temperature to remove trace amounts of wash buffer. The cRNA filter cartridge was then transferred to a new cDNA elution tube and 200µl of nuclease free water (preheated to 55°C) was added. The cRNA filter cartridge was then incubated at 55°C for 10 minutes and centrifuged at 10,000xg for 1.5 minutes at room temperature to elute cRNA into the collection tube.

2.4.2.12 Illumina Gene Expression Direct Hybridization Assay

Direct hybridization allows addition and detection of labelled cRNAs to a HumanHT-12 v4 Expression beadchip (Illumina) for analysis using the iScan reading system (Illumina). Each expression bead chip contains 50-mer sequence specific probes, which, by complementary binding to cRNA samples, allows

quantification of whole genome expression. First, labelled high quality cRNA strands are hybridized to the bead chip overnight before washing to remove unbound cRNAs. Analytical probes are then bound to the hybridized bead chip which allows for differential detection of gene expression signatures in a sample. All reagents used in this process are obtained from Illumina unless otherwise stated.

2.4.2.12.1 Hybridization to the bead chip

To normalize samples, 5µl nuclease free water and 10µl HYB reagent were added to each sample. To humidify bead hybridization chambers, 200µl HCB was added to each of the buffer reservoirs. Bead chips were then placed in the hybridization chambers and 15µl of sample was added to the corresponding wells. Chips were then incubated in the Hybridization oven for 14 hours at 58°C with the rocker speed set at 5.

2.4.2.12.2 Washing beadchip

To ensure the removal of unbound sample and prepare samples for detection, a series of washes were performed at described in Table 24. The high temperature wash was performed using a Hybex microsample incubator (Scigene) whilst room temperature washes were performed using a Pyrex staining dish housed in an orbital shaker (Sigma Aldrich). Finally, a block step was performed by loading 4ml Block E4 buffer into a wash tray. Each chip was then transferred to the loaded wash tray and placed on a rocking platform (VWR) at medium rocking speed for 10 minutes.

Table 24– Beadchip wash steps

Wash	Reagent	Time	Temperature
High Temp Wash	1x High temp wash buffer (Diluted 1:10)	10 minutes	55°C
1st Room Temp Wash	E1BC buffer	5 minutes (Shaken)	Room temperature
Ethanol Wash	100% ethanol	10 minutes	Room temperature
2nd Room Temp Wash	E1BC buffer	5 minutes (Shaken)	Room temperature

Table 24 details the reagents used during the Beadchip wash stage.

2.4.2.13 Detecting differential signals on array

To detect the differential signals generated by the analytical probes that have been hybridized to the chip, chips were transferred to a wash tray containing 4ml block E4 buffer containing Cy3-streptavidin (1µg/ml). Using wash tray lids to protect samples from light, chips were placed on a rocking platform for 10 minutes at medium speed. To remove unbound reagents, a third room temperature wash was performed by submerging chips in 250µl Wash E1BC buffer at room temperature for 5 minutes. Chips were then dried by centrifuging at 1,400 RPM for 4 minutes before being transferred to the iScan for imaging of the array. Once the arrays were scanned, files were generated detailing a number of sample and array metrics. These files were used to normalise and QC the gene expression data in a series of stages.

2.4.2.14 Gene expression data normalisation

To ensure an accurate representation of the data, all probes and samples underwent normalisation and QC collectively but final output featured gene

expression data for 4 probes in the selected region, across 22 CD8+ and 22 CD4+ lymphocyte samples. Gene expression normalisation and quality control was performed by Dr Darren Plant at the Arthritis Research UK Centre for Genetics and Genomics. All analysis was performed using bead array, limma, illuminaHumanv4.db, corpcor and pcaMethods packages installed through R version 3.0.2 (Dunning et al. 2007;Sarembaud et al. 2007;Wettenhall and Smyth 2004) unless otherwise stated.

2.4.2.15 Calculation of the signal to noise ratio across arrays

To identify how each array performed the signal to background noise ratio was calculated for 45 samples across all 4 arrays and visualised using a scatterplot. Generally, the expected value of signal to noise is 5-14 but a value of less than 2 indicates a sample should be removed from further analysis.

2.4.2.16 Calculation of intensity signals across probes

The average signal intensity of both regular and negative control probes across all arrays was calculated from the raw array data. Background correction, NEQC quantile normalization and a $\log^2$ conversion was then performed using the values from the negative control probes, to account for differences between array performances. This was visualised using boxplots showing differences between raw and normalized data.

2.4.2.17 Calculation of the proportion of probes expressed by each sample

The proportion of expressed probes was calculated for the CD8+ and the CD4+samples. A t test was then performed to identify if the proportion of expressed probes was different for each cell type. The number of probes which were expressed in at least one sample was identified and probes which were not

expressed at all across all arrays were removed from further analysis.

2.4.2.18 Matching probes to hg19 genome build

To ensure that the array annotation is as accurate as possible, the remaining probes were mapped to hg19 genome build using the UCSC genome browser. Probes were scored as "Perfect, good, bad and no match," depending on their mapping to transcripts. Unmatched and poor probes were removed from further analysis.

2.4.2.19 Identification of sample outliers

To identify how similar the samples within the sample groups (CD8+ and CD4+) were, multidimensional scaling plots (MDS plots) were generated. If samples are within the same group (CD4+ or CD8+) it is expected that they will cluster together, whilst sample outliers should be removed from further analysis.

2.4.2.20 Principal components analysis

To identify factors which may be contributing to sample variance and therefore could contribute to batch effects, principal components analysis (PCA) was performed. Any principal component (PC) which contributed to more than 10% variance was considered significant and was adjusted for in subsequent analysis.

2.4.2.21 Array weighting

As the performance of specific arrays can be affected by variation in chemistry and sample quality; array weighting was performed to account for this. This process allows poorer quality arrays to be downgraded in the analysis but still used therefore increasing the overall power of the study. Array weights were

determined using the limma package in R version 3.0.2 and used in the generation of the final probe expression values.

2.4.2.22 Cell specific eQTL analysis

To identify if an associated variant represented an eQTL with a gene in T lymphocyte subsets, linear regression was performed in STATA v.11.2. Genotype data was obtained from the data generated in section 2.4.1.2. Gene expression data was extracted from probes which were less than 400kb from the gene of interest. Linear regression was performed using genotype at the SNP of interest as a covariate.

3.0 Results

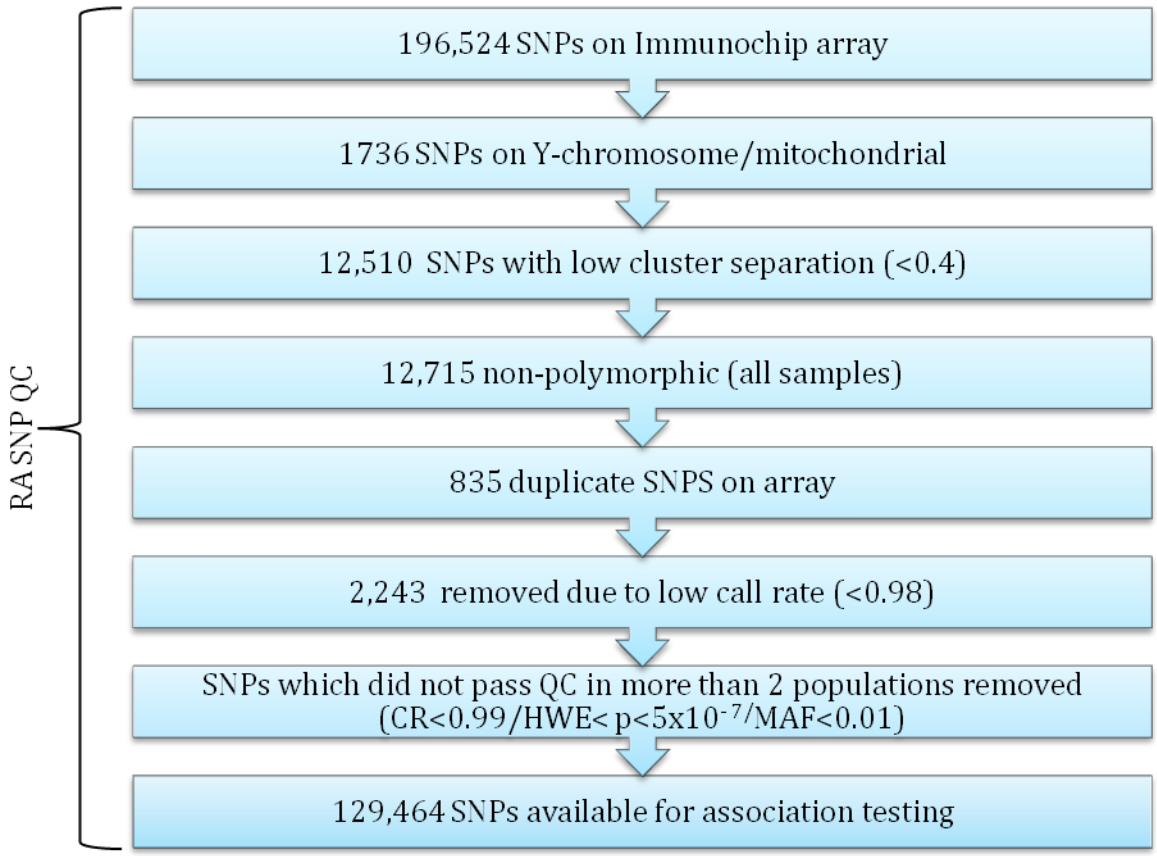
3.0 Results

3.1 Inflammatory arthritis overlap

3.1.1 Immunochip SNP and sample QC

All SNP and sample quality control (QC) was performed by Dr John Bowes, Dr Anne Hinks and Dr Joanna Cobb of the Arthritis Research UK Centre for Genetics and Genomics. Figure 21 is a summary of the number of SNPs removed at each stage of the QC for each disease. Figure 22 is a summary of the number of samples removed at each stage of the sample QC. The total number of SNPs and samples used in the association analysis is shown in Table 25.

Figure 21– SNP QC for each disease



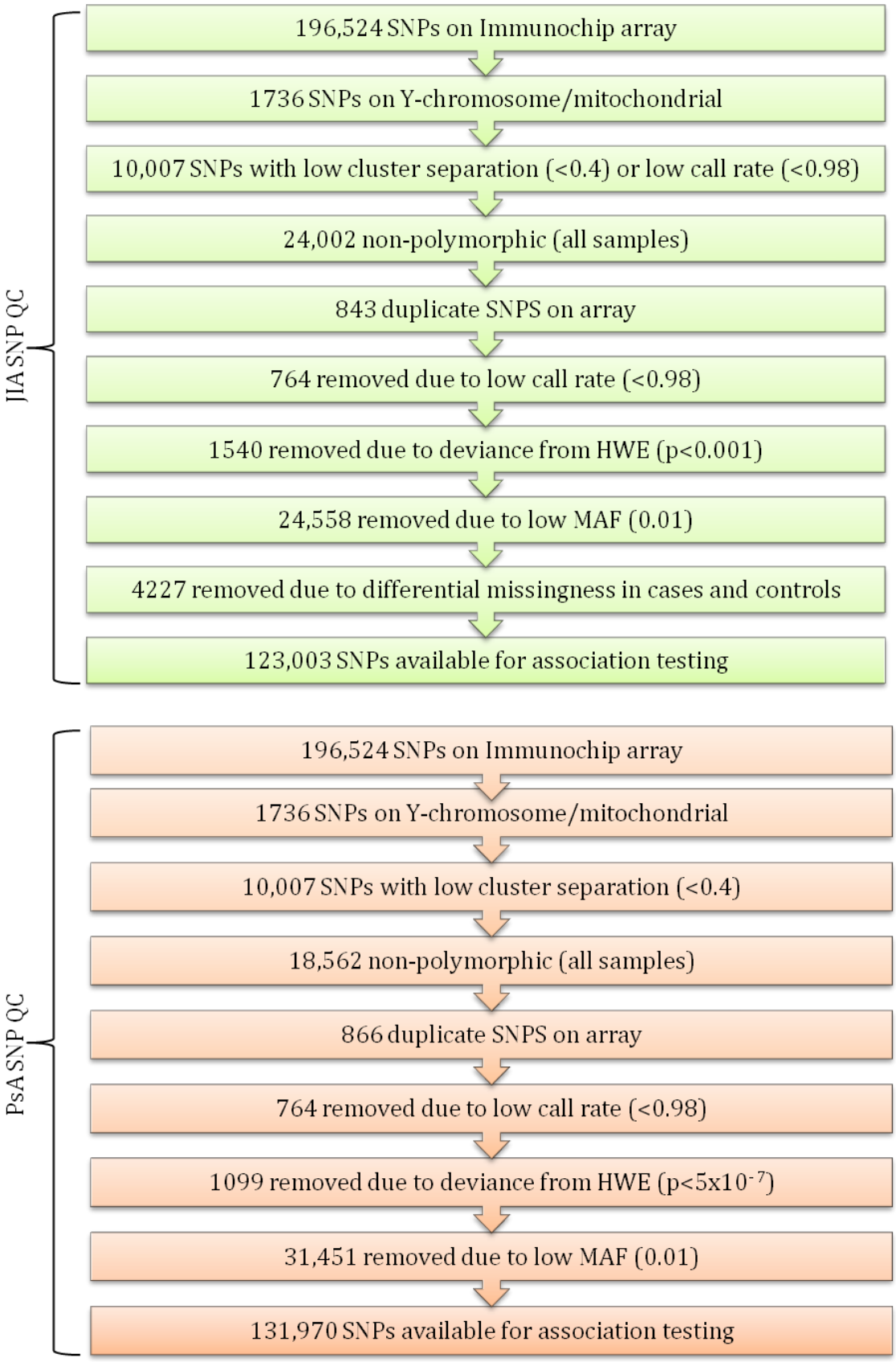
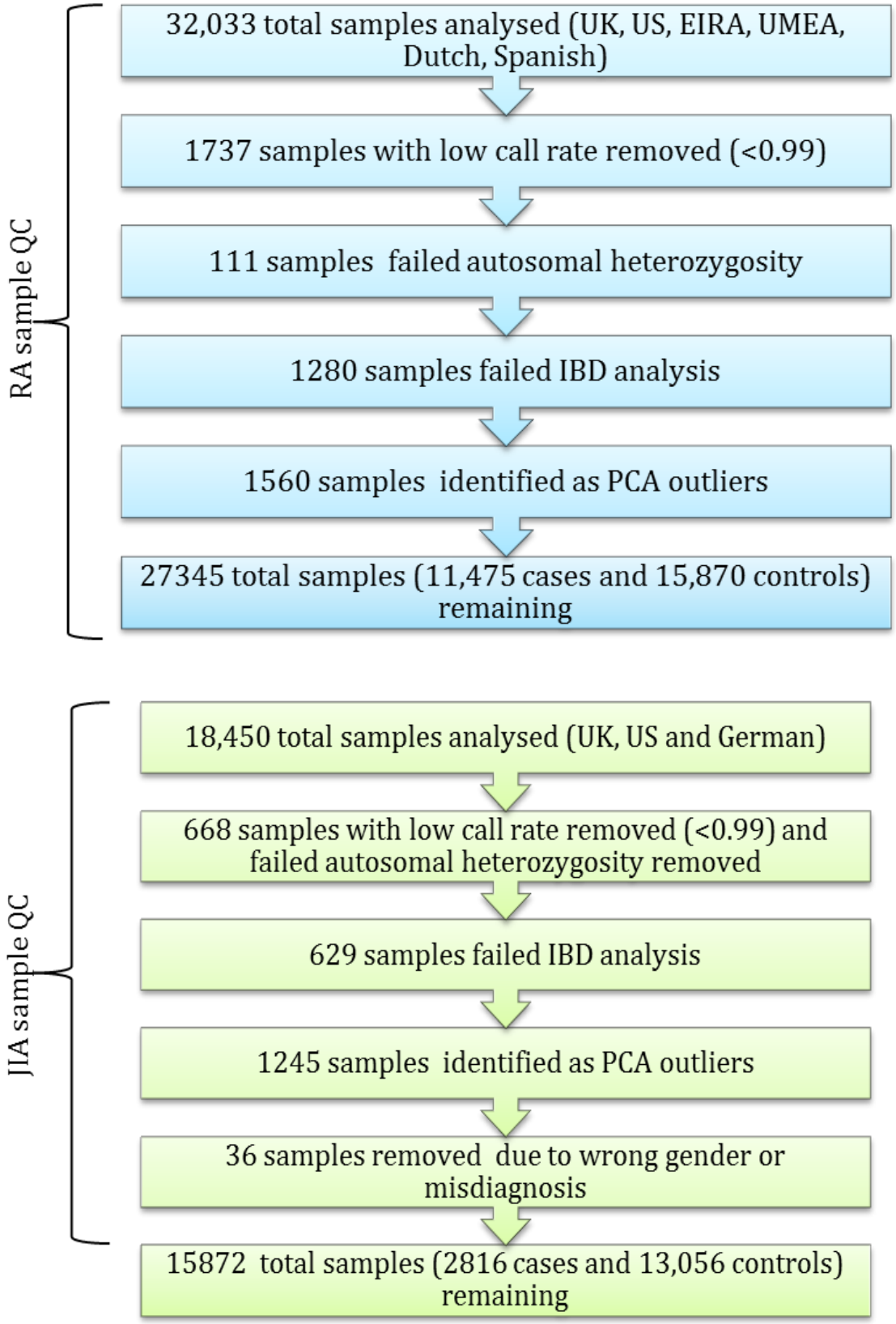


Figure 21 shows the SNP QC stages used for RA, JIA and PsA. To ensure a robust dataset was obtained SNPs which did not reach the QC threshold in the following catagories were removed: SNPs on the y/mitochondrial chromosome; SNPs with a low cluster separation; non-polymorphic SNPs; duplicate SNPs on the array; SNPs with a low call rate; SNPs which deviated from HWE; SNPs with a low MAF. HWE = Hardy Weinberg Equilibrium, MAF = Minor allele frequency.

Figure 22– Sample QC for each disease



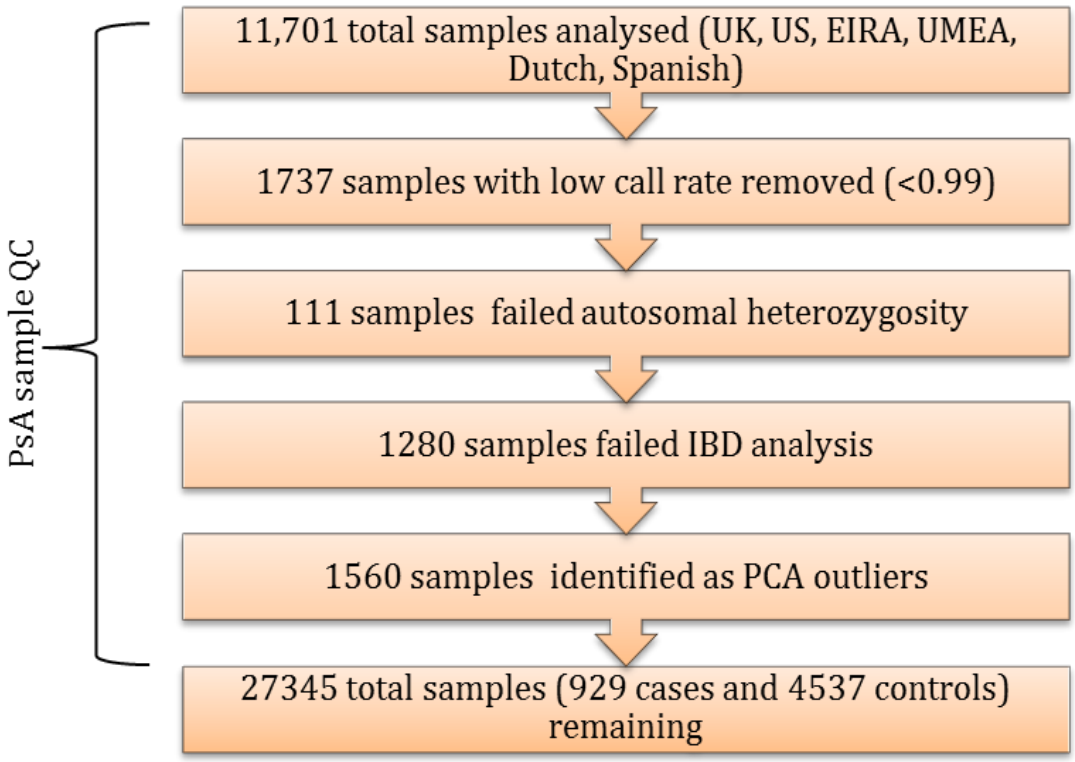


Figure 22 shows the sample QC stages used for RA, JIA and PsA. To ensure a robust dataset was obtained samples which did not reach the QC threshold in the following categories were removed: samples with a low call rate; samples with excessive autosomal heterozygosity; samples which failed IBD and samples identified as outliers using PCA. IBD = Identity by descent, PCA = Principal components analysis.

Table 25– Summary of total SNP and samples available for each disease post QC

Disease	SNPs	Cases	Controls
RA	129,464	11,475	15,870
JIA	123,003	2816	13,056
PsA	131,970	929	4537

Table 25 shows the number of SNPs, cases and controls which were available for analysis across RA, JIA and PsA. RA = Rheumatoid arthritis, JIA = Juvenile idiopathic

3.1.2 Power of each cohort to detect genetic effects

To identify the power of each of the individual studies and, therefore, the likelihood to detect genuine genetic effects, power calculations were performed. Table 26 shows the power of each of the individual studies to detect common (MAF>0.05) and rare (MAF>0.01) genetic effects with ORs of 1.2. In RA in the likelihood to detect rare SNPs was moderate at 59% but for common variants is 99%, indicating that this is a well powered study. In JIA the power to detect common variants is high (81%) but the study is underpowered when looking at rarer SNP (25%). In contrast, the PsA study is underpowered to detect both common and rare SNPs, with power of 37% and 11% respectively.

Table 26 – Power to detect genetic effects with OR = 1.2

Disease	Rare (MAF>0.01)	Common (MAF>0.05)
RA	59%	99%
JIA	25%	81%
PsA	11%	37%

Table 26 shows the power of each cohort to detect genetic effects at MAF of 1% and 5%. RA = Rheumatoid arthritis, JIA = Juvenile idiopathic arthritis, PsA = Psoriatic arthritis, MAF = minor allele frequency.

3.1.3 Calculating the number of inflammatory arthritis overlapping regions

Association testing was performed by Dr John Bowes, Dr Anne Hinks and Dr Joanna Cobb of the Arthritis research UK Centre for Genetics and Genomics. Overlap analysis was performed using each of the datasets generated.

In total 50 genetic regions contained SNPs associated with more than type of IA (IA; $p<1x10^{-3}$). Four of these regions reached genome wide significance ($p<5x10^{-8}$) in more than 1 disease, with the remaining regions reaching suggestive

significance ($p < 1x10^{-3}$). Table 27 shows the 50 regions and their respective index SNPs, p values and odds ratios across the diseases. In particular 14 regions shown in bold had a significance level of $p < 1x10^{-5}$ in more than 1 type of IA. Association plots for 2 of these regions are shown in Figure 24 and Figure 25.

Many regions such as *RUNX1, TYK2*, *IL6R, RASGRP1* and *IRF8* are novel IA associations and will require replication in an independent cohort but for now provide an insight into the shared genetics of these diseases. 10 regions (*TYK2, EOMES, CTLA4, RUNX1, IL2RA, PTPN2*, *IGSF3/CD2*, *RAB5B/ERBB2*, *IL2/IL21*, *and IL23R*) were associated across the 3 diseases inclusively but the majority of genetic overlap was observed between RA and JIA; Figure 23 shows the distribution of overlapping regions between the diseases. There were a total of 31 regions shared between RA and JIA, suggesting that the greatest overlap is shared between these diseases. In addition, 18 genetic regions were also associated with PsA and either RA or JIA indicating that overlap also exists between these diseases.

Table 27 – Regions associated with multiple types of Inflammatory Arthritis

Region	Chr	RA SNP	RA p	RA OR	JIA SNP	JIA p	JIA OR	PsA SNP	PsA p	PsA OR
PTPN22	1	rs2476601	$1.6x10^{-67}$	1.60	rs6679677	$3.19x10^{-25}$	1.59			
MMEL1	1	rs28532547	$1.82x10^{-8}$	0.90	rs1001620	$6.49x10^{-4}$	1.11			
IL6R	1	rs8192284	$8.38x10^{-8}$	0.91	rs11265608	$1.55x10^{-7}$	1.28			
PTPRC	1	rs2014863	$1.66x10^{-5}$	1.08	rs61829344	$1.69x10^{-4}$	0.81			
IGSF3/CD2	1	rs798000	$2.36x10^{-5}$	1.08	rs12725472	$9.85x10^{-4}$	1.12	rs77421743	0.0003201	1.83
NCF2	1	rs17849502	$2.75x10^{-5}$	1.18	rs7531089	$6.61x10^{-4}$	1.11			
CD247	1	rs840016	0.0006422	0.94	rs2056626	$6.84x10^{-5}$	0.88			
IL23R	1	rs12145984	0.0006482	0.92	rs17129835	$1.82x10^{-4}$	1.12	rs56920441	$3.86x10^{-8}$	1.35
RUNX3	1				rs4648881	$4.66x10^{-7}$	1.16	rs4649038	0.0002215	1.22
LCE3B/LCE3A	1				rs11205044	$5.03x10^{-4}$	0.9	rs10888503	0.0008668	0.83
STAT4	2	rs7574865	$2.47x10^{-9}$	1.13	rs10174238	$1.2x10^{-13}$	1.29			
CTLA4	2	rs3087243	$3.31x10^{-9}$	0.90	rs231725	$1.53x10^{-5}$	1.15	rs11571312	$3.51x10^{-5}$	1.34
AFF3	2	rs11123811	$7.63x10^{-9}$	1.11	rs7580200	$2.36x10^{-6}$	1.15			
NAB1	2	rs10931468	$9.13x10^{-7}$	1.14	rs10931468	$1.07x10^{-6}$	1.23			

SPRED2	2	rs6546146	$1.61x10^{-5}$	0.92	rs268122	$2.7x10^{-4}$	0.9			
DNASE1L3	3	rs35677470	$8.97x10^{-7}$	1.17	rs35677470	$2.48x10^{-5}$	1.25			
CCR9/CCR 2 Chemo	3	rs17078454	$3.78x10^{-6}$	1.11	rs62625034	$3.16x10^{-7}$	0.77			
IL12A	3	rs4680536	$1.19x10^{-5}$	0.92	rs2366643	$4.63x10^{-4}$	0.9			
EOMES	3	rs9880772	0.0006779	1.06	rs9862284	$6.22x10^{-4}$	1.11	rs733302	0.000586	1.92
IL2/IL21	4	rs62323881	0.0005067	1.12	rs1479924	$6.24x10^{-11}$	0.8	rs62324211	$2.45x10^{-5}$	
ANKRD55 /IL6ST	5	rs10065637	$6.01x10^{-16}$	0.83	rs10065637	$7.69x10^{-11}$	0.77			1.46
SLC22A4 /SLC22A5	5				rs4705862	$1.02x10^{-8}$	0.84	rs2278398	$8.84x10^{-5}$	
ERAP1/ER AP2	5				rs27300	$2.13x10^{-5}$	1.14	rs116764930	0.0002434	1.23
TNFAIP3	6	rs6920220	$2.15x10^{-12}$	1.16	rs5029924	$2.86x10^{-6}$	1.42			1.81
BACH2	6	rs56258221	$1.12x10^{-6}$	1.12	rs72928038	$3.09x10^{-4}$	1.15			
PRDM1	6	rs7739286	0.0002004	0.92	rs12205855	$5.91x10^{-5}$	1.23			
LRRC16A	6	rs742132	0.0003653	1.07	rs7740510	$5.01x10^{-6}$	0.85			

IKZF1	7	rs59538194	$8.69x10^{-5}$	1.12	rs35295940	$8.73x10^{-4}$	1.35			
AMPH/TARP	7	rs10278233	0.0002378	0.92				rs2392581	0.0008267	
BLK	8	rs4840565	$4.15x10^{-7}$	1.10	rs62490933	$9.86x10^{-4}$	0.81			0.84
PVT1	8	rs13281279	0.0001462	0.93	rs13281279	$1.51x10^{-4}$	0.89			
IL7	8	rs9298320	0.0006159	1.08				rs7822255	0.0005697	
TRAF1/C5	9	rs917770	$8.01x10^{-6}$	0.92	rs4837811	$1.69x10^{-4}$	1.13			1.21
IL2RA	10	rs7073236	$9.89x10^{-6}$	1.08	rs7909519	$8.00x10^{-10}$	0.72			
TREH	11	rs11217040	$3.31x10^{-6}$	0.90				rs10790255	0.0006756	
RAB5B/ERBB3/STAT2	12	rs10876870	0.0001489	0.93	rs1614219	$7.23x10^{-4}$	0.82	rs2020854	$6.97x10^{-5}$	0.80
PTPN11	12	rs7299227	0.0008308	0.77	rs17630235	$3.11x10^{-8}$	1.18			0.61
COG6	13	rs7993214	$4.98x10^{-5}$	0.93	rs7993214	$1.61x10^{-7}$	0.84			
ZFP36L1	14	rs7146217	0.0007464	0.94	rs12434551	$3.62x10^{-8}$	0.85			
RASGRP1	15	rs8043085	$3.53x10^{-9}$	1.13	rs6495986	$2.13x10^{-4}$	0.88			
IRF8	16	rs13330176	$3.07x10^{-6}$	1.10	rs2280381	$3.96x10^{-5}$	0.88			
CIITA	16	rs8056450	0.0005213	0.91	rs12598246	$3.58x10^{-4}$	1.12			

SMARCE1	17	rs723729	0.0008659	1.06				rs757412	$3.96x10^{-5}$	
NOS2	17				rs34913965	$1.51x10^{-6}$	0.85	rs4795067	0.0008984	1.37
PTPN2	18	rs7241016	$1x10^{-5}$	1.11	rs2847293	$1.44x10^{-12}$	1.31	rs11302687	0.0001132	1.20
TYK2	19	rs34536443	$3.40x10^{-13}$	0.70	rs34536443	$1.00x10^{-10}$	0.56	rs12720356	0.0007397	1.69
RUNX1	21	rs9979383	$8.06x10^{-7}$	0.91	rs9979383	$1.05x10^{-7}$	0.85	rs99793	0.0006116	0.71
IL2RB	22	rs9607418	$5.65x10^{-6}$	1.09	rs2284033	$1.55x10^{-8}$	0.84			0.83
UBE2L3	22	rs2266959	$7.11x10^{-6}$	1.10	rs2266959	$7.30x10^{-9}$	1.24			
MTMR3	22	rs77378082	$4.52x10^{-5}$	0.74	rs5763631	$6.18x10^{-4}$	1.39			

Table 27 shows the association results of the 50 overlapping regions. For each region the associated SNP, p value and odds ratio are shown for each disease. RA = Rheumatoid arthritis, JIA = Juvenile idiopathic arthritis, PsA = Psoriatic arthritis, p = p value, OR = odds ratio

Figure 23– Distribution of overlap between diseases

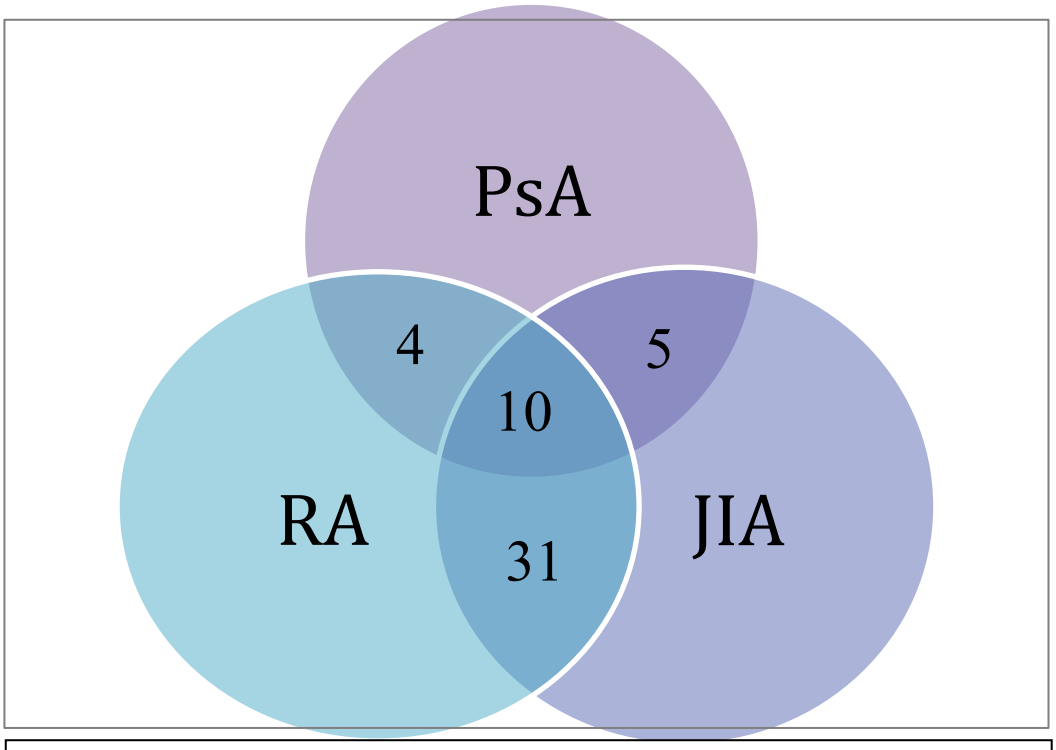


Figure 23 shows the regions associated with more than 1 type of IA. In total 50 regions are associated with more than 1 type of inflammatory arthritis with greatest genetic overlap present between RA and JIA. RA = rheumatoid arthritis, JIA = juvenile idiopathic arthritis, PsA = psoriatic arthritis.

3.1.4 Identification of correlation between SNPs in overlapping regions

To identify whether different or identical SNPs in genetic region are associated with disease, LD between each index SNP was calculated for the 50 regions. In regions which are associated with 3 diseases, LD between each of the 3 SNPs was calculated. Table 28 shows the correlation between each of the index SNPs in the overlapping regions. In 14 regions the SNP associated was either identical or highly correlated (r^2>0.8). Furthermore in all these regions, similar odds ratios were observed indicating this may be the same effect which is contributing to different diseases.

Figure 24 shows an association plot for the *RUNX1* region, in which an identical SNP is associated with each of the 3 diseases. In 9 regions the SNPs associated with each disease are different but moderately correlated by LD (r^2>0.4<0.8). In 6 regions (*PTPN2*, *SPRED2, IL2/IL21, CD247, EOMES* and *ZFP36L1*) the direction of effect is similar across the diseases but in 3 regions (*CTLA4, CIITA* and *IL23R*) the effect directions are different across diseases. In 32 regions there is limited or no LD observed between the SNPs (r^2<0.4). Interestingly in many of these regions the risk allele is opposing across the associated diseases. This indicates that it may represent a different effect within the same genetic region which is contributing to disease susceptibility. Figure 25 shows the *IL2RA* region, as an example of a region in which a different SNP is associated with RA and JIA, with low correlation observed between the SNPs. Notably in 5 regions (*TYK2, CTLA4, PTPN2, IL2/IL21* and *CD247*) which are associated across the 3 diseases the index SNPs in 2 diseases are the same or highly correlated whilst the index SNP in the third disease shows no correlation. This may indicate that in some regions the effects are the same across some diseases but different in the others or that multiple effects exist within the region.

Figure 24 – *RUNX1* region association plots

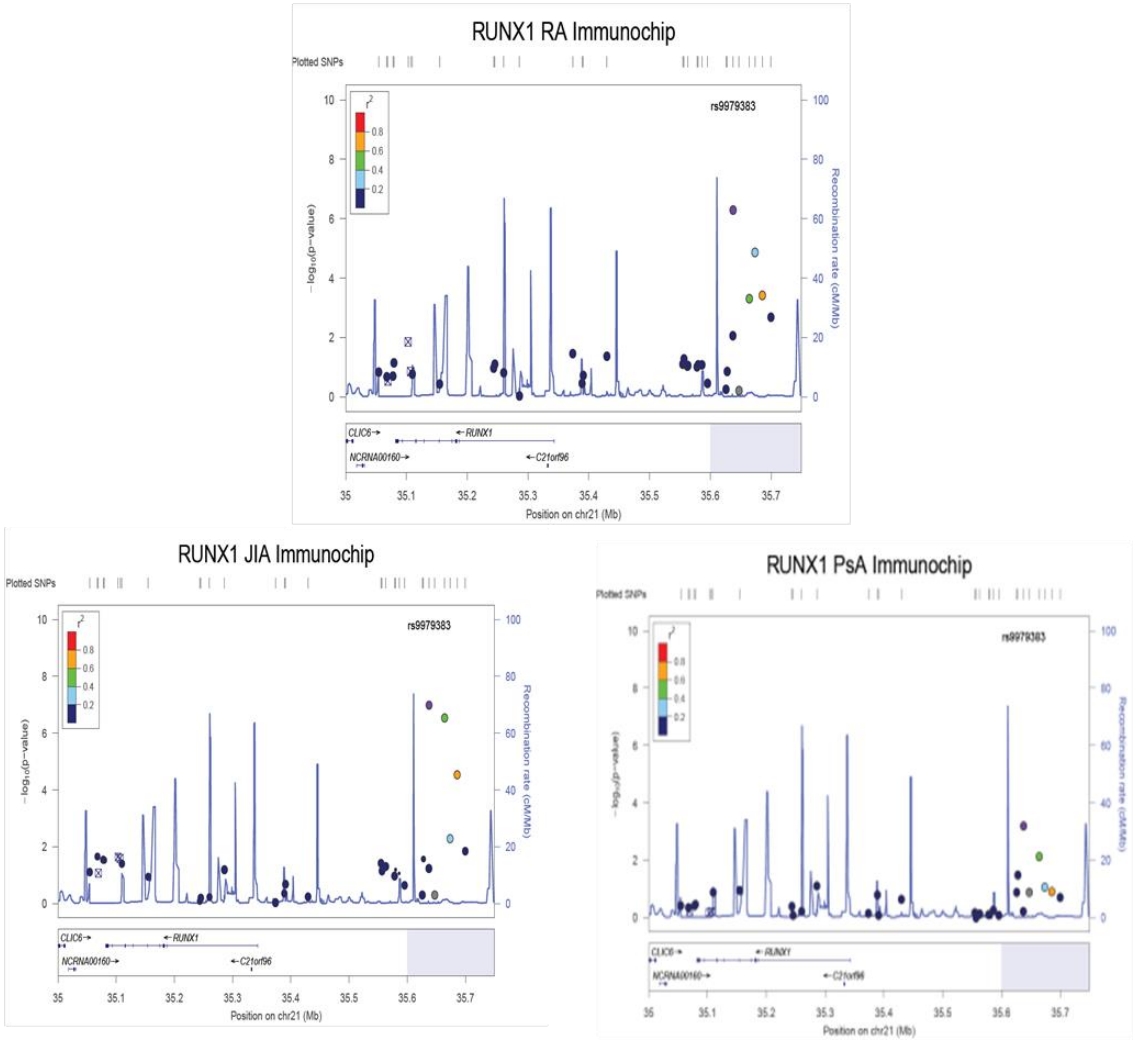


Figure 24 shows association plots for the *RUNX1* region for RA (top), JIA (bottom left) and PsA (bottom right). In each plot the x-axis represents the base position across the genome whilst the y-axis represents the $-\log_{10}$ of the p-value. In each plot a dot represents a SNP with the colour coding representing LD with the annotated index SNP. LD = linkage disequilibrium.

Figure 25– IL2RA region association plots

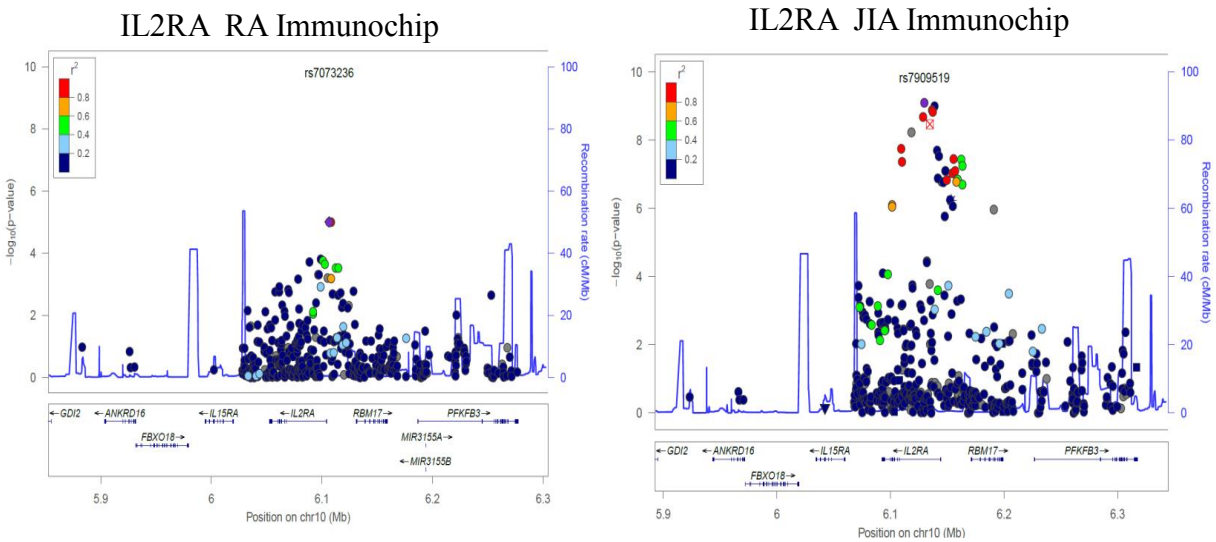


Figure 25 shows association plots for the *IL2RA* region for RA (left) and JIA (right). in each plot the x-axis represents the base position across the genome whilst the y-axis represents the $-\log_{10}$ of the p-value. In each plot a dot represents a SNP with the colour coding representing LD with the annotated index SNP. LD = linkage disequilibrium.

Table 28 – Correlation between index SNPs in overlapping regions

Same SNP/Highly correlated ($r^2>0.8$)	Different SNP/moderately correlated ($r^2>0.4<0.8$)	Different SNP/low correlation ($r^2<0.4$)
PTPN22	CTLA4 (RA/JIA)	TNFAIP3
ANKRD55/IL6ST	PTPN2 (RA/JIA)	CTLA4 (RA/PsA and JIA/PsA)
TYK2 (RA and JIA)	SPRED2	RASGRP1
STAT4	IL2/IL21 (RA/PsA)	MMEL1
AFF3	CIITA	IL6R
RUNX1	CD247	BLK
DNASE1L3	IL23R (RA/JIA)	IRF8
NAB1	EOMES (RA/JIA)	TREH
BACH2	ZFP36L1	CCR9/CCR2 Chemo
UBE2L3		IL2RB
IL12A		TRAF1/C5
COG6		IL2RA
PVT1		PTPN2 (RA/PsA and JIA/PsA)
LCE3B/LCE3A		PTPRC
		IGSF3/CD2

	NCF2
	MTMR3
	IKZF1 TYK2 (RA/JIA and PsA)
	RAB5B/ERBB3/STAT2
	PRDM1
	AMPH/TARP
	LRRC16A
	IL2/IL21 (RA/JIA and JIA/PsA)
	IL23R (RA/PsA and JIA/PsA)
	PTPN11
	SMARCE1
	RUNX3
	SLC22A4/SLC22A5
	ERAP1/ERAP2
	NOS2
	EOMES (RA and PsA/JIA and PsA)

Table 28 shows the correlation between the SNPs in the 50 overlapping regions. RA = Rheumatoid arthritis, JIA = Juvenile idiopathic arthritis, PsA = Psoriatic arthritis.

3.1.5 Selecting a functionally promising region for further analysis

To prioritize a promising overlapping region for further functional analysis a number of factors were considered including the number of types of IA a region was associated with, the p value size in each disease and whether the same/highly correlated SNP was associated with each disease. This was performed to detect any associations, which appeared to be similar between the 3 diseases, and therefore would be good candidates for investigation common biological effects.

Of the 50 overlapping regions identified in section 3.1.3, 9 were associated across the 3 diseases (*TYK2, EOMES, CTLA4, RUNX1, IL2RA, PTPN2, CD2/IGSF3, ERBB2, IL2/IL21 and IL23R*). Of these 10 regions in only 2 (*RUNX1* and *TYK2*) was the SNP identical or highly correlated by LD (r^2>0.8) indicating that the same genetic effect may be associated with each disease. When the strength of p values were considered the RUNX1 was the most associated. Furthermore the direction of effect is the same with rs9979383 conferring protection (OR = 0.83-0.9) across the 3 diseases providing further evidence that this region represents a true overlapping disease association.

To provide some evidence of the functional contribution of this SNP to disease susceptibility, focused bioinformatics analysis was performed using a variety of databases. In addition a literature search was performed to identify previous associations and potential biological functions of the *RUNX1* region. To account for the fact that the index SNP rs9979383 may be correlated with a more functionally promising SNP, all bioinformatics analysis was performed on rs9979383 and a perfect (r^2 = 1) proxy rs8129030.

3.1.6 RUNX1 functional annotation

The genomic region where rs9979383 and rs8129030 reside may be intergenic as they map ~300kb upstream of most isoforms of the *RUNX1* gene itself but can also be considered intronic as they lie in intron 7 in the longest splice variant of *RUNX1*. They are also in close proximity to *LOC100506403* which is a long non-coding RNA (lincRNA).

3.1.6.1 RUNX1 eQTL analysis

When entered as a search query in the Genevar database, no significant cis-eQTL with rs9979383 or rs8129030 was found ($p<1x10^{-3}$), although this was limited by poor coverage of the region on the genotyping and microarrays. When entered into the SCAN eQTL database, a trans-eQTL was identified between both rs8129030, rs9979383 and the retinol binding protein 5 (*RBP5*) gene on chromosome 12 in a Hapmap CEU cell line ($p=3x10^{-5}$). Table 29 summarises these findings. *RBP5* is involved in the intracellular transport of retinol (Vitamin A) which has been shown to be associated with bone fragility (Conaway et al. 2011).

Table 29– *RUNX1* eQTL analysis

SNP ID	Chromosome	Position	Gene	Population	p
rs8129030	21	35634458	*RBP5*	CEU	$6x10^{-5}$
rs9979383	21	35637631	*RBP5*	CEU	$3x10^{-5}$

Table 29 shows the eQTL analysis results from the SCAN eQTL database which contained a significant trans-eQTL with RBP5. P = p value, CEU = CEPH (Utah Residents with Northern and Western European Ancestry).

3.1.6.2 *RUNX1* Transcription factor binding analysis

To identify whether the rs9979383 or rs8129030 lie in regions which alter TF binding, these were entered as search queries in the UCSC genome browser. Figure 26 shows the region surrounding rs9979383 as a track on the UCSC genome browser with histone modification and TFBS shown. It can be seen that rs9979383 lies within a region which carries an enhancer associated histone methylation mark (H3K4Me1) in NHEK (keratinocyte cell line) and has been shown to interact with a number of TFs such as the p65 *(*RELA) w As a TF analysis could be performed to identify the contribution of the associated SNP to *RUNX1* binding As a TF analysis could be performed to identify the contribution of the associated SNP to *RUNX1* binding hich is an essential part of the nuclear factor kappa-light-chain-enhancer of activated B cells (NFKB) pathway. The findings indicate that the SNP

may lie in an enhancer region of the *RUNX1* gene and may be a potential expression quantitative trait locus (eQTL) yet to be identified in the bioinformatics databases examined previously.

Figure 27 shows the region surrounding rs8129030 as a track on the UCSC genome browser. Although this region does not carry the same histone marks as rs9979383, it does exhibit strong binding with the transcriptional repressor CTCF (CTCF), which is involved in regulating the 3D structure of chromatin in a number of cell lines (Rubio et al. 2008). As the studies which populate this database are performed in a small number of samples, further analysis is required to identify the presence of TFBS in more relevant cell types.

Figure 26– rs9979383 region TF binding

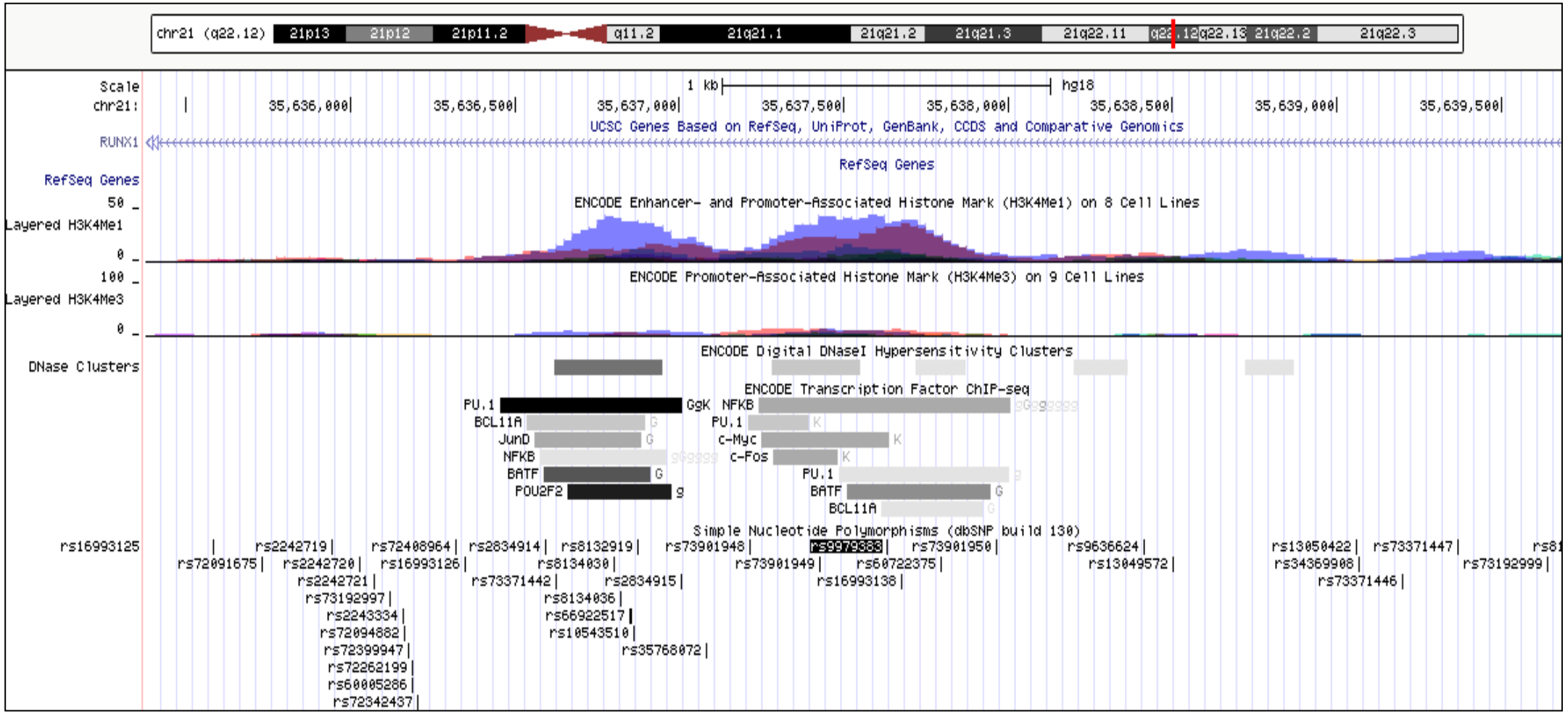


Figure 26 is a UCSC genome browser plot showing the transcription factor potential of rs9979383. Tracks showing histone modifications, DNaseI hypersensitivity and transcription factor binding potential are shown with rs9979383 highlighted in black.

Figure 27– rs8129030 region TF binding

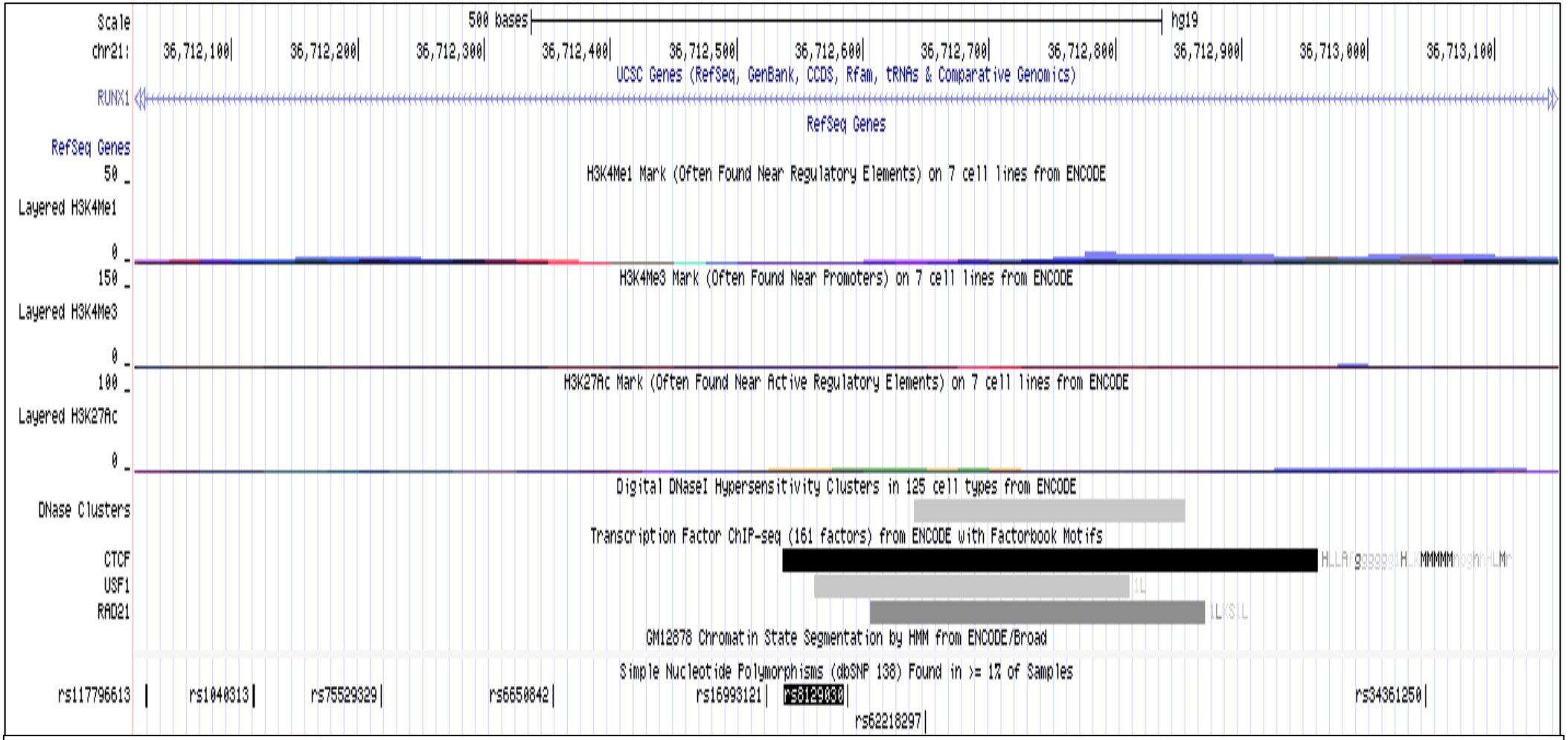


Figure 27 is a UCSC genome browser plot showing the transcription factor potential of rs8129030. Tracks showing histone modifications, DNaseI hypersensitivity and transcription factor binding potential are shown with rs8129030 highlighted in black.

3.1.6.3 RUNX1 literature search

To gain an insight into the potential biological role of a genetic association in this region Runt related transcription factor (*RUNX1*) and its alias acute myeloid leukaemia 1 (*AML1*) were entered as keyword search queries into NCBI PubMed (http://www.ncbi.nlm.nih.gov/pubmed). The literature search returned 2531 publications for *RUNX1* and 1850 for *AML1*. Publications were examined by eye with particular attention paid to those reporting the potential function of the product of *RUNX1* and it's involvement in the immune response. As it encodes a transcription factor (TF), *RUNX1* has been shown to act on various genes involved in T lymphocyte generation, bone formation and has previously been implicated in other immune mediated diseases.

RUNX1 has been shown to be an important regulator for haematopoiesis, especially during embryonic development. In a mouse model knockout of RUNX1, the absence of *RUNX1* is lethal as a consequence of a lack of foetal liver haematopoiesis (Okuda et al. 1996). When adult transgenic mice lacking RUNX1 were generated, haematopoiesis was fully functional in the myeloid compartment but megakaryocyte and lymphocyte development was inhibited, indicating *RUNX1* is essential for these pathways in adulthood (Ichikawa et al. 2008). *RUNX1* has been shown to be important in chondrogenesis by promoting cell maturation whilst regulating production of MMPs. *RUNX1* expression has been shown to be dysregulated in osteoarthritis, indicating downregulation of *RUNX1* could potentially result in joint hypertrophy characteristic of this disease (Yano et al. 2013).

RUNX1 appears to be particularly important for lymphocyte development, particularly in cell polarization. Binding sites for *RUNX1* have been shown to essential for gene silencing in thymocytes, driving CD8+ lineage commitment (Taniuchi et al. 2002) whilst it has also been shown to activate Class I MHC expression in vivo (Howcroft et al. 2005). In CD4+ lymphocytes *RUNX1* has been shown to repress GATA3 expression to skew cells away from the Th2 lineage

(Komine et al. 2003). It has been shown to form a complex with Th1 master regulation Tbet, which inhibits the *RUNX1* mediated transcription of the gene encoding RORγt, preventing differentiation into Th17 cells (Lazarevic et al. 2011). *RUNX1* knock down in CD4+ T lymphocytes in the mouse been shown to result in spontaneous autoimmune lung disease driven by IL17 and IL21, characteristic of Th17 cells (Lazarevic et al. 2011;Wong et al. 2012). In Treg cells *RUNX1* also plays an important role, interacting with FoxP3+ to suppress transcription of inflammatory cytokines such as IL2 and IFNγ (Ono et al. 2007), highlighting the crucial role that RUNX1 plays in regulation of immune development.

With respect to inflammatory disease RUNX1 binding sites have been previously associated with multiple autoimmune diseases including RA, PsV and SLE (Alarcon-Riquelme 2003). Association within the *RUNX1* region in 3 types of inflammatory arthritis is therefore an interesting observation (Eyre et al. 2010;Hinks et al. 2013). Bowes et al. Manuscript in preparation). Although the mechanism by which *RUNX1* has not been characterized in these diseases it could be hypothesised that a role in lymphocyte or chondrocyte development could be implicated. As *RUNX1* represents a TF capable of a number of roles across different cell types, it is therefore important that this TF is investigated further in a specific tissue of relevance.

Collectively these observations this makes *RUNX1* a strong candidate for further genetic and functional investigations. As the SNP is not located within the coding region of *RUNX1*, it is appropriate to first investigate whether the variant is directly affecting the *RUNX1* gene. As this SNP is located upstream of *RUNX1*, it was hypothesised that it may affect transcription of the *RUNX1* gene and therefore was subjected to eQTL gene expression studies described in section 2.5. In particular experiments looking at eQTL analysis in T lymphocytes were designed due to the essential role of RUNX1 in CD4/CD8 lineage and the crucial involvement of T lymphocytes in inflammatory arthritis (described in Table 1).

3.2 Replication of overlapping associations

3.2.1 Selection of genetic regions for replication

To investigate whether newly identified overlapping associations represented true genetic associations, 7 regions from section 3.1 were selected for replication in an independent RA cohort using the Sequenom genotyping platform. Table 31 shows a summary of the association results for the 9 selected regions in the Immunochip project across the 3 diseases.

Of the 7 regions selected, 4 regions (*IGSF3/CD2, CD28/ICOS/CTLA4, EOMES and RAB5B/ERBB3/STAT2*) were associated across the 3 types of IA, 2 regions (*IL12A and MTMR3)* were associated with RA and JIA and only a single region (*LCE3B/LCE3A*) was associated exclusively with JIA and PsA in the original comparison of Immunochip data.

3.2.2 SNP assay design

A multiplex assay of 12 SNPs was designed to include index SNPs from the different diseases across all 7 of the overlapping regions selected for replication. Table 32 shows the SNPs which were included from each region in the assay. In several cases the RA disease association had already been replicated in independent cohorts (and localised using the Immunochip), therefore did not require further replication. Additionally, in some cases the index SNP could not be tolerated within the Sequenom multiplex assay so a highly correlated proxy ($r^2 > 0.9$) was included instead.

3.2.3 Subjects

Clinical and demographic features of the 3879 RA samples and 2561 controls genotyped from the United Kingdom Rheumatoid Arthritis genetics group (UKRAG) consortium are shown in Table 30. Although clinical features are listed, no stratified analysis was performed using these features.

Table 30– Demographics for 3879 cases and 2561 controls.

Age of onset	n	% cohort
16-20	52	0.81
20-29	184	2.86
30-39	429	6.66
40-49	611	9.49
50-59	686	10.65
60-69	556	8.63
70-79	296	4.60
80-89	3	0.05
Unknown	3033	47.10
Sex	**n**	**% cohort**
Male	1593	24.74
Female	3213	49.89
Unknown	1044	16.21

Table 30 shows the age and sex demographics of the cases and controls from the UKRAG cohort. The numbers of samples of percentage of cohort are shown in each case. N = number of samples.

Table 31 –Immunochip regions selected for overlap replication

Region	Chr	RA SNP	RA p	RA OR	JIA SNP	JIA p	JIA OR	PsA SNP	PsA p	PsA OR
IGSF3/CD2	1	rs798000	2.36E-05	1.08	rs12725472	0.000985	1.12	rs77421743	0.00032	1.83
LCE3B/LCE3A	1				rs11205044	0.000503	0.9	rs10888503	0.000867	0.83
CD28/ICOS/CTLA4	2	rs3087243	3.31E-09	0.90	rs231725	1.53E-05	1.15	rs11571312	3.51E-05	1.34
IL12A	3	rs4680536	1.19E-05	0.92	rs2366643	0.000463	0.9	rs13065738	0.05236	1.11
EOMES	3	rs9880772	0.000678	1.06	rs9862284	0.000622	1.11	rs733302	0.000586	1.92
RAB5B/ERBB3/STAT2	12	rs10876870	0.000149	0.93	rs1614219	0.000723	0.82	rs2020854	6.97E-05	0.61
MTMR3	22	rs77378082	4.52E-05	0.74	rs5763631	0.000618	1.39	rs74612035	0.01673	1.26

Table 31 shows the regions selected for the Immunochip overlap replication. RA = Rheumatoid arthritis, JIA = Juvenile idiopathic arthritis, PsA = Psoriatic arthritis, p = p value, OR = odds ratio

Table 32– SNPs included on Immunochip overlap replication

Region	Chr	SNP(s) genotyped		
IGSF3/CD2	1	rs10494164 (JIA proxy for rs798000)	rs77421743	
LCE3B/LCE3A	1	rs11205044		
CD28/ICOS/CTLA4	2	rs231725		
EOMES	3	rs11129295 (RA proxy for rs9880772)	rs9862284	
IL12A	3	rs587422 (RA and JIA proxy for rs4680536/rs2366643)		
RAB5B/ERBB3/STAT2	12	rs11171739 (RA proxy for rs10876870)	rs67594137 (JIA proxy for rs1614219)	rs74703593 (PsA proxy for rs2020854)
MTMR3	22	rs5763631	rs76317766 (RA proxy for rs77378082)	

Table 32 shows the 12 SNPs selected from the 7 regions in the overlap replication. The region, chromosome and SNPs genotyped are shown. Chr = chromosome, SNP = single nucleotide polymorphism.

3.2.4 Power calculations pre-QC

Prior to genotyping, power calculations were performed to predict if sufficient sample size was available to detect genetic effects. For common SNPs (MAF>0.05) it was determined that a sample size of 3879 cases and 2561 controls with effect sizes of between 1.1-1.5, the study had 22%-99% power to detect these effects, respectively. Therefore, the sample size was sufficiently powerful to detect genetic effects with larger effect sizes (OR = 1.3-1.5) but was much less well powered to detect more modest genetic effects (OR = 1.1-1.2).

3.2.5 Genotyping using the Sequenom MassARRAY platform

3879 RA cases and 2562 healthy controls were genotyped using the Sequenom MassARRAY platform.

3.2.6 Assigning SNP genotypes

SNP genotypes were assigned using the Typer 4.0 genotyping software.

Figure 28 shows examples from 2 SNPs from the multiplex assay and how genotypes were assigned. In each case homozygous major alleles are coloured blue, heterozygous samples are yellow and homozygous minor allele samples are coloured green. SNPs that could not be confidently assigned to a cluster, including negative controls, are coloured red. All SNPs were checked manually for accurate calling and deviation from HWE ($p<0.001$). At this stage, two SNPs were excluded due to poor assay performance.

Figure 28– SNP calls from Typer 4.0

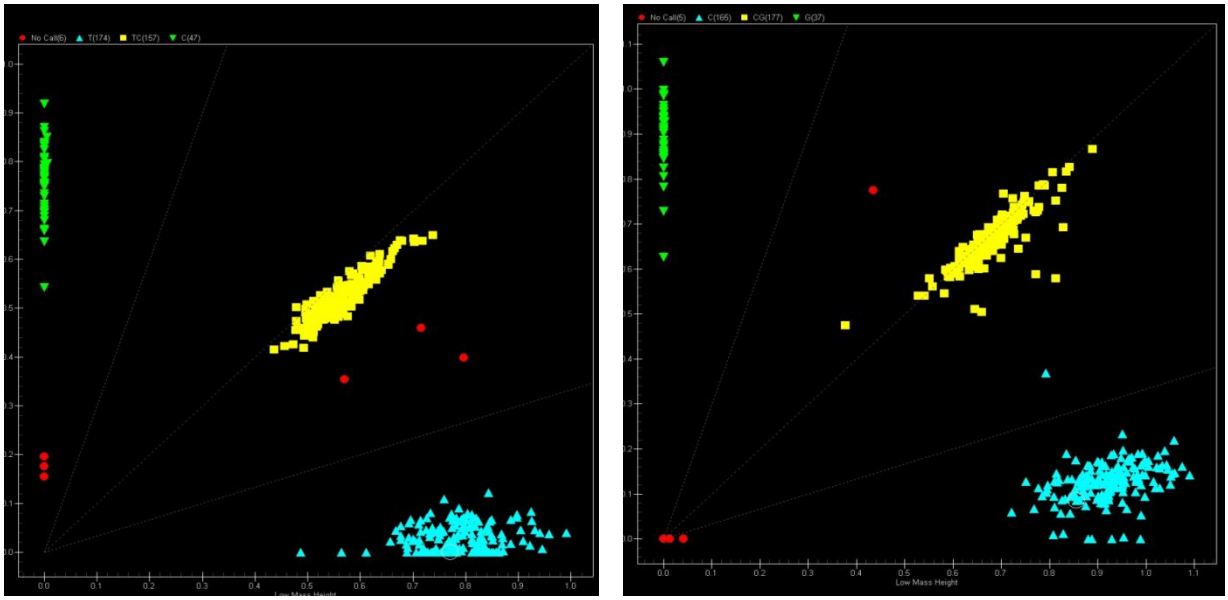


Figure 28 shows Sequenom genotyping clusters defined using Typer program. In the plots each sample is represented by a shape with blue and green representing homozygous genotypes and yellow representing heterozygous genotypes. The red shapes represent samples which could not be assigned a genotype for this SNP.

3.2.6.1 Issues with DNA quality

Once genotype calling was performed using the Sequenom Typer 4.0 genotyping software, it was noted that some samples performed much better than others did with an unusually large number of samples failing completely. As quantification and QC data on this extensive sample cohort was not available prior to genotyping, gel electrophoresis data from the PCR products of each assay was examined. A random number of samples from across each plate were selected and gel electrophoresis performed for both assays. With strong amplification a series of bright bands are expected of uniform size with missing bands indicating failure to amplify or sample contamination. In Figure 29 good quality DNA samples are shown with strong bands present across the gel. This indicates that the PCR amplification of the DNA was successful and this is reflected in the high call rate for samples and low frequency of failing SNPs. In comparison in Figure 30 a low quality DNA plate is shown with a large number of weaker and missing bands. This is reflected in the lower call rates present in the genotyping with a larger number of failing SNPs compared to Figure 29. Loss of sample quality may be the consequence of a number of factors including evaporation through faulty lids, long term storage in water or sample contamination.

Figure 29 – High quality DNA gel and genotyping

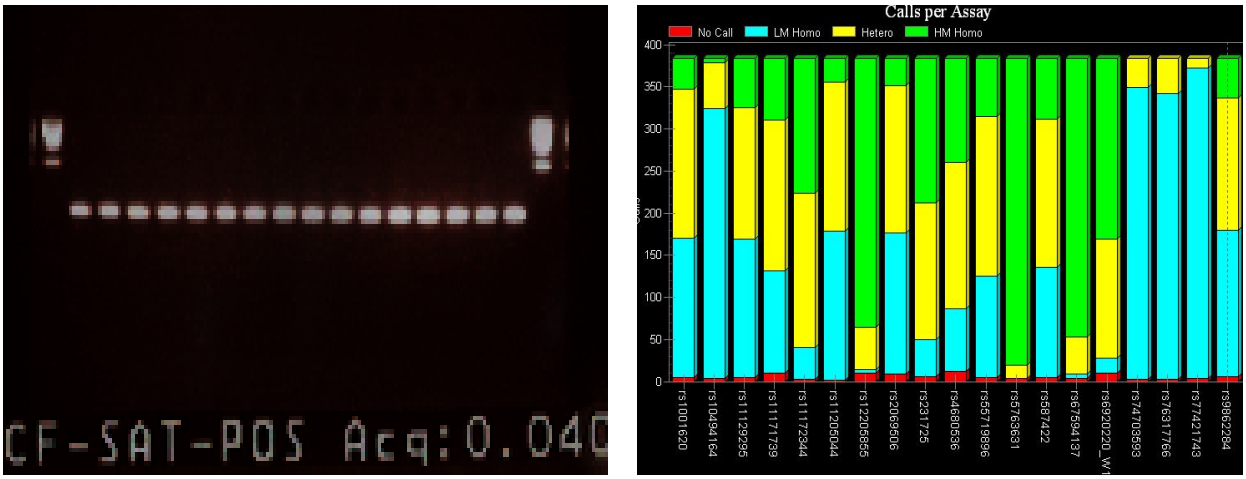


Figure 29 shows the high rate of successful genotyping when high quality DNA was used. In the left hand plot the gel electrophoresis from the high quality DNA is shown, with uniform PCR products present in each well. In the right hand plot the proportion of SNP calls are shown for each SNP which are detailed on the x-axis. For each SNP the proportion of homozygous calls are shown in blue/yellow, the heterozygous calls shown in green and the no calls shown in red.

Figure 30– Low quality DNA gel and genotyping

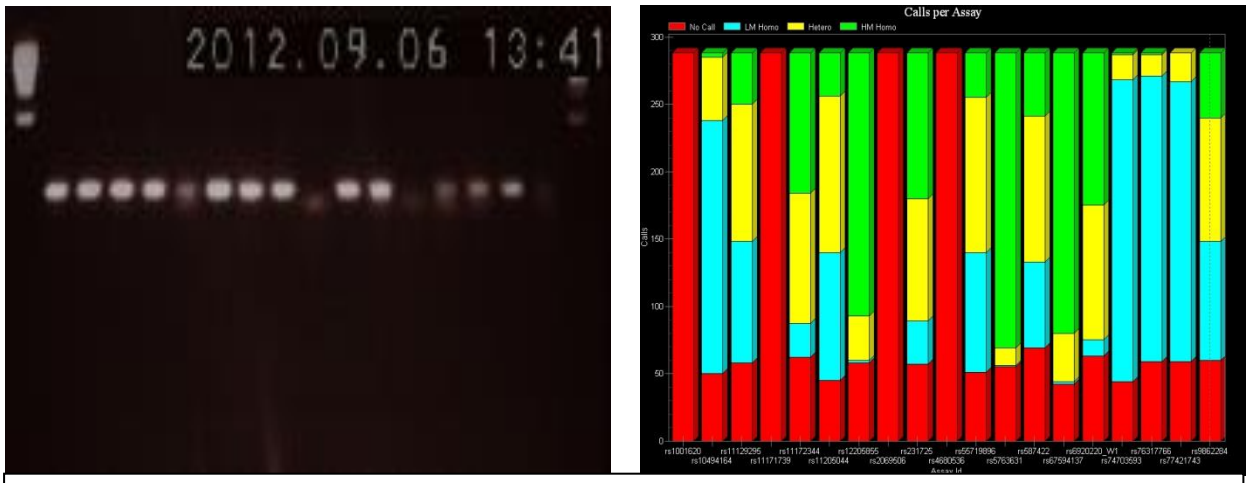


Figure 30 shows the low rate of successful genotyping when low quality DNA was used. In the left hand plot the gel electrophoresis from the low quality DNA is shown, with less uniform and missing PCR products in each well. In the right hand plot the proportion of SNP calls are shown for each SNP, which are detailed on the x-axis. For

3.2.7 Sample and SNP QC

12 SNPs were genotyped in 3879 cases and 2561 controls in total. Due to the low DNA quality of some samples (See section 3.2.6), two stages of QC were employed. This was to achieve an accurate representation of the true genotyping quality and prevent skewing of results by completely failing samples. Initially in stage I, low quality DNA samples and SNP assays which completely failed genotyping (<50% call rate) were removed from analysis. In stage II, more stringent QC was performed on the remaining samples. After SNP and sample QC, data for 9 SNPs genotyped in 2595 cases and 1636 controls were available for association analysis. Both QC stages I and stage II are summarized in Figure 37.

Figure 31 – SNP and sample QC summary

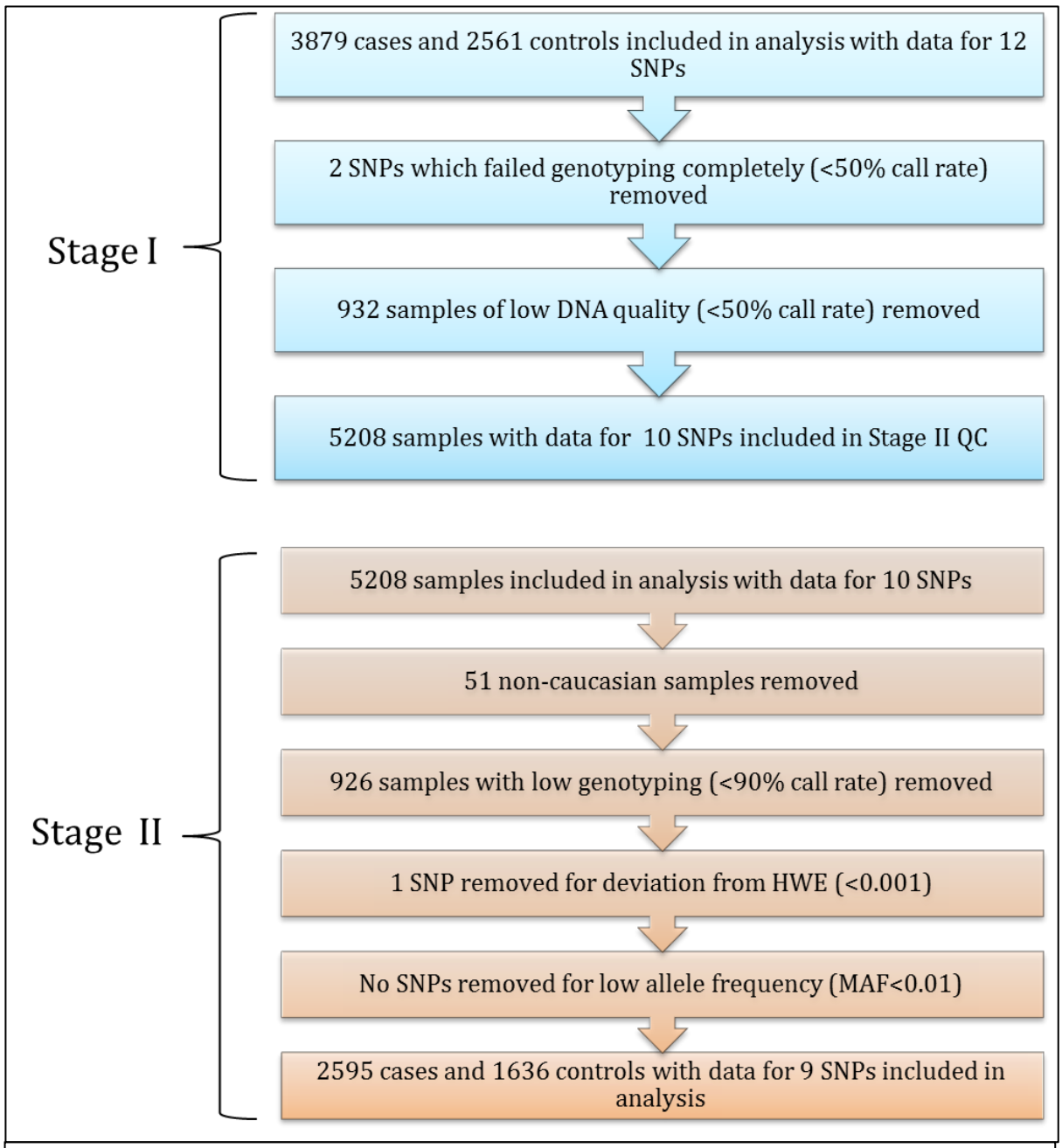


Figure 31 shows a flowchart of the SNP and sample QC employed. In the first stage samples and SNPs which had very low call rates were removed. In the second stage samples and SNPs were removed using the following criteria: non-caucasian samples; low genotyping calls, deviation from HWE; low MAF. HWE = Hardy Weinberg Equilibrium; MAF = Minor allele frequency.

3.2.8 Post-QC power calculations

Post sample QC, power calculations were performed to predict if sufficient sample size was available to detect genetic effects once a large number of samples were removed. For common SNPs (MAF>0.05) it was determined that a sample size of 2595 cases and 1636 controls with effect sizes of between 1.1-1.5 had a power of 16%-99% to detect these effects respectively. In this number of samples there was 99% power to detect a large effect size of 1.5 but the loss of samples did result in a large reduction in power to detect modest effects.

3.2.9 Association testing

The allelic association results from the 9 SNPs are shown in Table 33. The 2 SNPs in bold text (rs231725 and rs576361) represent those with an allelic association p value of less than 0.05. Although these SNPs were not the RA index SNPs from the Immunochip study, rs231725 was associated with RA in the Immunochip study ($p=1.37x10^{-5}$) with an OR in the same direction but this could not be investigated for rs5763631 as this SNP was removed during the RA Immunochip QC described in Figure 21 . The SNP rs231725 is an intergenic SNP which lies downstream of the Cytotoxic T-Lymphocyte Antigen 4 (*CTLA4*) gene whilst rs5763631 is an intronic SNP which lies within the Myotubularin-related protein 3 (*MTMR3*) gene. No other significant associations were identified although several SNPs show trend towards association ($p<0.07$), which with greater power might be identified at a significant level.

Table 33 – Allelic association results from Overlap Replication

CHR	SNP	Position	Gene	Frequency cases	Frequency controls	P	OR	95% CI
1	rs77421743	117249473	IGSF3/CD2	0.01485	0.01743	0.3554	0.84	0.6017-1.2
1	rs11205044	152593437	LCE3B/LCE3A	0.3491	0.3298	0.06827	1.09	0.9935-1.196
2	**rs231725**	**204740675**	**CD28/ICOS/CTLA4**	**0.3513**	**0.3188**	**0.0021**	**1.16**	**1.054-1.27**
3	rs9862284	27800325	EOMES	0.3634	0.3435	0.06297	1.09	0.9953-1.196
3	rs587422	159729805	IL12A	0.4196	0.4159	0.7398	1.02	0.9289-1.109
12	rs67594137	56374318	RAB5B/ERBB3/STAT2	0.08192	0.0842	0.711	0.97	0.8283-1.137
12	rs74703593	56585248	RAB5B/ERBB3/STAT2	0.0453	0.0422	0.4996	1.08	0.8686-1.335

22	rs76317766	30312987	MTMR3	0.05558	0.05291	0.5983	1.05	0.8679-1.279
22	**rs5763631**	**30347633**	**MTMR3**	**0.0166**	**0.02509**	**0.00644**	**0.66**	**0.483-0.8902**

Table 33 shows the association results of the overlap replication. For each SNP, position, gene, allele frequency in cases and controls, p value, odds ratio and 95% confidence interval are shown. p = p value, OR = odds ratio, 95% CI = 95% confidence interval.

3.3 RUNX1 replication and fine mapping

3.3.1 Defining the region for fine mapping

Although the *RUNX1* region represented a novel overlapping region it was not included on the overlap replication as it was selected as a region for further genetic analysis. This was performed by fine mapping the region to identify all genetic effects within the region.

As the overlapping SNP in the *RUNX1* region (rs9979383) lies within two peaks of high recombination it is likely that the association in this region is located between these points. This region was selected for fine mapping and covered 150kb of chromosome 21 (35600kb-35750kb in hg18/NCBI36). Figure 32 shows an LD plot showing the extent of SNP LD around rs9979383 and the region selected for fine mapping.

Figure 32 – *RUNX1* region selected for fine mapping

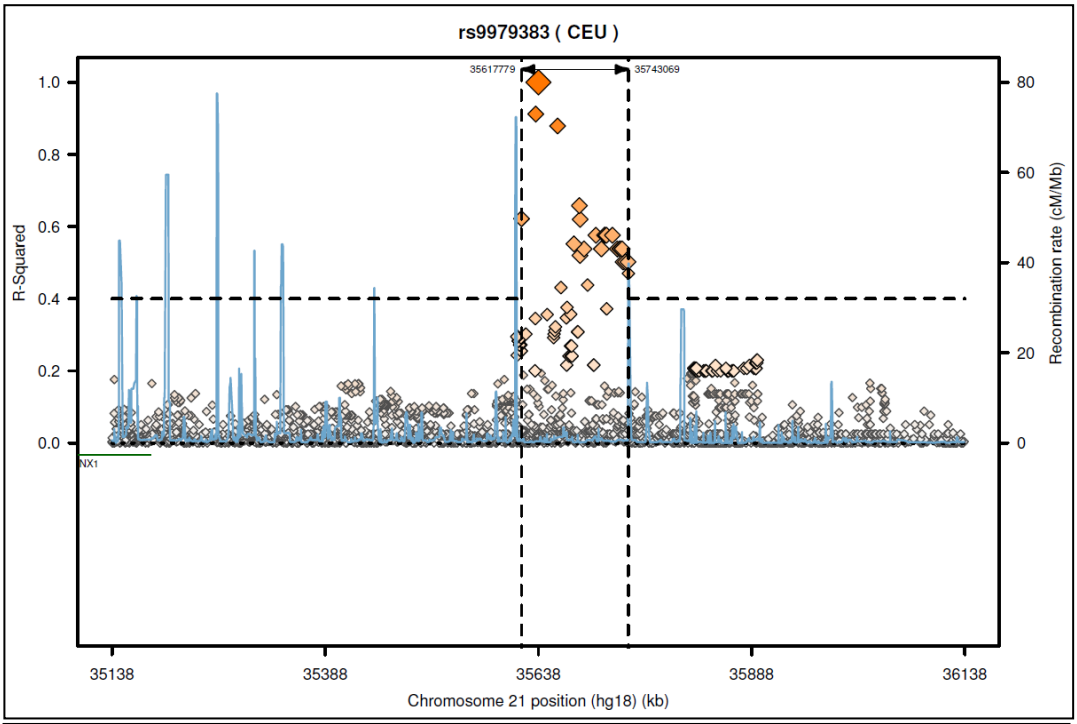


Figure 32 shows LD around the selected SNP rs9979383. The x-axis represents the base position on chromosome 21 whilst the y-axis represents the level of LD using r^2. Based on this the region selected for fine mapping is shown by the vertical dotted line. LD = linkage disequilibrium.

3.3.2 Calculating *RUNX1* SNP coverage on the Immunochip array

The tag SNP coverage of the *RUNX1* fine mapping region on the Immunochip array was calculated for both common (MAF > 0.05) and low frequency (MAF > 0.01) variants.

Table 34 shows the total number of SNPs in the Utah residents with Northern and Western European ancestry (CEU) 1000 genomes (July 2010 release) within the *RUNX1* fine mapping region and the calculated number and percentage of these variants captured on the Immunochip array at r^2 = 0.8 and r^2=0.9. At 6 and 7% the Immunochip array provided very low coverage of this region so further fine mapping was necessary to identify the true genetic effect in the region.

Table 34 – SNP capture of the RUNX1 region on the Immunochip array

MAF	r^2	Total SNPs	SNPs on Immunochip array	Coverage (%)
>0.05	0.8	322	24	7
>0.01	0.8	600	37	6
>0.05	0.9	322	23	7
>0.01	0.9	600	36	6

Table 34 shows the coverage of the RUNX1 region on the Immunochip array. Coverage was calculated for different MAFs and using different r^2 values. MAF = minor allele frequency.

3.3.3 Subjects

Clinical and demographic features of the 3491 RA samples and 2359 controls genotyped from the United Kingdom Rheumatoid Arthritis genetics group (UKRAG) were identical to those described in section 3.2.3.

3.3.4 Pre-QC power calculations

Prior to genotyping, power calculations were performed to predict if sufficient sample size was available to detect genetic effects. For common (MAF>0.05) SNPs it was determined that a sample size of 3491 cases and 2359 controls had a power of 20%-99% to detect effect sizes of between 1.1-1.5 . Therefore this sample size was sufficiently powerful for detecting genetic effects with larger effect sizes (OR = 1.3-1.5) but was much less well powered to detect smaller genetic effects (OR = 1.1-1.2)

3.3.5 Tag SNP selection and assay design

51 common (MAF >0.05) tag SNPs from CEU 1000 genomes (July 2010 release) were selected for genotyping using the Tagger function in Haploview v.4.2. Figure 33 shows the output from the tagger function. Table 35 shows the total number of SNPs from the (CEU) 1000 genomes (July 2010 release) within the *RUNX1* fine mapping region, the calculated number and percentage of these variants captured by the 51 SNPs at $r^2 = 0.8$ and $r^2=0.9$ and the percentage increase of coverage in comparison to the Immunochip Array.

To include all 51 SNPs in the genotyping assay, 2 genotyping assays were designed. Assay 1 contained 30 SNPs and assay 2 contained 21 SNPs.

Table 35 – SNP capture of the *RUNX1* fine mapping region with 2 assays

MAF	r^2	Total SNPs in region	SNPs on Assay	Coverage (%)	Increase (%)
>0.05	0.8	322	51	82	75
>0.01	0.8	600	51	44	38
>0.05	0.9	322	51	68	61
>0.01	0.9	600	51	36	30

Table 35 shows the coverage of the RUNX1 region with 2 assays. Coverage was calculated for different MAFs and using different r^2 values. MAF = minor allele frequency.

Figure 33– Haploview tagger results for 51 SNPs

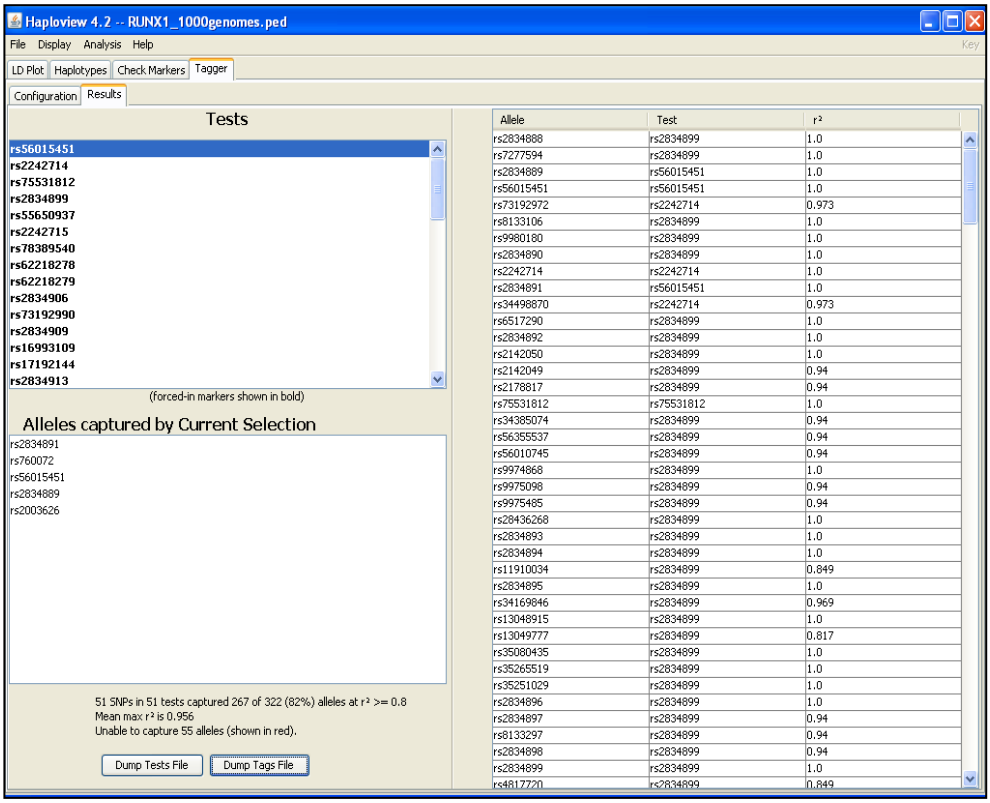


Figure 33 shows the Haploview tagger results for the *RUNX1* region. In the tests box (top left) all SNPs, which were tested, are listed with a single SNP highlighted in blue. In the box below (bottom left) the SNPs correlated with this highlighted SNP (r^2>0.8) are shown. In the right hand column a pairwise analysis of each of all the SNPs tested is shown, with the LD displayed in r^2.

3.3.6 Genotyping using the Sequenom MassARRAY system

3491 RA cases and 2359 healthy controls were genotyped using the Sequenom MassARRAY platform.

3.3.7 Calling SNP genotypes

SNP genotypes were called using the Typer 4.0 genotyping software (Sequenom). Figure 34 shows the calling of SNP rs9979383 from assay 1 and rs13052307 from assay 2. In each case homozygous major alleles are coloured blue, heterozygous samples are yellow and homozygous minor allele samples are coloured green. SNPs that could not be accurately assigned to a cluster were coloured red. All SNPs were checked manually for accurate calling and deviation from HWE ($p<0.001$). At this point 3 SNPs were excluded from further analysis due to poor assay performance.

Figure 34 – SNP calls from Typer 4.0

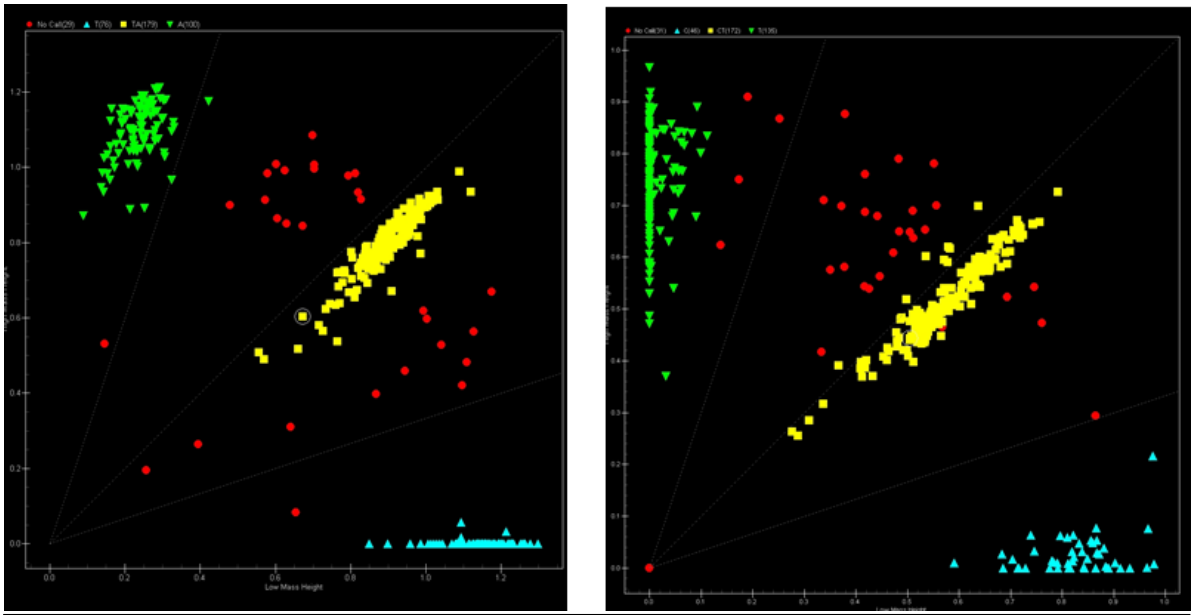


Figure 34 shows Sequenom genotyping clusters defined using Typer program. In the plots each sample is represented by a shape with blue and green representing homozygous genotypes and yellow representing heterozygous genotypes. The red shapes represent samples which could not be assigned a genotype for this SNP.

3.3.7.1 Issues with DNA quality

Once genotype calling was performed using the Sequenom Typer 4.0 genotyping software, it was noted that some samples performed much better than others with an unusually large number of samples failing completely across both assays. As this genotyping was performed on identical samples to the Immunochip overlap replication (section 3.2.3), the PCR product gels and genotyping results were examined as before. Figure 35 and Figure 36 are examples of high quality and low quality DNA samples respectively.

Figure 35– High quality DNA gel and genotyping

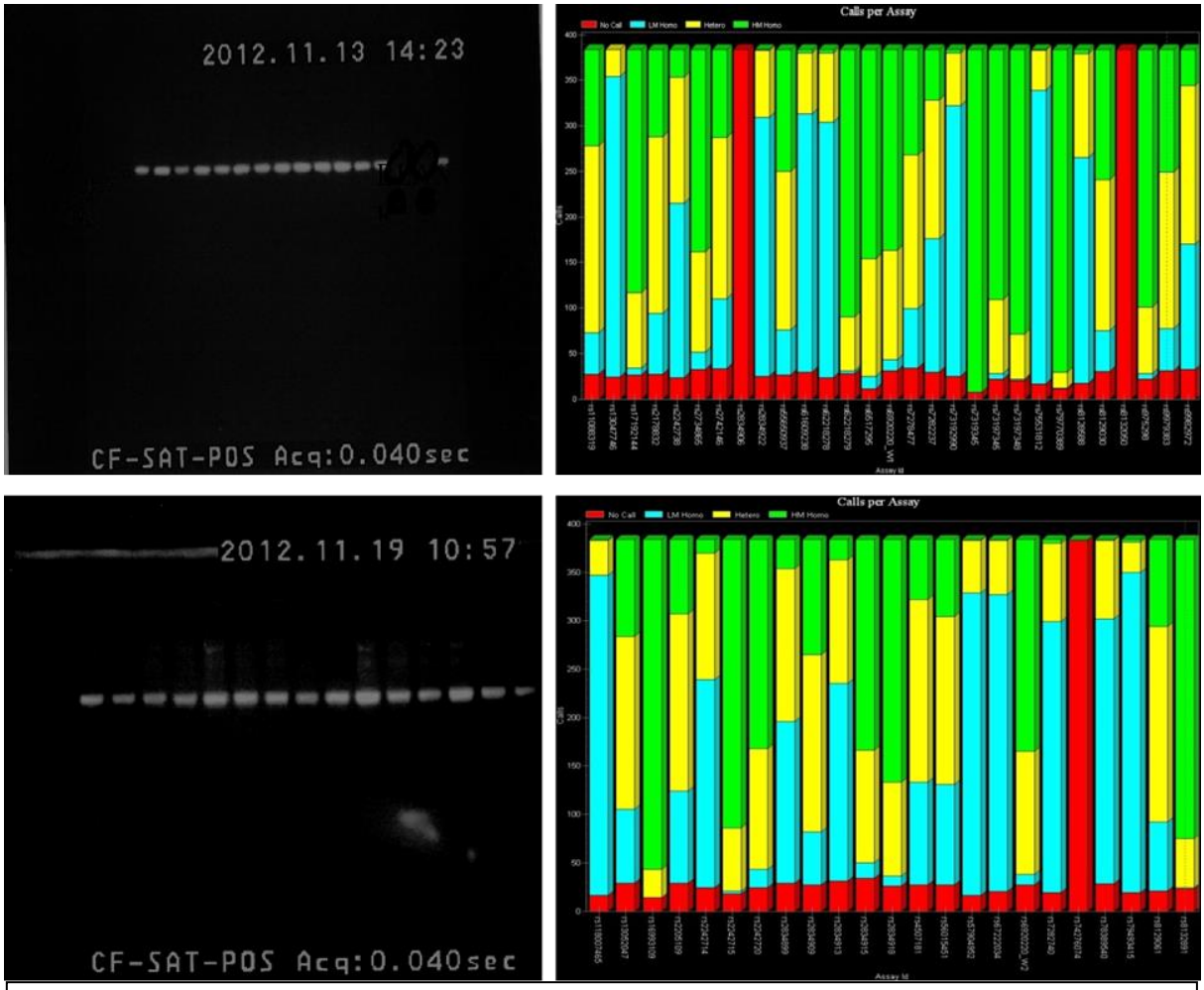


Figure 35 shows the high rate of successful genotyping when high quality DNA was used. In the left hand plot the gel electrophoresis from the high quality DNA is shown, with uniform PCR products present in each well. In the right hand plot the proportion of SNP calls are shown for each SNP which are detailed on the x-axis. For each SNP the proportion of homozygous calls are shown in blue/yellow, the heterozygous calls shown in green and the no calls shown in red.

Figure 36 – Low quality gel and genotyping

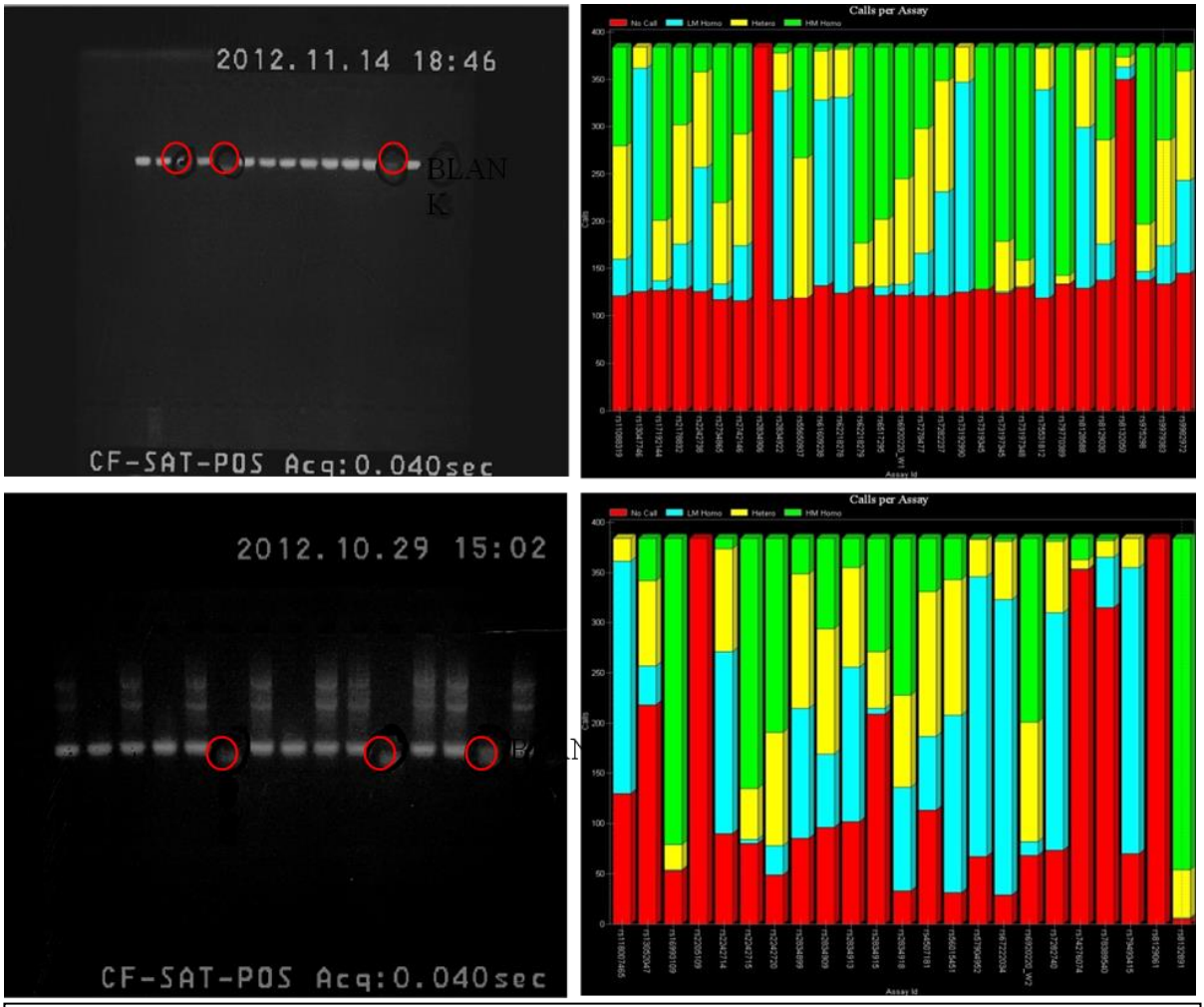


Figure 36 shows the low rate of successful genotyping when low quality DNA was used. In the left hand plot the gel electrophoresis from the low quality DNA is shown, with less uniform and missing PCR products in each well. In the right hand plot the proportion of SNP calls are shown for each SNP, which are detailed on the x-axis. For each SNP the proportion of homozygous calls are shown in blue/yellow, the heterozygous calls shown in green and the no calls shown in red.

3.3.8 Sample and SNP QC

51 SNPs were genotyped in 2359 cases and 3491 controls in total. Due to the low DNA quality of some samples (section 3.2.6), 2 stages of QC were employed. This was to achieve an accurate representation of the true genotyping quality and prevent skewing of results by completely failing samples. Initially in stage I, low quality DNA samples and SNP assays which completely failed genotyping (<50% call rate) were removed from analysis. In stage II, more stringent QC was performed on the remaining samples. After SNP and sample QC data for 42 SNPs

genotyped in 2359 cases and 1877 controls were available for association analysis. Both QC stages I and stage II are summarized in Figure 37.

Figure 37 – Genotyping QC stage I and stage II

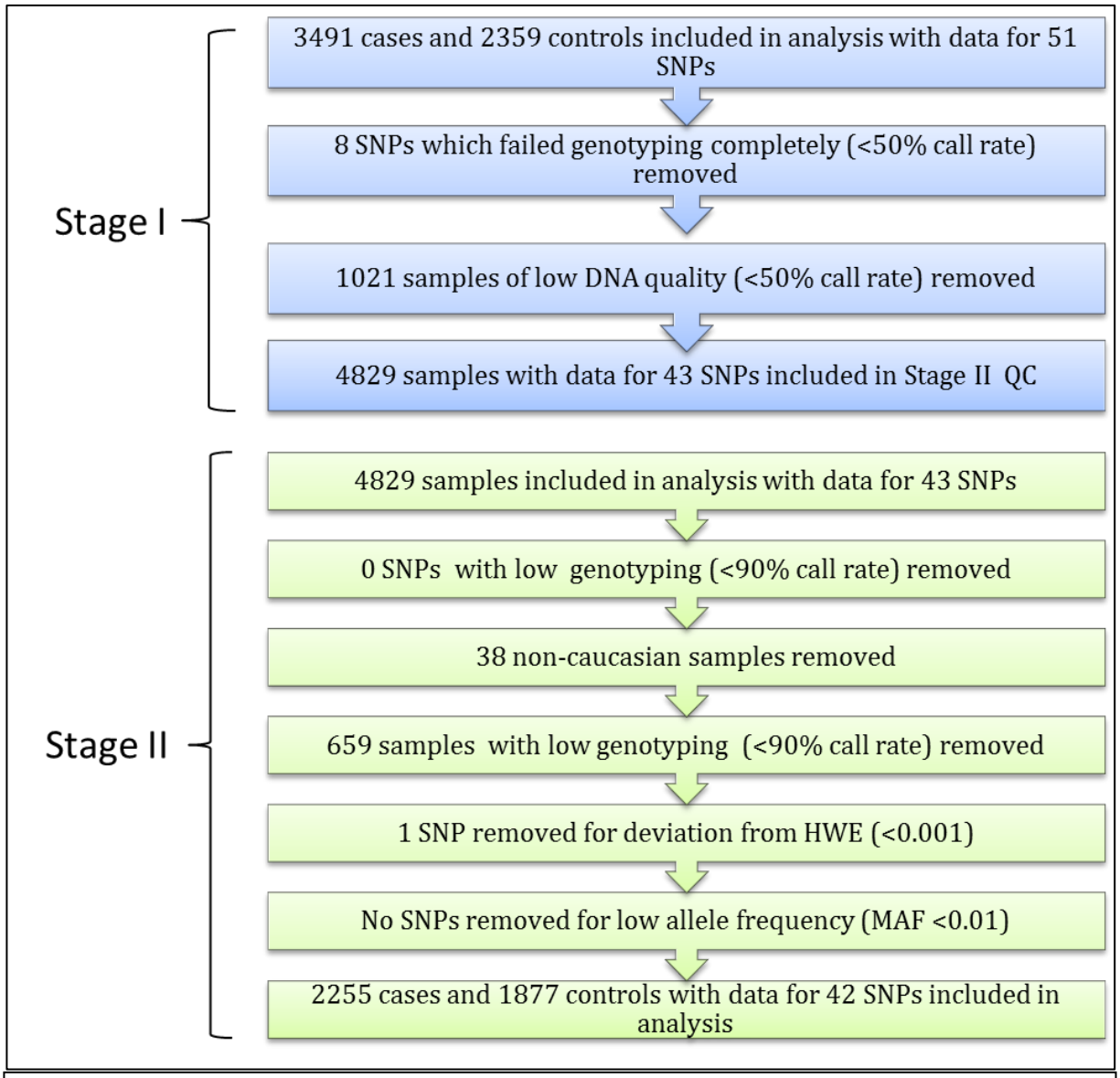


Figure 37 shows a flowchart of the SNP and sample QC employed. In the first stage samples and SNPs which had very low call rates were removed. In the second stage samples and SNPs were removed using the following criteria: non-Caucasian samples; low genotyping calls, deviation from HWE; low MAF. HWE = Hardy Weinberg Equilibrium; MAF = Minor allele frequency.

3.3.9 Post-QC power calculations

Post sample QC, power calculations were performed to predict if sufficient sample size was available to detect genetic effects once a large number of samples were removed. For common SNPs (MAF>0.05) it was determined that a sample size of 2255 cases and 1877 controls had a power of 16%-99% to detect effect sizes of between 1.1-1.5, respectively.

3.3.10 Association testing

The allelic association results from the *RUNX1* fine mapping genotyping are shown in Table 36. The index SNP from the Immunochip study rs9979383 is shown in bold and represents the most significantly associated SNP with an allelic p value of 0.026. No other significant associations ($p<0.05$) were identified indicating that the association in this region is most likely localised to rs9979383 in RA. Figure 38 shows a SNP association plot for the fine mapped region, with rs9979383 annotated as the most associated SNP. Table 37 shows the p values, odds ratios (ORs) and 95% confidence intervals (CI) for rs9979383 in the fine mapping study compared to the Immunochip results for RA, JIA and PsA. Figure 39 compares the OR of rs9979383 in this study with the OR from the RA, JIA and PsA Immunochip studies. For this study the OR for rs9979383 is identical to that of the Immunochip study, strengthening the likelihood that the signal is localised to this SNP or one in high LD with it.

Table 36 – Allelic association testing for *RUNX1* fine mapping genotyping

SNP	Pos	Frequency cases	Frequency controls	P	OR	95% CI
rs56015451	36675655	0.4566	0.4662	0.386	0.9622	0.8819-1.05
rs2242714	36677687	0.2358	0.2278	0.3957	1.046	0.9434-1.159
rs75531812	36679979	0.07059	0.06592	0.4047	1.076	0.9053-1.28
rs2834899	36685788	0.308	0.3058	0.8285	1.01	0.9197-1.11
rs2242715	36688993	0.112	0.1062	0.4065	1.061	0.9229-1.219
rs62218278	36690137	0.1048	0.1114	0.3393	0.9342	0.8124-1.074
rs62218279	36690797	0.09321	0.09288	0.9597	1.004	0.8644-1.166
rs73192990	36693635	0.1012	0.102	0.9122	0.992	0.8593-1.145
rs2834909	36695176	0.4414	0.4295	0.2811	1.05	0.9612-1.146
rs16993109	36698195	0.04037	0.04118	0.8541	0.9797	0.7869-1.22
rs17192144	36702985	0.1752	0.1666	0.3017	1.063	0.9469-1.192
rs2834913	36709666	0.2648	0.2593	0.5722	1.029	0.9319-1.136
rs79770389	36709836	0.01713	0.01801	0.7611	0.95	0.6828-1.322
rs8129030	36712588	0.3536	0.3738	0.05709	0.9162	0.8373-1.003
rs2242720	36714156	0.26	0.2471	0.1791	1.071	0.9691-1.183
rs9979383	**36715761**	**0.3571**	**0.3809**	**0.02662**	**0.903**	**0.8251-0.9882**

rs8132891	36720105	0.07571	0.07772	0.7321	0.972	0.8259-1.144
rs67222034	36724842	0.07684	0.07834	0.7995	0.9792	0.8325-1.152
rs975298	36726185	0.1533	0.1442	0.2482	1.075	0.9511-1.214
rs13047746	36726967	0.04007	0.04064	0.8958	0.9854	0.7905-1.228
rs2834918	36731196	0.1969	0.1905	0.4672	1.042	0.9331-1.163
rs2834922	36738670	0.1148	0.1086	0.3835	1.064	0.9258-1.222
rs61609238	36740965	0.1195	0.1123	0.3113	1.073	0.9365-1.228
rs9982972	36744868	0.3676	0.3485	0.07131	1.087	0.9928-1.19
rs4507181	36748497	0.4301	0.4508	0.0595	0.9194	0.8424-1.003
rs13052047	36748863	0.476	0.4572	0.08984	1.078	0.9883-1.177
rs118007465	36752815	0.05778	0.05818	0.9385	0.9927	0.8245-1.195
rs2742146	36753728	0.4762	0.4872	0.3208	0.957	0.8775-1.044
rs2898237	36763769	0.323	0.3372	0.1698	0.9375	0.855-1.028
rs73197345	36770120	0.1371	0.1406	0.6455	0.971	0.8565-1.101
rs7282740	36773278	0.1129	0.109	0.5769	1.04	0.9059-1.194
rs73197348	36773339	0.07609	0.07752	0.8078	0.98	0.833-1.153
rs7278477	36773939	0.4351	0.4253	0.3704	1.041	0.9536-1.136

rs2734865	36779287	0.2286	0.2305	0.8422	0.9896	0.8926-1.097
rs57904952	36785648	0.08111	0.08378	0.6604	0.9653	0.8244-1.13
rs6517295	36797477	0.1988	0.1954	0.6969	1.022	0.9158-1.141
rs8128588	36802005	0.173	0.1653	0.3763	1.057	0.9349-1.195
rs11088319	36803737	0.4113	0.3945	0.1216	1.072	0.9816-1.172
rs2242738	36809021	0.2888	0.2998	0.2784	0.9486	0.8622-1.044
rs2178832	36821898	0.4412	0.4552	0.203	0.9448	0.8657-1.031
rs79493415	36824271	0.05501	0.05401	0.8423	1.02	0.8422-1.234

Table 36 shows the association results of the overlap replication. For each SNP, position, gene, allele frequency in cases and controls, p value, odds ratio and 95% confidence interval are shown. The SNP in bold represents the SNP which has a $p<0.05$. p = p value, OR = odds ratio, 95% CI = 95% confidence interval.

Table 37– Association statistics for rs9979383 compared to the Immunochip Study

Study	P	OR	95% CI
UKRAG	0.026	0.903	0.8251-0.9882
RA I-chip	8.06×10^{-7}	0.91	0.8323-0.9321
JIA I-chip	1.05×10^{-7}	0.85	0.8-0.93
PsA I-chip	6.12×10^{-4}	0.854	0.7727-0.9432

Table 37 compares the association statistics for rs9979383 from the fine mapping study to the results from RA, JIA and PsA in the Immunochip study.

Figure 38 – Locus zoom plot of the *RUNX1* region

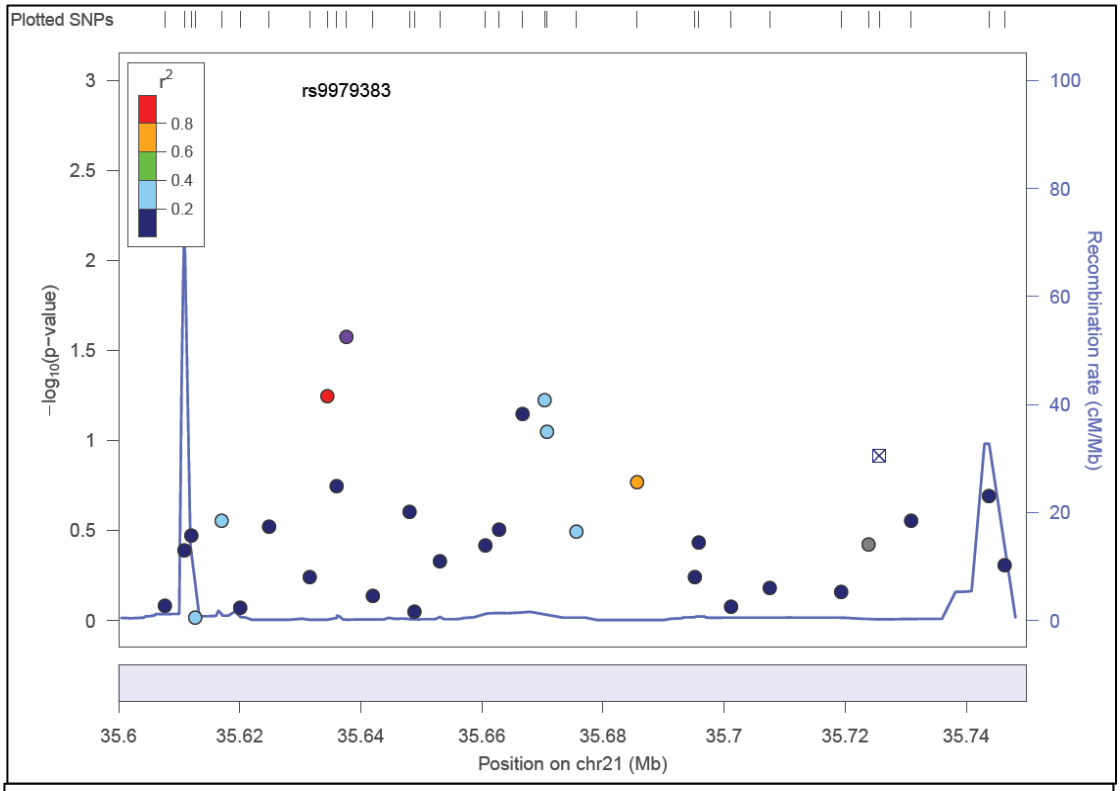


Figure 38 shows the association plot for the fine mapped *RUNX1* region. The x-axis represents the base position on chromosome 21 whilst the y-axis represents the $-\log_{10}$ of the p value. Each SNP is represented by a dot with the most associated SNP annotated. Colour coding of each SNP is determined by LD with the most associated SNP. SNP = single nucleotide polymorphism, LD = linkage disequilibrium.

Figure 39– Odds ratio forest plot

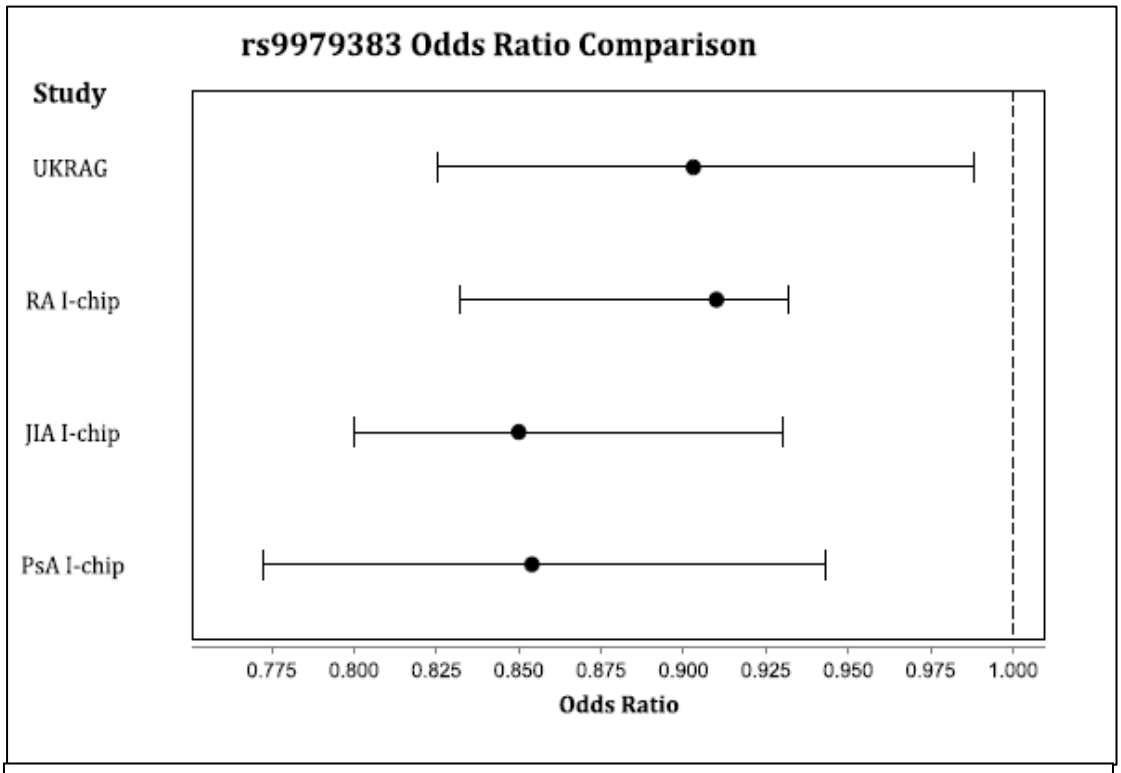


Figure 39 is a forest plot representing an odds ratio comparison between the UKRAG, RA Immunochip, JIA Immunochip and PsA Immunochip studies for rs9979383. In this plot the x-axis represents the odds ratio value. In each horizontal row the value of the odds ratio is represented by a dot whilst the bars represent the 95% confidence intervals.

3.3.11 Identification of multiple effects in the *RUNX1* region

Conditional logistic regression was performed to identify if rs9979383 was responsible for all genetic effects within the associated region. When logistic regression was performed conditioning on rs997383 no additional significant associations ($p<0.001$) were identified (Table 36). The association plot in Figure 40 shows the absence of any significant effects when conditioning on rs9979383.

Table 38**– Conditional logistic regression results**

SNP	Position	P	OR	95% CI
rs56015451	36675655	0.2871	0.9526	0.871-1.042
rs2242714	36677687	0.4238	1.044	0.9398-1.159
rs75531812	36679979	0.3201	1.092	0.9179-1.3
rs2834899	36685788	0.647	1.023	0.9294-1.125
rs2242715	36688993	0.4434	1.056	0.9182-1.215
rs62218278	36690137	0.213	0.9145	0.7944-1.053
rs62218279	36690797	0.3984	1.072	0.9117-1.262
rs73192990	36693635	0.4421	0.9432	0.8125-1.095
rs2834909	36695176	0.7658	0.9832	0.8792-1.099
rs16993109	36698195	0.52	0.9295	0.7438-1.161
rs17192144	36702985	0.8113	1.015	0.8972-1.149
rs2834913	36709666	0.6029	0.9713	0.8703-1.084
rs79770389	36709836	0.5932	0.9178	0.6699-1.257
rs8129030	36712588	NA	NA	NA-NA
rs2242720	36714156	0.6463	1.026	0.9182-1.147
rs9979383	36715761	NA	NA	NA-NA
rs8132891	36720105	0.4075	0.9308	0.7856-1.103
rs67222034	36724842	0.4166	0.933	0.7892-1.103
rs975298	36726185	0.6294	1.032	0.9074-1.174
rs13047746	36726967	0.6692	0.9519	0.7593-1.193
rs2834918	36731196	0.9546	0.9965	0.8834-1.124
rs2834922	36738670	0.8161	1.017	0.8808-1.175

rs61609238	36740965	0.7061	1.027	0.8924-1.183
rs9982972	36744868	0.1864	1.071	0.9673-1.186
rs4507181	36748497	0.3064	0.9429	0.8425-1.055
rs13052047	36748863	0.2761	1.059	0.9554-1.173
rs118007465	36752815	0.7317	0.967	0.7982-1.172
rs2742146	36753728	0.9694	0.9979	0.896-1.111
rs2898237	36763769	0.9702	0.9973	0.8641-1.151
rs73197345	36770120	0.3343	0.9387	0.8254-1.067
rs7282740	36773278	0.9109	1.008	0.8732-1.164
rs73197348	36773339	0.5355	0.949	0.8043-1.12
rs7278477	36773939	0.6017	1.025	0.935-1.123
rs2734865	36779287	0.4602	0.9604	0.8628-1.069
rs57904952	36785648	0.3945	0.9312	0.7903-1.097
rs6517295	36797477	0.6826	1.024	0.9154-1.145
rs8128588	36802005	0.8035	1.017	0.8926-1.158
rs11088319	36803737	0.2813	1.053	0.9584-1.158
rs2242738	36809021	0.06526	0.9081	0.8197-1.006
rs2178832	36821898	0.2587	0.949	0.8666-1.039
rs79493415	36824271	0.9994	1	0.8243-1.213
rs7282237	36824423	0.2776	0.9504	0.867-1.042

Table 38 shows the results from the conditional logistic regression analysis, when conditioning on rs9979383. For each SNP, the base position, p value, odds ratio and 95% confidence interval when conditioned on rs9979383 is shown.

Figure 40– Association plot when conditioned on rs9979383 showing no independent effects

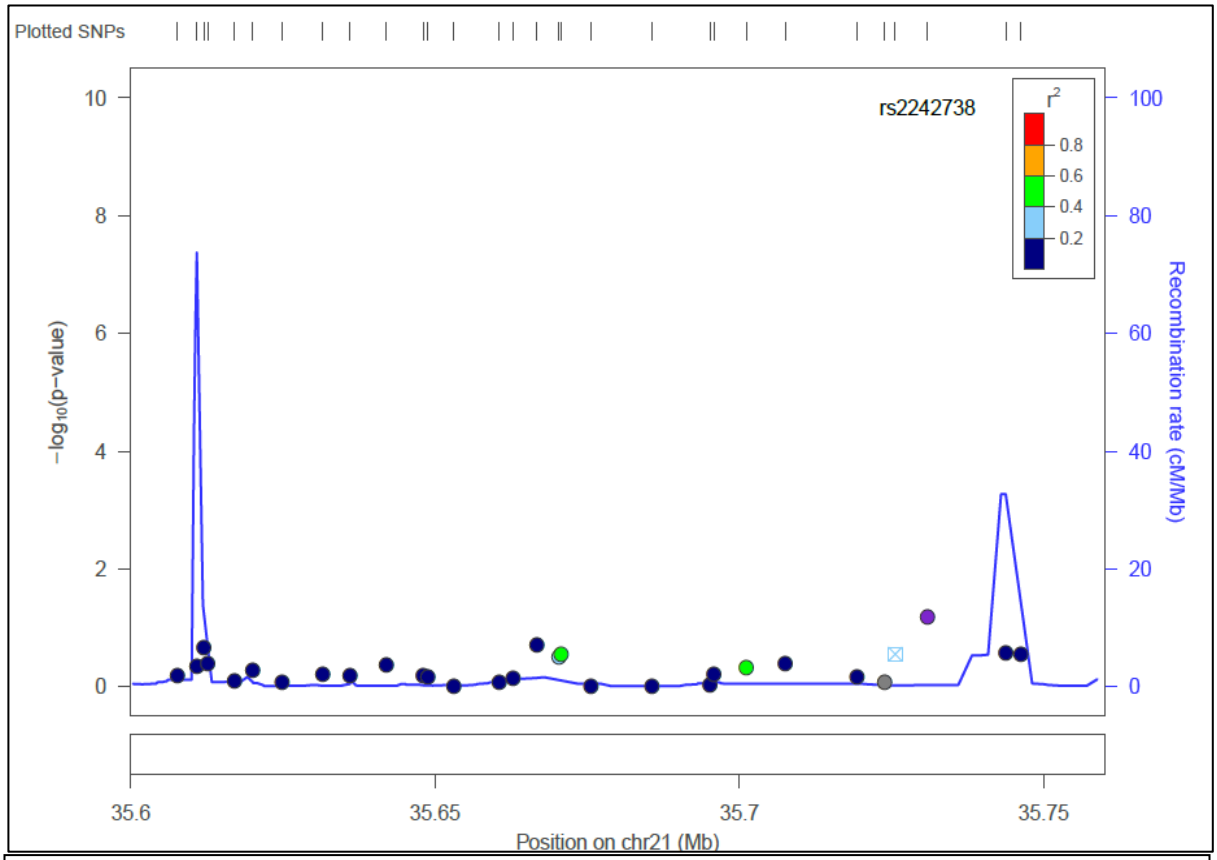


Figure 40 shows the association plot for the *RUNX1* region when conditional analysis had been performed. The x-axis represents the base position on chromosome 21 whilst the y-axis represents the $-\log_{10}$ of the p value when conditioned on rs9979383. Each SNP is represented by a dot.

3.4 Functional analysis of the *RUNX1* region

3.4.1 eQTL analysis of the *RUNX1* region in whole blood

3.4.1.1 Subjects

Clinical and demographic features of the 75 healthy controls selected for eQTL analysis from the national repository healthy volunteers (NRHV) cohort are shown in Table 37. All subjects provided both a matched blood DNA sample and blood RNA sample for analysis. Once extracted, all 75 DNA samples had a yield of greater than 25ng/µl required for Taqman genotyping.

Table 39– Demographics for 75 subjects

Age	n	% cohort
20-29	6	8.00
30-39	12	16.00
40-49	18	24.00
50-59	13	17.33
60-69	15	20.00
70-79	10	13.33
80-89	1	1.33
Sex	**n**	**% Cohort**
Male	18	25.71
Female	57	76

Table 39 shows the age and sex demographics of the healthy controls from the NRHV cohort. The numbers of samples of percentage of cohort are shown in each case. N = number of samples.

3.4.1.2 SNP genotyping using Taqman allelic discrimination assays

Genotyping was performed on 75 healthy controls using a Taqman allelic discrimination assay.

3.4.1.3 Calling of genotypes using the Quant studio RT-PCR software

Genotype calls for rs9979383 were performed using the QuantStudio™ 12K Flex Real-Time PCR Software. Figure 41 shows the SNP clustering of the samples into 3 distinct groups. Of the 75 samples genotyped, 73 genotyped successfully for rs9979383 with Table 38 summarizing the genotypic distribution of the 73 remaining samples.

Figure 41– Genotype calls using Quant studio RT-PCR software

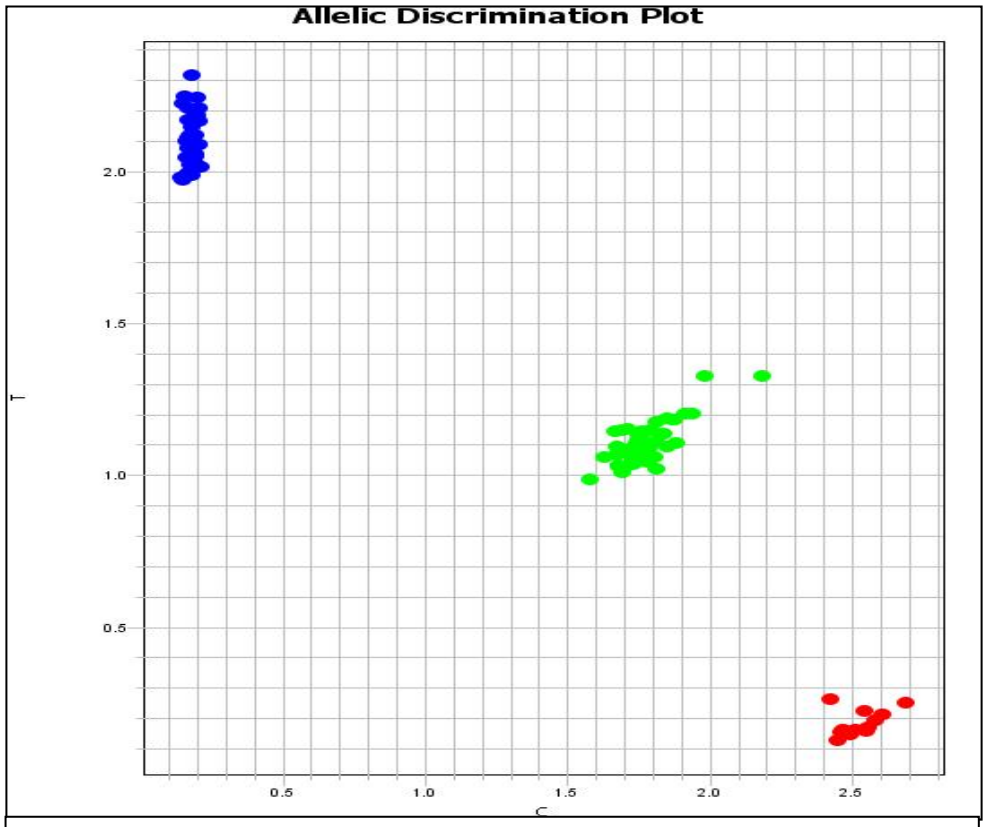


Figure 41 shows the genotype calls using the Quant studio RT-PCR software. The x and y values represent values of fluorescence which are used by the software to make genotype calls. Homozygous genotyping calls are represented in red and blue, whilst heterozygous genotyping calls are represented in green.

Table 40- Genotypic distribution of genotype calls in healthy controls

Genotype call	n	% cohort
Homozygous C/C	10	13.70
Heterozygous C/T	35	47.95
Homozygous T/T	28	38.36

Table 40 shows the genotype distribution at rs9979383 in the NRHV cohort. n= number of subjects

RUNX1 is a highly variable gene with 19 splice variants recorded in the Ensembl genome browser 72 as shown in Figure 41. As 14 of these variants share common exons 5-9 a gene expression assay was designed to capture these exons. In addition Glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) and Beta actin were selected as endogenous controls due to a recommendation by the manufacturer and 2 assays were designed to capture expression of these genes. Total RNA was extracted from the 75 subjects as described in the Appendix.

Figure 42– ENSEMBL gene browser showing *RUNX1* splice variants

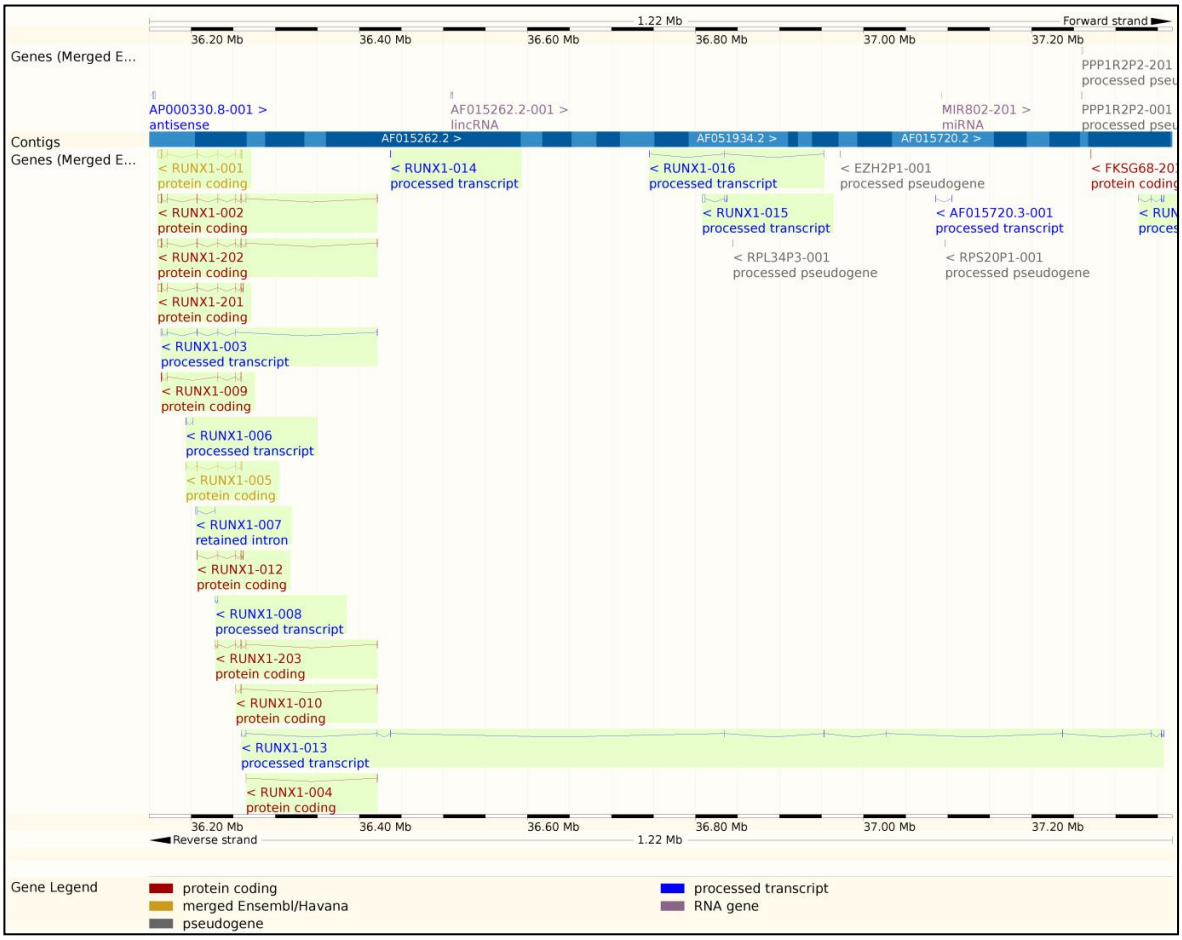


Figure 42 shows the identified splice variants for *RUNX1* according to ENSEMBL (http://www.ensembl.org/index.html). Each transcript is colour coded in relation to its function with protein coding transcripts in red and processed transcripts in blue.

3.4.1.6 Total RNA quality control

All samples were run on the Nanodrop N1000 and the Bioanalyzer to calculate total yield and RIN values. In total, 72 samples had a total yield of >1µg RNA and a RIN of greater than 5, therefore were suitable for cDNA conversion. Figure 43 shows an example Electropherogram and gels from 2 of the healthy control samples. The presence of strong 18s and 28s peaks on the Electropherogram (left) and two strong 18s and 28s bands on the gel (right) without the presence of

additional artefacts is indicative of a good quality RNA sample. In both these cases the RIN of each sample was greater than 8.

Figure 43– Electropherogram and gels from 2 healthy controls samples.

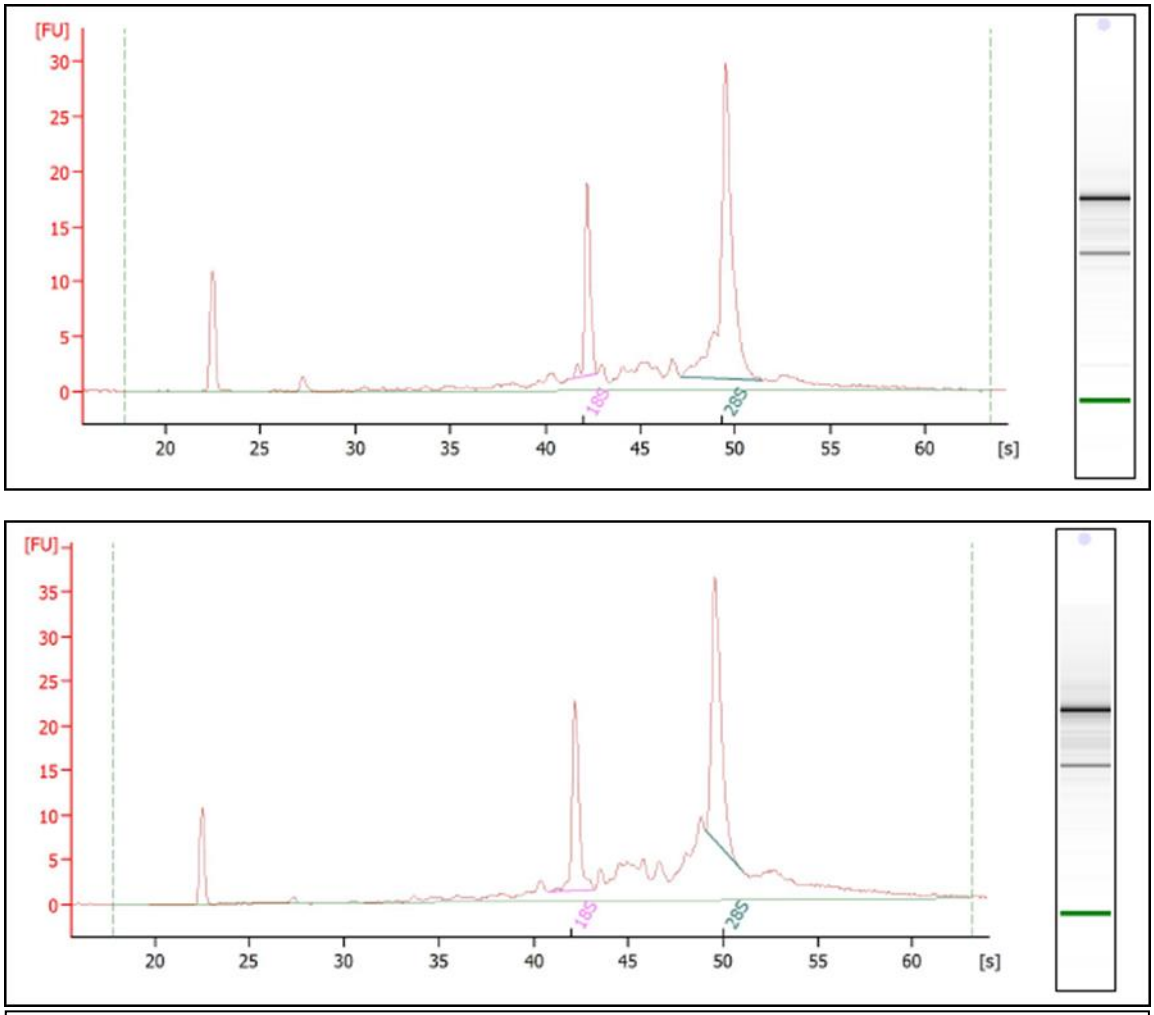


Figure 43 shows the electropherogram peaks from 2 RNA samples. On the x-axis is the time in seconds from the initiation of the run whilst the y-axis represents the fluorescence units. The presence of defined 18S and 28S peaks combined with low background noise is used to determine integrity and concentration of RNA samples.

3.5.1.7 cDNA conversion of RNA samples

72 samples of 50ng/µl concentration were successfully converted to cDNA for gene expression analysis.

3.4.1.8 Gene expression analysis of *RUNX1* and endogenous controls

Gene expression analysis for *RUNX1* and 2 endogenous controls (*GAPDH* and *ACTNB*) was performed for 72 samples in triplicate. Figure 44 shows an amplification plot for these 3 genes showing that all 3 genes are expressed in cDNA from these whole blood samples. Figure 44 is a summary of the QC performed and the number of samples, which were available for eQTL analysis. In total 70 samples with both genotype and gene expression data were available for analysis.

Figure 44 – Amplification plot for *RUNX1* and housekeeping genes in whole blood

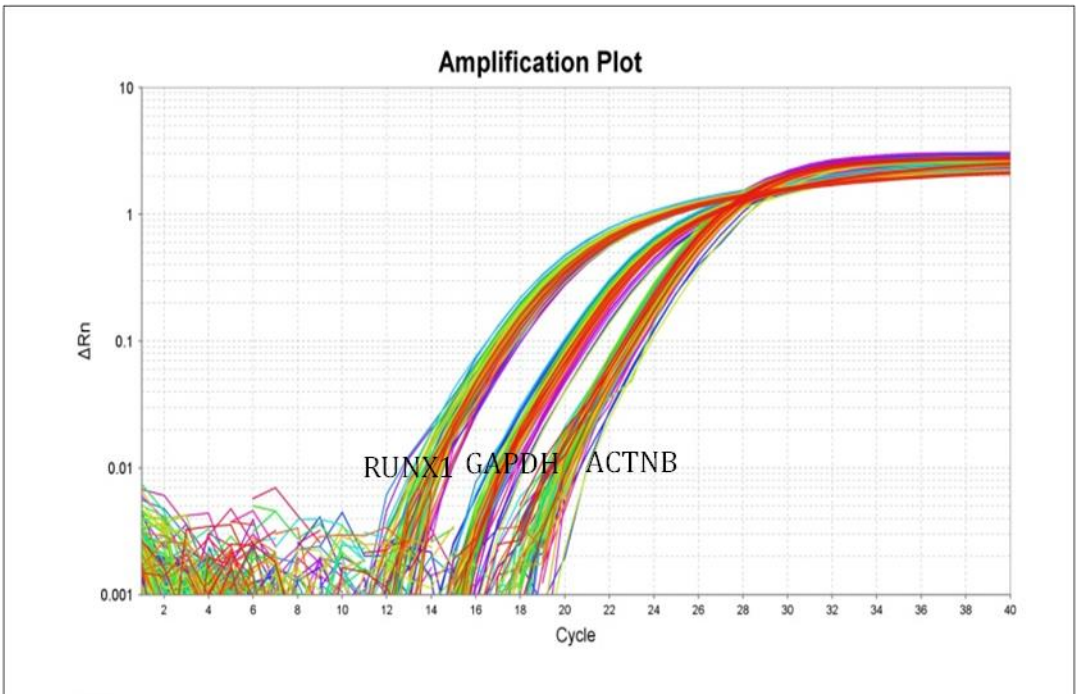


Figure 44 shows the amplification plot for *RUNX1*, *GAPDH* and *ACTNB* in whole blood, confirming expression of these genes in this tissue type. The x-axis represents the number of PCR cycles that have occurred whilst the y-axis represents the ΔRn fluorescent signal at each time point. The ΔRn represents the magnitude of signal generated, which can be used to determine the quantity of PCR product in each sample.

Figure 45– QC summary for eQTL analysis

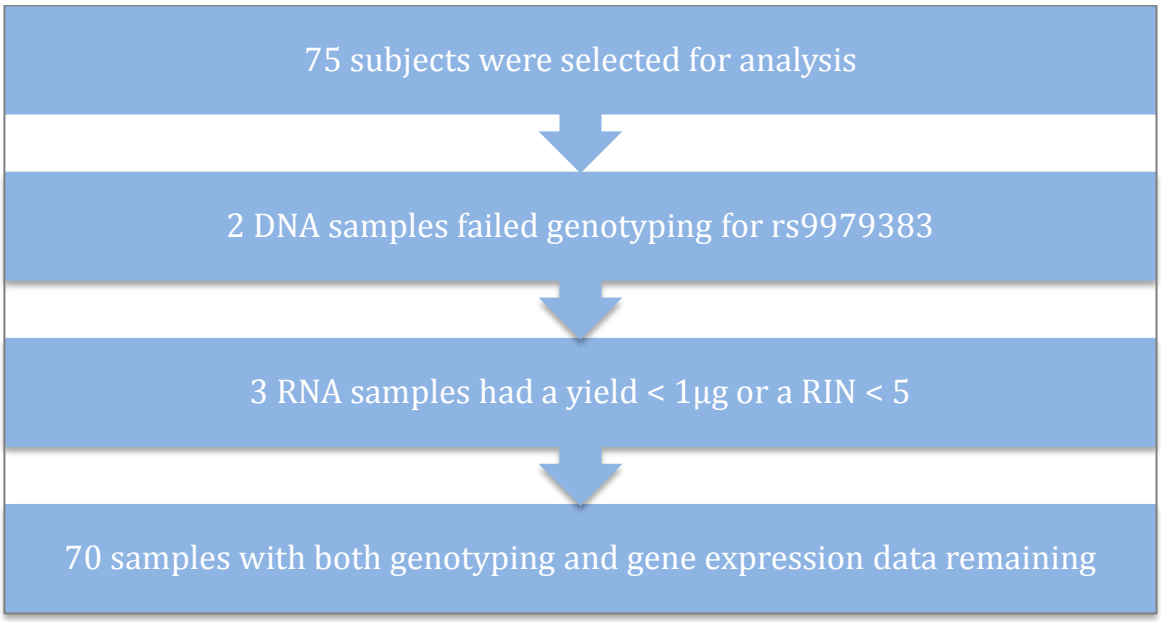


Figure 45 shows the QC stages employed for the eQTL analysis and the samples that passed each stage of QC. These included removing samples, which failed genotyping, and samples, which had a low RIN value. Only samples that had RNA and DNA were taken forward for analysis. RIN = RNA integrity number.

3.4.1.9 Whole blood eQTL analysis

In the 70 samples available for analysis, *RUNX1* genotype and expression was successfully measured but when correlated with genotype at rs9979383, no significant eQTL was identified ($p = 0.92$).
Figure 46 shows *RUNX1* gene expression normalised to *GAPDH* and *ACTNB* stratified by genotype at rs9979383.

3.4.1.10 Summary of eQTL analysis in whole blood

These results show that there is no significant evidence for an eQTL between rs9979383 and *RUNX1* in whole blood in healthy controls. This may be due to a lack of eQTL in the *RUNX1* region, that rs9979383 correlates with expression of

another gene or that this eQTL is masked in whole blood and therefore requires further investigation using cell specific methods.

Figure 46– eQTL analysis of *RUNX1* region in whole blood

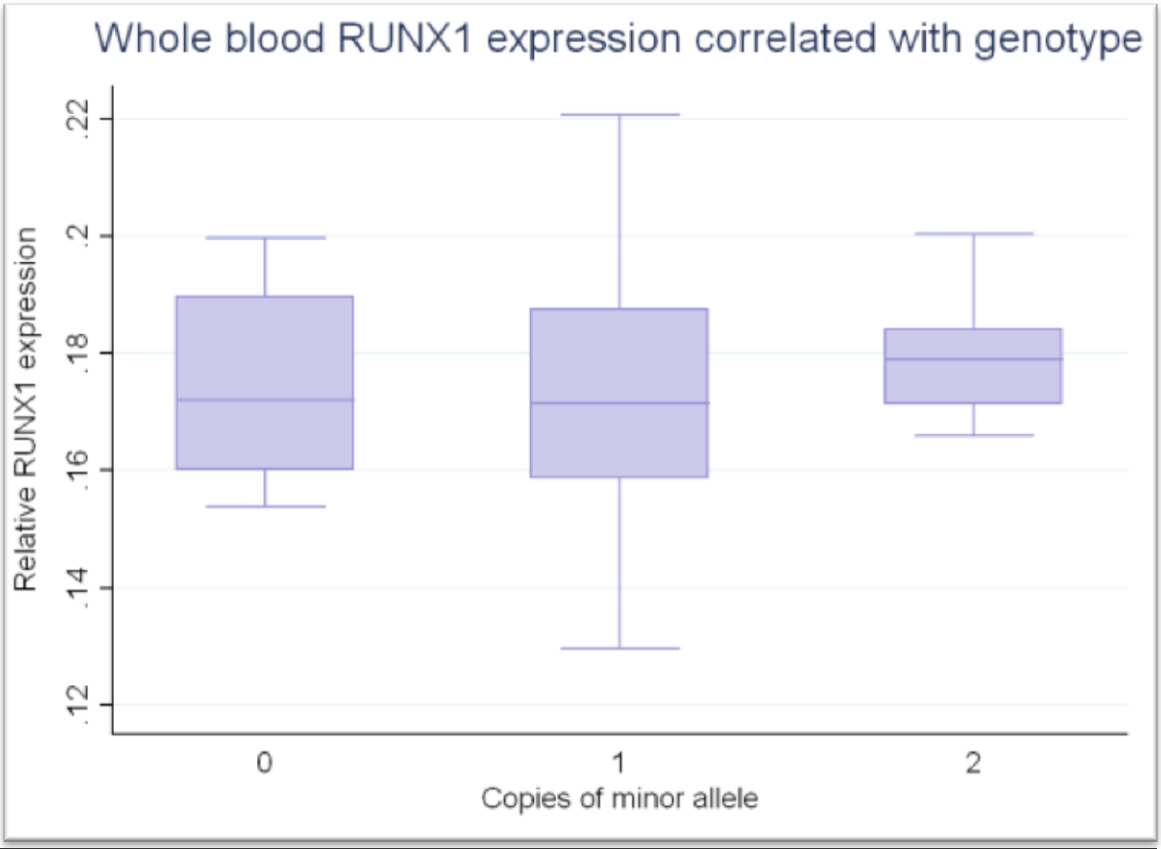


Figure 46 shows the eQTL analysis between rs9979383 and *RUNX1*. The x-axis represents the genotype groups: 0 indicates major allele homozygote; 1 heterozygote and 2 indicates minor allele homozygote. The y-axis represents the relative *RUNX1* expression (normalised to 2 endogenous controls).

3.4.2 eQTL analysis of the *RUNX1* region in T lymphocytes

3.4.2.1 Subjects

Clinical and demographic features of the 23 healthy controls selected for cell specific eQTL analysis from the National repository healthy volunteers (NRHV) cohort are shown in Table 41. All subjects provided both a matched blood DNA sample and blood sample for cell specific expression analysis. Although sex and age was available for these samples, no stratified analysis was performed.

Table 41– Demographics of 23 samples from the NRHV cohort

Age	n	% cohort
20-29	4	17.39
30-39	5	21.74
40-49	9	39.13
50-59	5	21.74
Sex	**n**	**% cohort**
Male	8	34.78
Female	15	65.22

Table 41 shows the age and sex demographics of the healthy controls from the NRHV cohort. The numbers of samples of percentage of cohort are shown in each case. n = number of samples.

3.4.2.2 Genotyping of samples

Genotypes for rs9979383 were obtained from the Taqman genotype data generated in section 3.4.1. Complete genotypes for 23 of the samples were available for analysis. Table 42 describes the genotype distribution for these samples.

Table 42– Genotype distribution for 23 healthy controls

Genotype call	n	% cohort
Homozygous C/C	2	8.69
Heterozygous C/T	11	47.83
Homozygous T/T	9	39.13
No genotype	1	4.35

Table 42 shows the genotype distribution at rs9979383 in the NRHV cohort. n= number of subjects

3.4.2.3 Sample collection for PBMC extraction

For peripheral blood mononuclear cells (PBMCs) extractions 2x10ml EDTA tubes of peripheral blood were sampled from 23 subjects and PBMCs extracted. Due to the large number of blood samples required for this study, blood drawing was performed in sessions across 26 days with all extractions being performed within 2 hours of blood sampling.

3.4.2.4 Cell count and viability checks

Cell counts and viability checks were performed for each of the 23 PBMC samples. Figure 47 shows the total PBMC yield from each sample. Samples which are highlighted in red had noticeably low blood volumes (less than 15ml total) so limited PBMC yields were expected from these samples. All samples had a PBMC yield of between $1.31x10^7$ and $4.2x10^7$ cells with an average yield of $2.43x10^7$ PBMCs per 20ml blood sample, indicating some variability in the number of PBMCs obtained from healthy controls. All PBMC samples had a cell viability of greater than 90% and therefore were suitable for cell separation in section 3.4.2.6.

Figure 47– PBMC yield from healthy control bloods

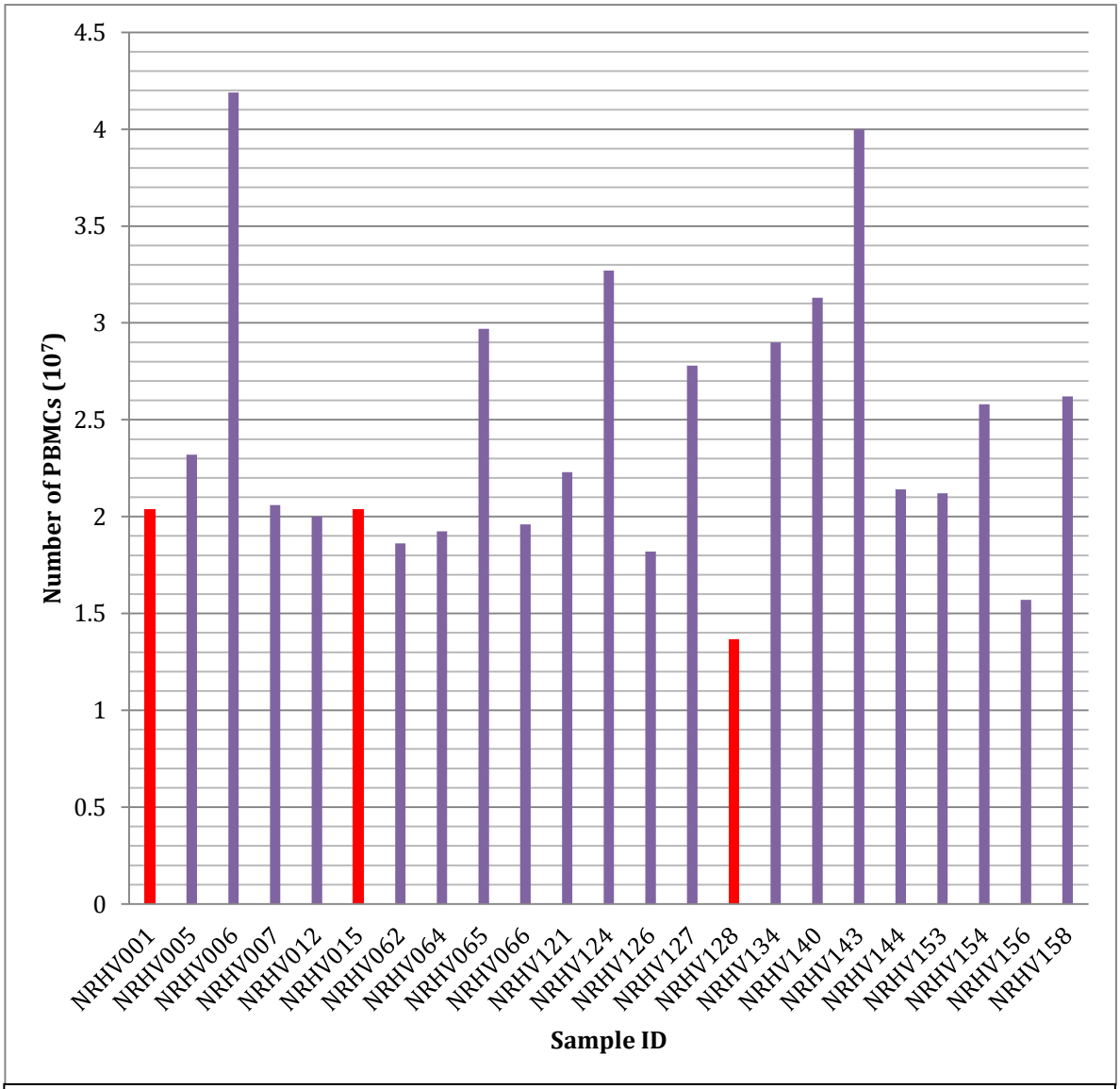


Figure 47 shows the PBMC yield from each of the 23 samples. The x-axis represents the sample ID whilst the y-axis represents the yield of PBMCs (10^7). Low volume blood samples (less than 15mls) are highlighted in red. PBMC = peripheral blood mononuclear cell.

3.4.2.5 Cryopreservation and thawing of PBMCs

As samples were collected in sampling sessions across 26 days, cryopreservation was utilised to maintain PBMC yield and viability for cell separation. All samples were cryopreserved for no more than 31 days and were thawed in batches of 6 samples to minimise experimental variability between samples. Post-cryopreservation, samples were immediately processed for cell separation to optimise yield and viability of cells.

3.4.2.6 Separation of PBMCs into T lymphocyte subsets

In order to obtain both CD4+ and CD8+ T lymphocytes from each sample a triple phase cell separation was performed using the positive selection strategy described in section 2.4.2.6.1. Separations were performed in groups of 6 samples in parallel to minimise batch effects. Primarily, CD8+ T lymphocytes were separated from each thawed PBMC sample. CD8+ lymphocytes were successfully separated from all 23 samples and a cell count performed. Figure 49 shows the CD8+ cell yield from each PBMC sample. The average cell count obtained was $1.37x10^6$ CD8+ cells per sample.

Post CD8+ separation, a CD14+ separation was performed on the CD8- negative fraction whilst the CD8+ fraction was collected for flow cytometry and total RNA extracted. This was performed to remove CD14+ monocytes which may contaminate the CD4+ lymphocyte population during the final positive selection. This is due to their low expression of the CD4 marker. The CD14+ fraction was then discarded and a final CD4+ positive selection performed using the CD14- fraction.

In the final separation CD4+ cells were obtained from all 23 samples and a cell count performed. Figure 48 shows the cell count for the CD4+ cells obtained from each sample. The average cell count obtained from each sample was $1.46x10^6$

CD4+ cells. The CD4- fraction was discarded and the CD4+ fraction collected for flow cytometry and total RNA extraction.

Figure 48– CD4+ lymphocyte yield

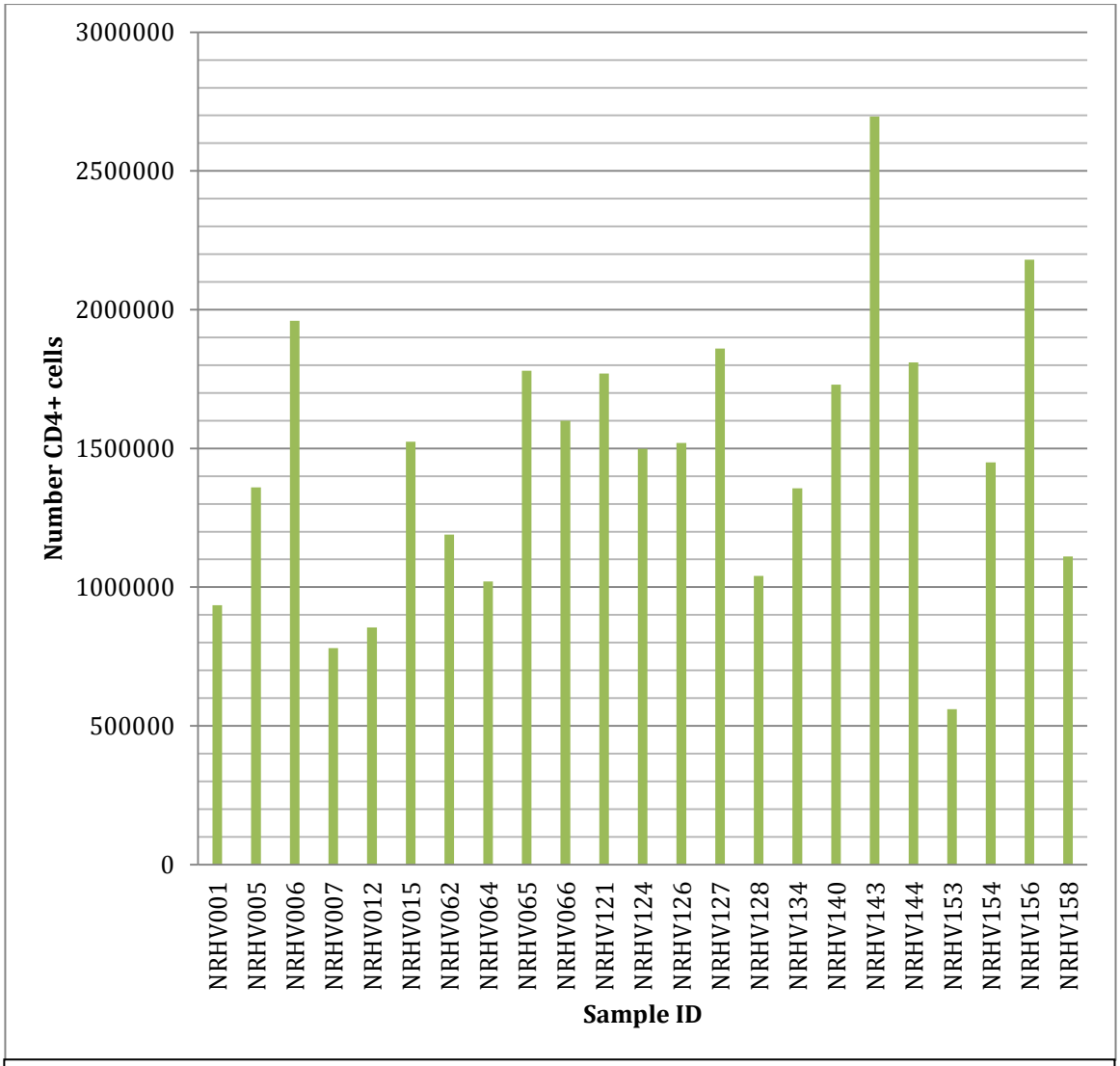


Figure 48 shows the CD4+ lymphocyte yield from each of the 23 samples. The x-axis represents the sample ID whilst the y-axis represents the yield of CD8+ lymphocytes.

Figure 49 CD8+ lymphocyte yield

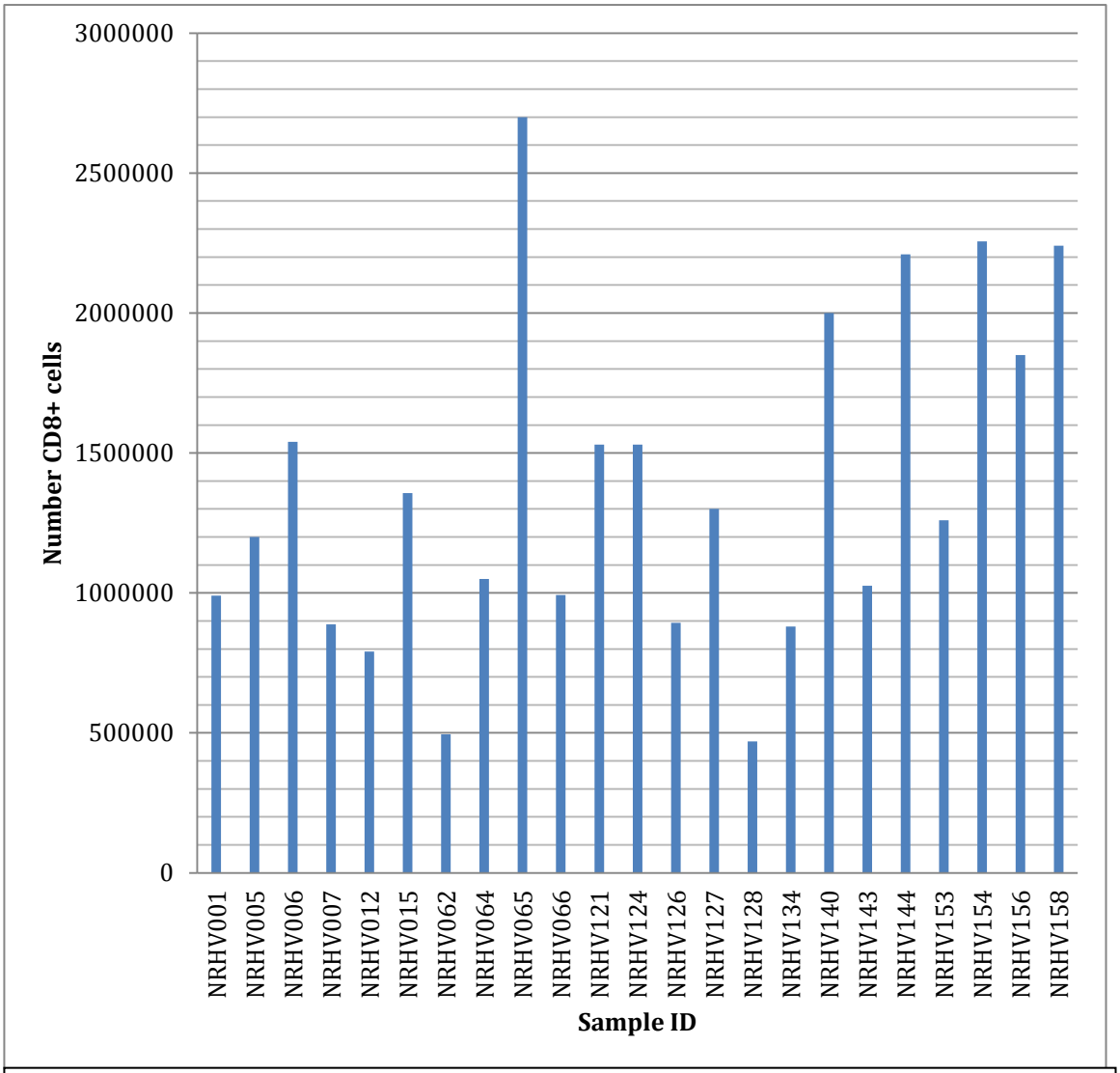


Figure 59 shows the CD4+ lymphocyte yield from each of the 23 samples. The x-axis represents the sample ID whilst the y-axis represents the yield of CD8+ lymphocytes.

3.4.2.7 Assessment of viability and cell purity using flow cytometry

Post cell separation, flow cytometry was used to assess the purity of cell populations obtained and ensure they were homogenous enough for accurate gene expression analysis. Furthermore, these techniques were used to assess the viability of the cells collected during the separation and ensure they were sufficiently viable for further analysis.

To assess the viability of the cell populations, a dead cell stain was used to discriminate between live and dead cells within the samples. The percentage of live cells in each sample was then calculated for each of the CD4+ and CD8+ populations from each healthy control sample. Viability could not be calculated for 2 of the CD8+ samples and 6 of the CD4+ samples due to sample and reagent limitations so the performance of other samples were used as a indication of how efficient the separation technique used was. Figure 50 shows histogram peaks of the viability of CD8+ and CD4+ cell populations across selected healthy control samples. Figure 50 A is a histogram peak showing the percentage of live and dead cells in a CD8+ sample with the x axis representing the uptake of the cell viability stain and the y axis representing the cell count. In the left hand peak, live cells that have an intact cell membrane are counted whilst in the right hand peak dead cells, which have taken up the viability dye, are counted (R1). Figure 50 B shows histogram peak representing the uptake of the dead cell stain in a CD4+ sample. As before the left hand peak represents cells that did not take up the dead cell stain whilst in the right hand peak the cells that took up the dead cell stain are counted (R4).

Figure 51 and Figure 52 show the overall viability of CD8+ and CD4+ cell populations across the maximum number of samples. In the CD8+ cells (Figure 51) the average viability was 80% whilst in the CD4+ cells the average viability in the CD4+ cells was 67% (Figure 52). This is not unusual as the cells have been previously cryopreserved and have undergone a series of centrifugation steps in section 3.4.2.6. The CD4+ lymphocytes had a lower average viability than the CD8+ lymphocytes, which may be a result of the CD4+ separation being performed last. At this point the cells would have passed through the column several times and would have undergone several centrifugation steps, which would have increased the risk of damage to the cell membranes.

Figure 50– Plots showing viability of CD8+ and CD4+ lymphocytes

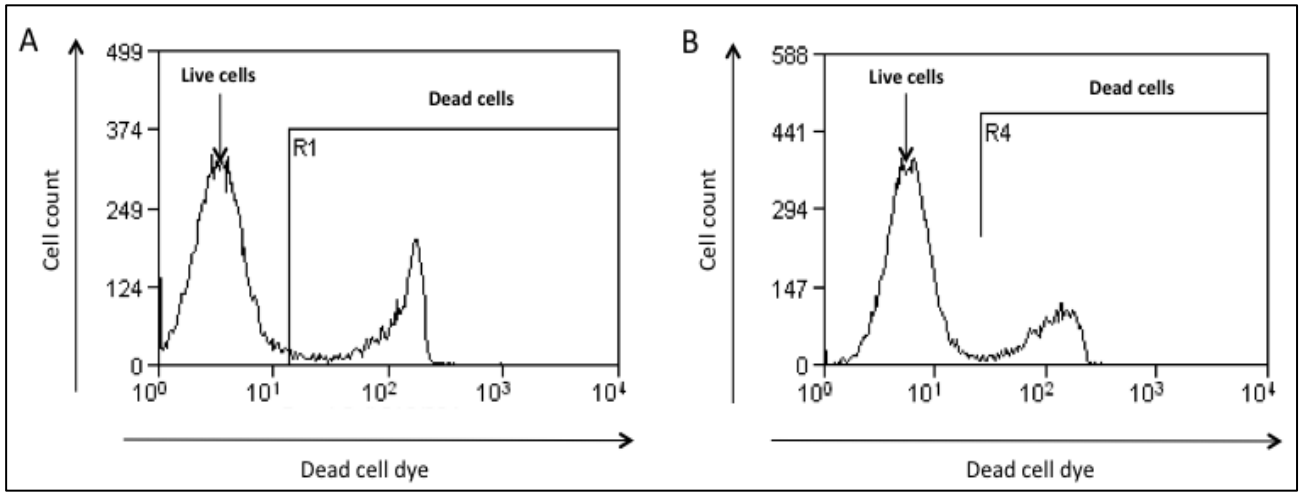


Figure 50 shows the viability of cells in 2 lymphocyte samples treated with dead cell viability stain. The x axis represents fluorescence due to uptake of the dead cell stain whilst the y axis represents the cell count A) Histogram of a selected CD8+ lymphocyte sample, with the right hand peak representing dead cells which took up the dead cell stain (R1) whilst the left hand peak represents viable cells which did not take up the stain. B) Histogram of a selected CD4+ lymphocyte sample, with the right hand peak representing dead cells which took up the dead cell stain (R4) whilst the left hand peak represents viable cells which did not take up the stain.

Figure 51- CD8+ lymphocyte viability across all samples

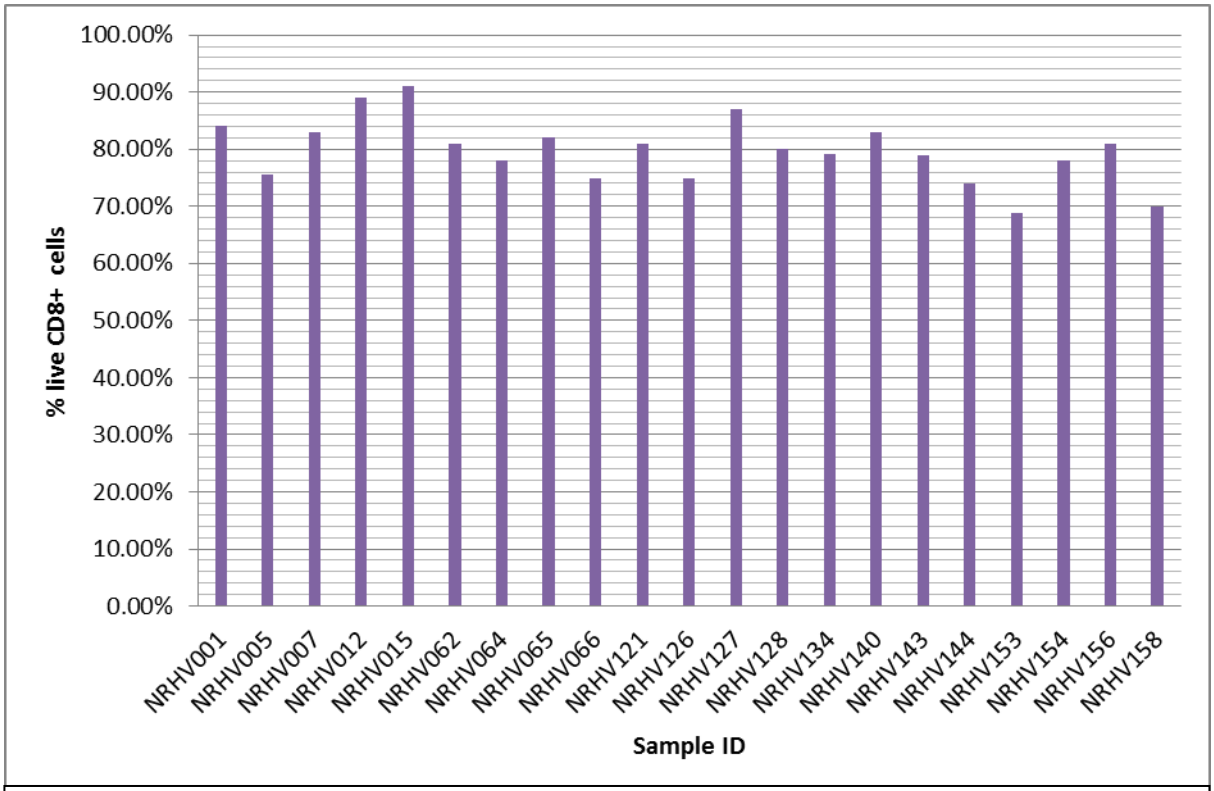


Figure 51 shows a bar chart of CD8+ lymphocyte viability from 21 of the 23 samples with the x-axis representing the sample name whilst the y-axis represents the percentage of cells that were viable in each sample. Viability was expressed as the percentage of cells, which did not take up the dead cell stain, compared to the total number of cells analysed.

Figure 52- CD4+ lymphocyte viability across all samples

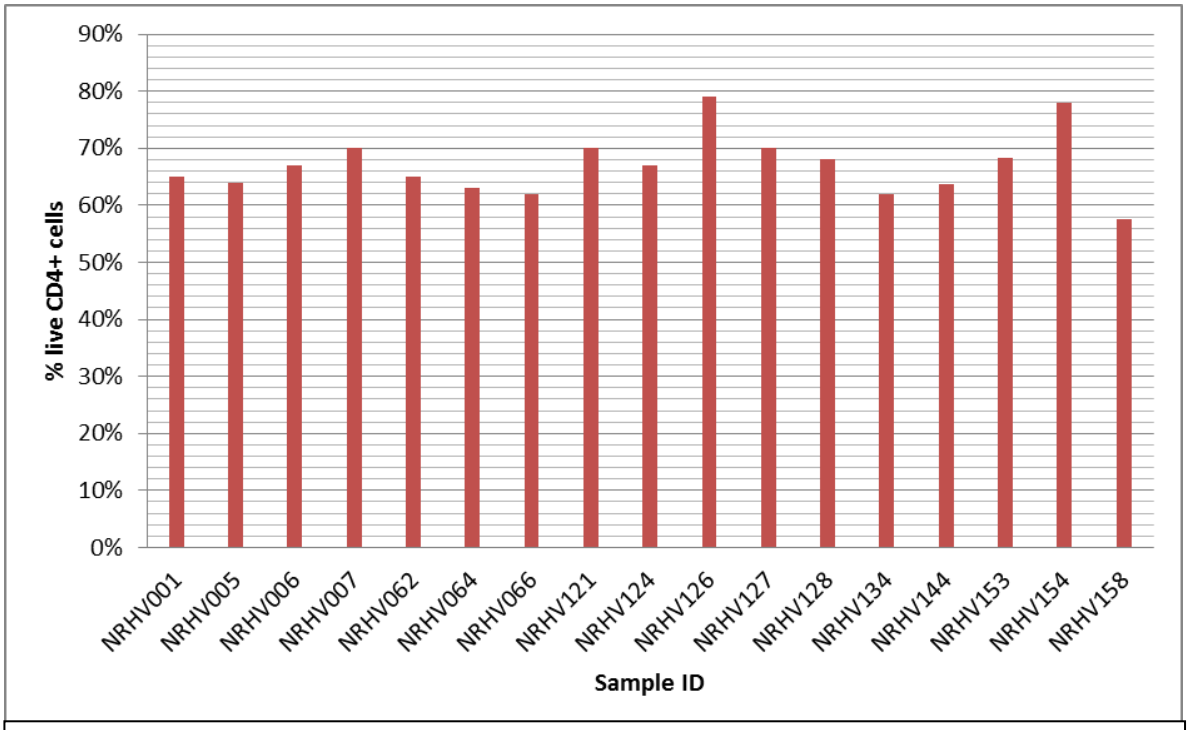


Figure 52 shows a bar chart of CD4+ lymphocyte viability from 17 of the 23 samples with the x-axis representing the sample name whilst the y-axis represents the percentage of cells that were viable in each sample. Viability was expressed as the percentage of cells, which did not take up the dead cell stain, compared to the total number of cells analysed.

3.4.2.8 Purity of cell populations

To assess the purity of the total cell populations obtained during the cell separations, the expression of CD8+ or CD3+CD4+ was examined. This was then calculated as a percentage of the total number of cells counted to give an overall sample purity. Figure 53 and show some examples of the flow cytometry scatter plots and histograms obtained from the analysis of the cells across selected healthy control samples. Figure 53 A + B shows the flow cytometry output for 2 of the CD8+ lymphocyte samples. This is shown as a single stain histogram with the x-axis representing increasing levels of CD8-APC expression and the y-axis representing the cell count. Prior to this analysis cells were gated as described in section x. In the right hand peak a threshold has been set (R4) to capture the number of cells, which are positive for CD8-APC expression. The purity of the cell population was then calculated using the number of cells expressing CD8 compared to the number of cells analysed in total.

Figure 53– Histogram plots showing CD8+ population purity

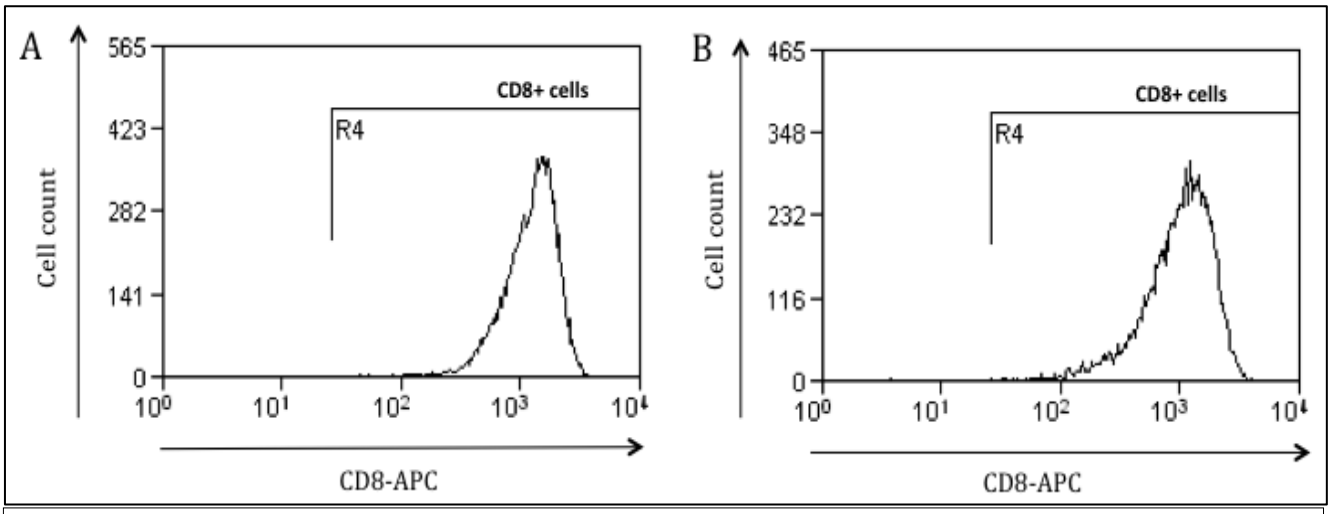


Figure 53 shows the expression of CD8-APC across 2 samples which have been gated as previously described in section 2.4.2.7.3. A-B) Histograms showing the expression of CD8-APC in the x axis and the cell count in the y axis. The right hand peak (R4) represents the cells, which are positive for CD8-APC whilst the presence of a peak in the left hand side would represent cells, which did not express CD8-APC.

Figure 54 represents an example of the flow cytometry output from 2 of the CD4+ lymphocyte samples double stained with CD3-APC/CD4-PE (Figure 54 A+B) and 2 different samples double stained with CD3-APC/CD4-VioBlue (Figure 54 A+B). As these were double stained, this is shown as a dot plot with each dot representing a single cell. The x-axis represents increasing levels of CD3 expression whilst the y-axis represents increasing levels of CD4 expression, with the position of each dot indicating the co-expression levels of these. In both sets a quadrant threshold was set determining positivity for both CD3 and CD4, which are characteristic of CD3+CD4+ T lymphocytes (PE = R6, Vioblue = R11). The number of cells in this quadrant was then used to calculate the purity of the CD4+, which is the number of cells expressing CD3 and CD4 compared to the total number of cells analysed.

Across all the CD4+ plots (Figure 54; A-D) a low number of CD3+CD4- can be detected in the CD3+CD4- quadrant (A+B=R5; C+D=R10). These may represent CD3+CD8+ cells not removed by the CD8+ positive selection or the presence of CD3+CD4- accessory cells which are involved in lymphoid development and B cell responses (Lane et al. 2005). As this population consists of very few cells, it is unlikely they have an aberrant effect on the results.

Figure 54– Plots showing CD4+ population purity using PE and Vioblue flourochromes

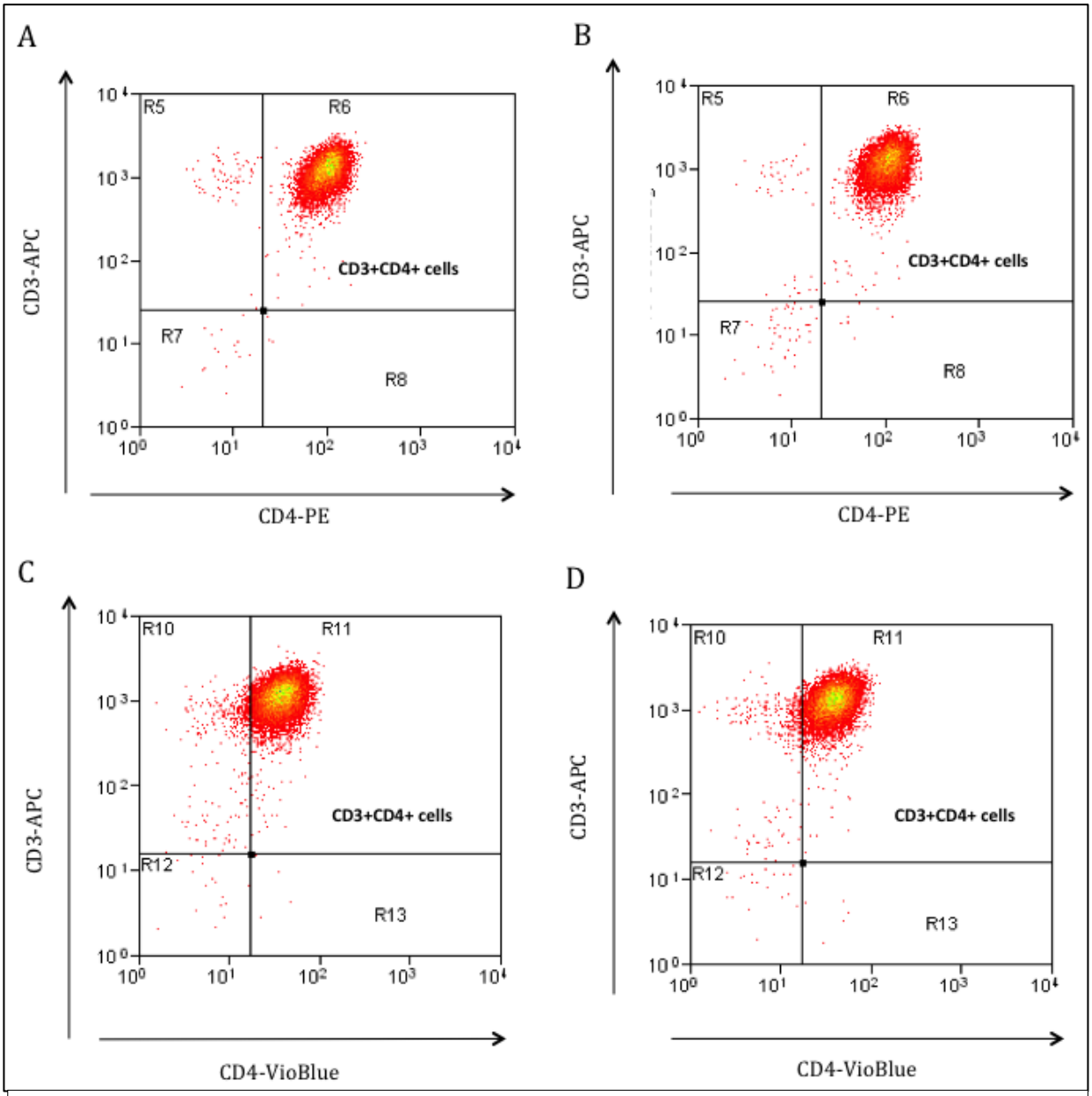


Figure 54 shows the expression of CD3 and CD4 on cells, which have been gated, as previously described in section 2.4.2.7.3. A+B) Dot plots of 2 CD4+ samples from the first 6 samples stained with CD3-APC and CD4-PE. The x-axis represents expression of CD4-PE whilst the y-axis represents expression of CD3-APC. The position of each dot is a representation of the co-expression levels of CD3 and CD4, with each dot representing a single cell. A quadrant threshold was generated for positivity for both CD3 and CD4 (R6). This was used to calculate the purity of each sample. C+D) Dot plots of 2 CD4+ samples from the first 6 samples stained with CD3-APC and CD4-VioBlue. The x-axis represents expression of CD4-VioBlue whilst the y-axis represents expression of CD3-APC. The position of each dot is a representation of the co-expression levels of CD3 and CD4, with each dot representing a single cell. A quadrant threshold was generated for positivity for both CD3 and CD4 (R11). This was used to calculate the purity of each sample.

In Figure 55 and Figure 56 the purity of CD8+ and CD4+ populations across all samples is shown using bar charts. In these plots the x-axis represents the sample ID whilst the y-axis represents the purity of the samples. This was calculated using the number of cells expressing CD8 or CD3-CD4 and calculating as a percentage of the total cells analysed. The average purity in the CD8+ populations (Figure 55) was 99.4% whilst the average purity in the CD4+ cells (Figure 56) was 95.31%. In both cases these represent very pure cell populations with minimum contamination by other cell types. The decrease in purity of CD4+ cell populations compared to CD8+ may be a result of the CD4+ cells being separated in the final separation as described previously. As a result all CD8+ and CD4+ samples were considered of sufficient purity for total RNA extraction and gene expression analysis.

Figure 55 – CD8+ lymphocyte purity

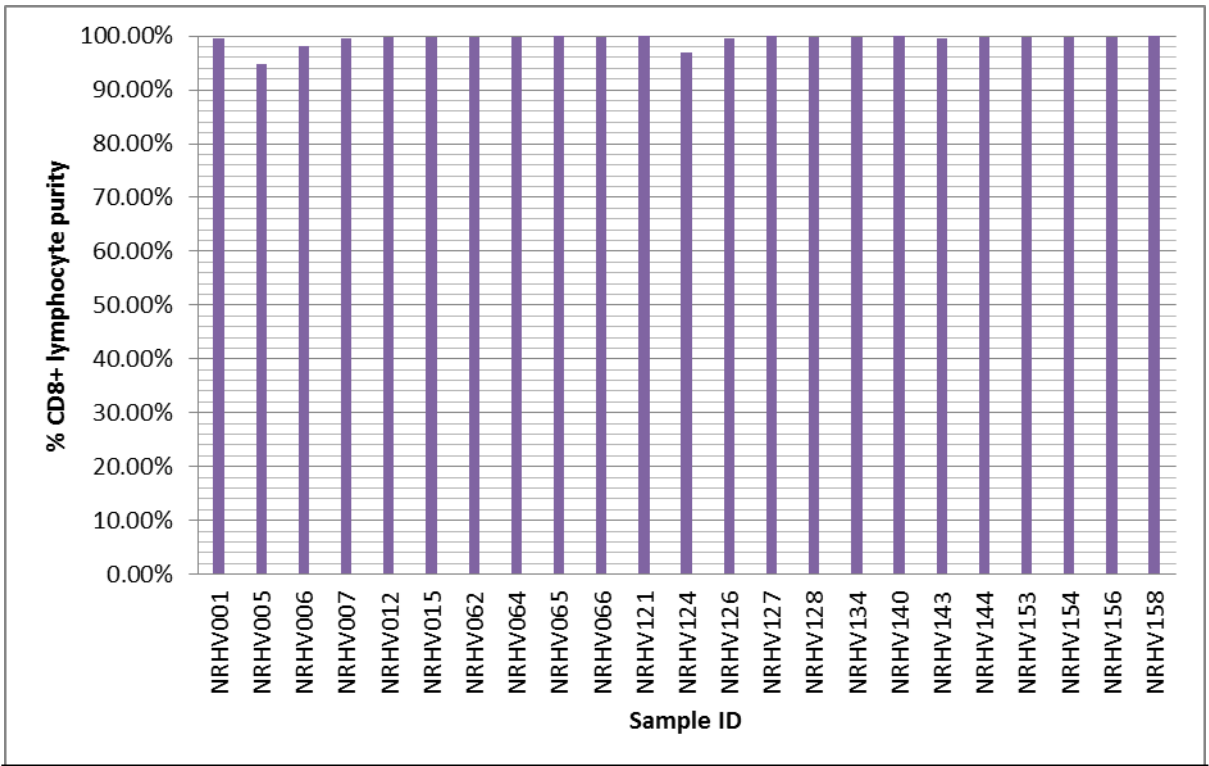


Figure 55 shows a bar chart of CD8+ lymphocyte purity across all samples, with the x-axis representing the sample name whilst the y-axis represents the samples purity. Purity was calculated using the number of cells expressing CD8 and calculating as a percentage of the total cells analysed.

Figure 56– CD4+ lymphocyte purity

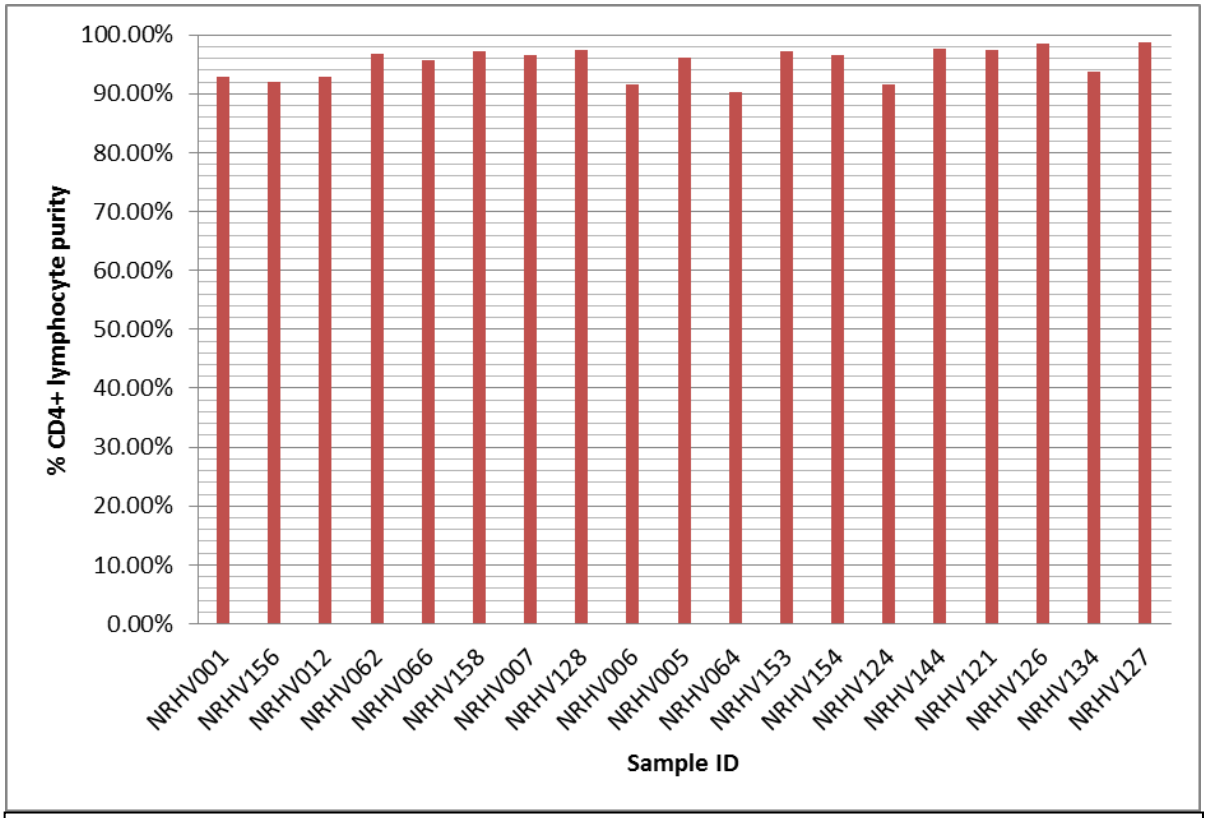


Figure 56 shows a bar chart of CD4+ lymphocyte purity across all samples, with the x-axis representing the sample name whilst the y-axis represents the samples purity. Purity was calculated using the number of cells expressing CD3-CD4 and calculating as a percentage of the total cells analysed.

3.4.2.9. Extracting total RNA from cell subset suspensions

46 total RNA samples (23 CD8+ and 23 CD4+) were successfully extracted from the cell suspensions.

3.4.2.10 RNA quality control

Total RNA was extracted from the 46 cell suspensions (23 CD8+ and 23 CD4+ samples). All samples were run on the Nanodrop N1000 and the Bioanalyzer to

calculate total yield and RIN values. Table 43 shows the total RNA yield for each sample, the 260/280 ratio, the 260/230 ratio and RIN value for each sample. All samples had a total yield of greater than 400ng required for gene expression analysis. 44 of the samples had a RIN of greater than 5, therefore were suitable for cRNA conversion in section 3.4.2.11. Samples, which did not meet these parameters, are highlighted in red. NRHV128_CD8 sample had a RIN of 4.1 so was considered low quality but was still converted to cRNA whilst NRHV012_CD4 had a RIN, which could not be calculated and was removed from further analysis. It was noted that the 260/230 ratios in the samples were much lower than expected for RNA (normally ~2). It was determined that this was due to the low quantity of RNA being analysed combined with the presence of minute amounts of Trizol carryover from the RNA extraction. Although noted it was not expected that this would affect the performance of the gene expression analysis as the carryover was minimal. Figure 57 shows an example Electropherogram from 2 of the healthy control samples. During the electropherogram, the presence of 18s and 28s ribosomal RNAs are measured, with the x axis representing time and the y axis representing the quantity of these subunits. The presence of strong 18s and 28s peaks on the Electropherogram without the presence of additional artefacts is indicative of a good quality RNA sample. In both these cases the RIN of each sample was greater than 8.

Table 43- Characteristics of extracted RNA

Sample ID	CD4+ RNA yield	260/230	260/280	CD4+ RIN	CD8+ RNA yield	260/230	260/280	CD8+ RIN
NRHV001	1532ng	1.39	1.76	6	1239ng	1.36	1.74	6.2
NRHV005	1840ng	1.48	1.78	8.5	2438ng	1.55	1.81	8.5
NRHV006	1838ng	1.72	1.82	5	1669ng	1.71	1.78	5.2
NRHV007	1345ng	1.59	1.92	8.2	1940ng	1.69	1.79	8.8
NRHV012	993ng	1.09	1.71	N/A	1395g	1.48	1.73	8.2
NRHV015	2382ng	1.71	1.821	5.1	1865ng	1.6	1.74	7.9
NRHV062	1027ng	1.22	1.74	6	1492ng	1.37	1.73	6.3
NRHV064	877ng	1.19	1.72	8.5	746ng	1.1	1.8	8.5
NRHV065	4200ng	2.01	1.91	7.6	2440ng	1.71	1.78	9.3
NRHV066	537ng	1.24	1.69	8.2	1392ng	1.87	1.64	8.9
NRHV121	2575ng	1.75	1.78	8.8	3292ng	1.87	1.82	8.9
NRHV124	2243ng	1.74	1.78	7.7	2775ng	1.74	1.78	7.2
NRHV126	2607ng	1.7	1.8	8.4	1965ng	1.5	1.85	8.2
NRHV127	3970ng	1.86	1.88	8.7	2580ng	1.75	1.78	8.1
NRHV128	720ng	1.11	1.68	8.4	532ng	1.09	1.72	4.1
NRHV134	1183ng	1.51	1.78	7.1	1300ng	1.44	1.75	8.5
NRHV140	573ng	1.31	1.76	8	2325ng	1.83	1.87	8.7
NRHV143	1238ng	1.53	1.74	8	1728ng	1.6	1.74	8.4
NRHV144	2560ng	1.58	1.86	8	3118ng	1.64	1.75	8.4
NRHV153	2135ng	1.7	1.83	8.3	1698ng	1.72	1.78	9

NRHV154	2144ng	1.55	1.9	7.7	3118ng	1.79	1.73	8.5
NRHV156	988ng	1.28	1.77	7.8	1390ng	1.51	1.72	8.3
NRHV158	1236ng	1.13	1.73	8.1	2170ng	1.66	1.79	8.5

Table 43 shows the RNA yield from each of the CD8 and CD4 samples. Details of RNA yield, 260/230 ratio, 260/280 ratio, and RIN are given for each cell type. RIN = RNA integrity number,

Figure 57– Bioanalyzer traces of 2 healthy control samples

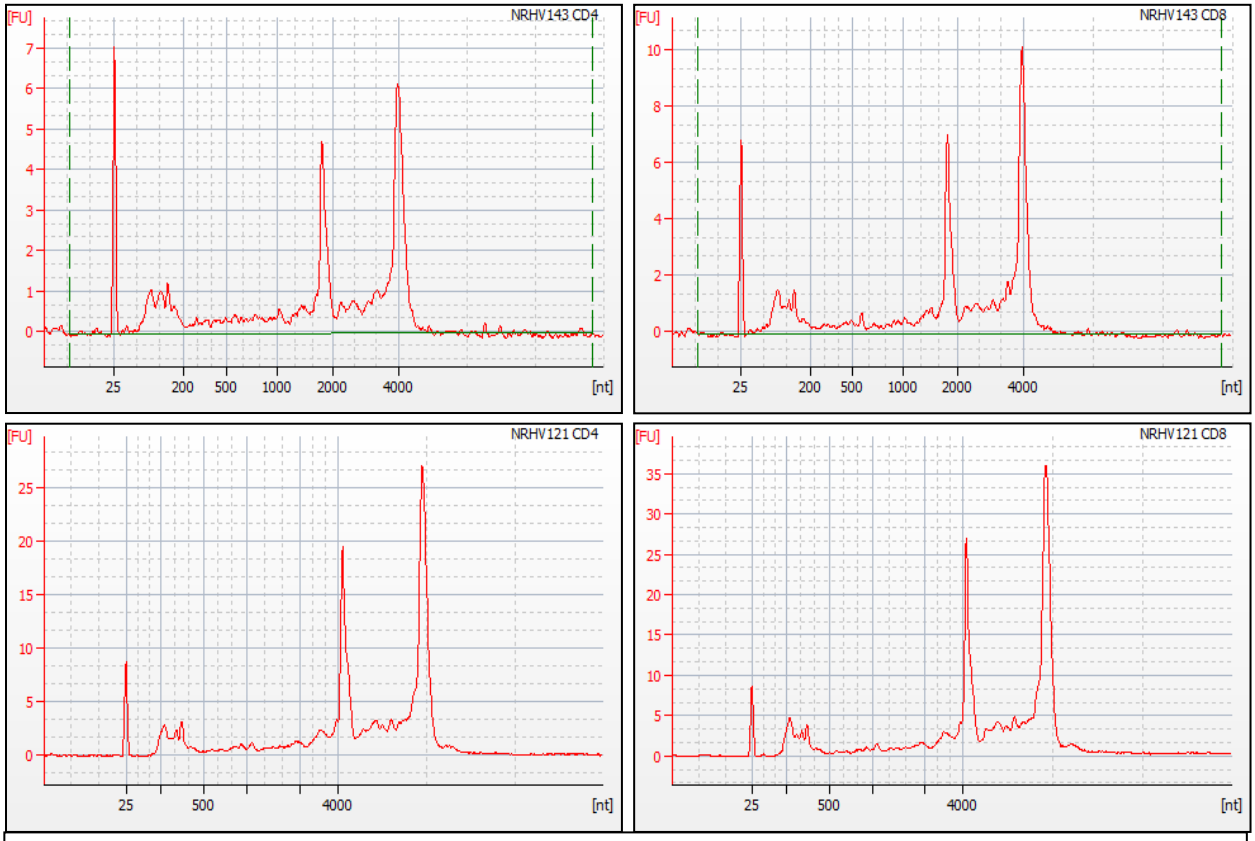


Figure 57 shows the electropherogram peaks from 2 CD4+ (left) and 2 CD8+ (right) RNA samples. On the x-axis is the time in seconds from the initiation of the run whilst the y-axis represents the fluorescence units. The presence of defined 18S and 28S peaks combined with low background noise is used to determine integrity and concentration of RNA samples.

3.4.2.11 DNase treatment of Total RNA

45 samples (23 CD8+ and 22 CD4+) were treated with a DNase treatment to remove any genomic DNA contamination.

3.4.2.12 RNA amplification using Illumina TotalPrep Amplification Kit

400ng of total RNA from 45 samples (23 CD8+ and 22 CD4+) was successfully converted to cRNA for gene expression analysis.

3.4.2.13 Illumina Gene Expression Direct Hybridization Assay

750ng of cRNA from 45 samples (23 CD8+ and 22 CD4+) were successfully hybridised to the Illumina HumanHT expression array for gene expression analysis.

3.4.2.14 Detecting differential signals on array

Array metric data for 45 (23 CD8+ and 22 CD4+) samples across 47,000 probes was successfully generated. 1 CD4+ sample did not scan successfully and therefore had to be removed from further analysis.

3.4.2.15 Gene expression normalization and QC

Once the arrays were scanned, files were generated detailing a number of sample and array metrics. These files were used to normalise and QC the gene expression data in a series of stages. Figure 58 details each normalisation and QC stage employed and the number of samples/array probes that were removed at each stage.

Figure 58– Gene expression normalisation and QC

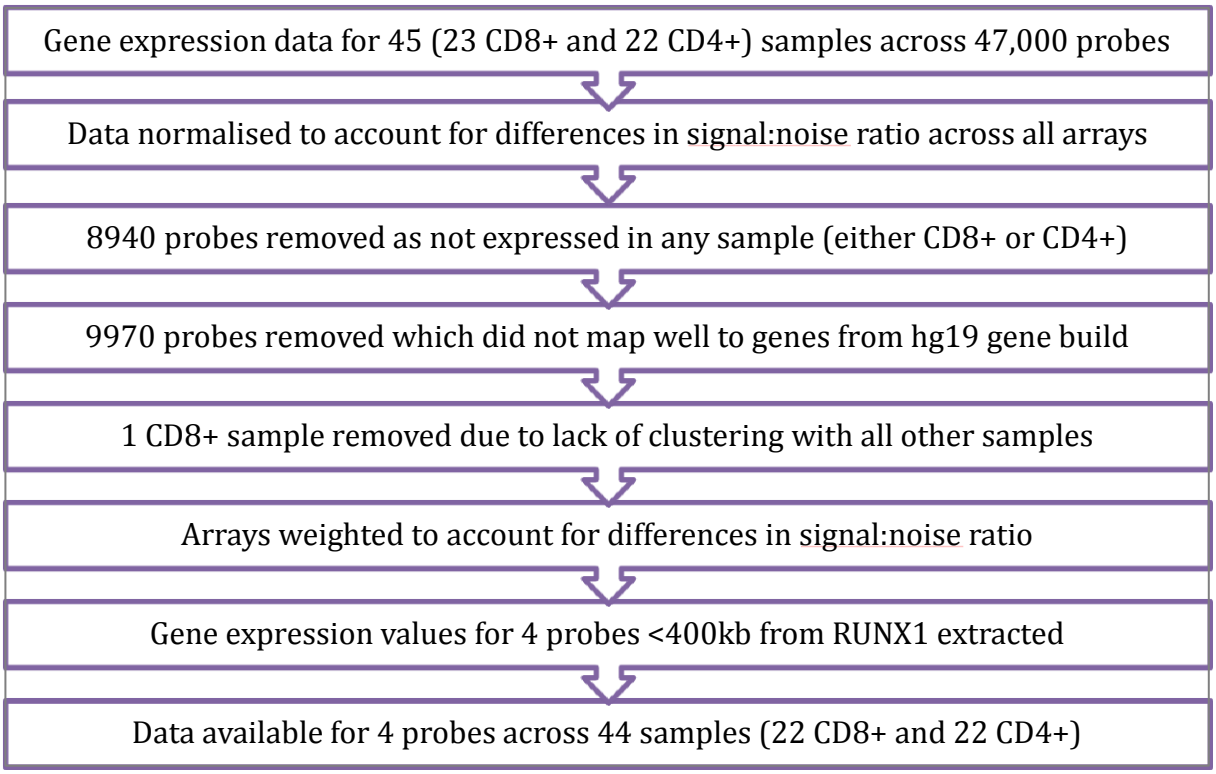


Figure 58 shows the normalisation and QC workflow for the gene expression array data. Data was first normalised to account for inter-sample differences in signal to noise ratios. Samples and probes were removed based on the following sequential criteria: probes which were not expressed in any sample; probes which did not map well to the hg19 gene build; samples which did not cluster neatly on the MDS plot. Each array was then weighted based on it's performance and the probes in the *RUNX1* region extracted for analysis. hg = human genome, MDS = multidimensional scaling.

3.4.2.16 Calculation of the signal to noise ratio across arrays

Signal to noise ratios were calculated for all 45 samples across all arrays. Figure 59 shows the signal to noise ratio for each sample, with all samples achieving a ratio of more than 2, although some are notably lower than expected.

Figure 59 – Signal to noise ratios for 45 samples

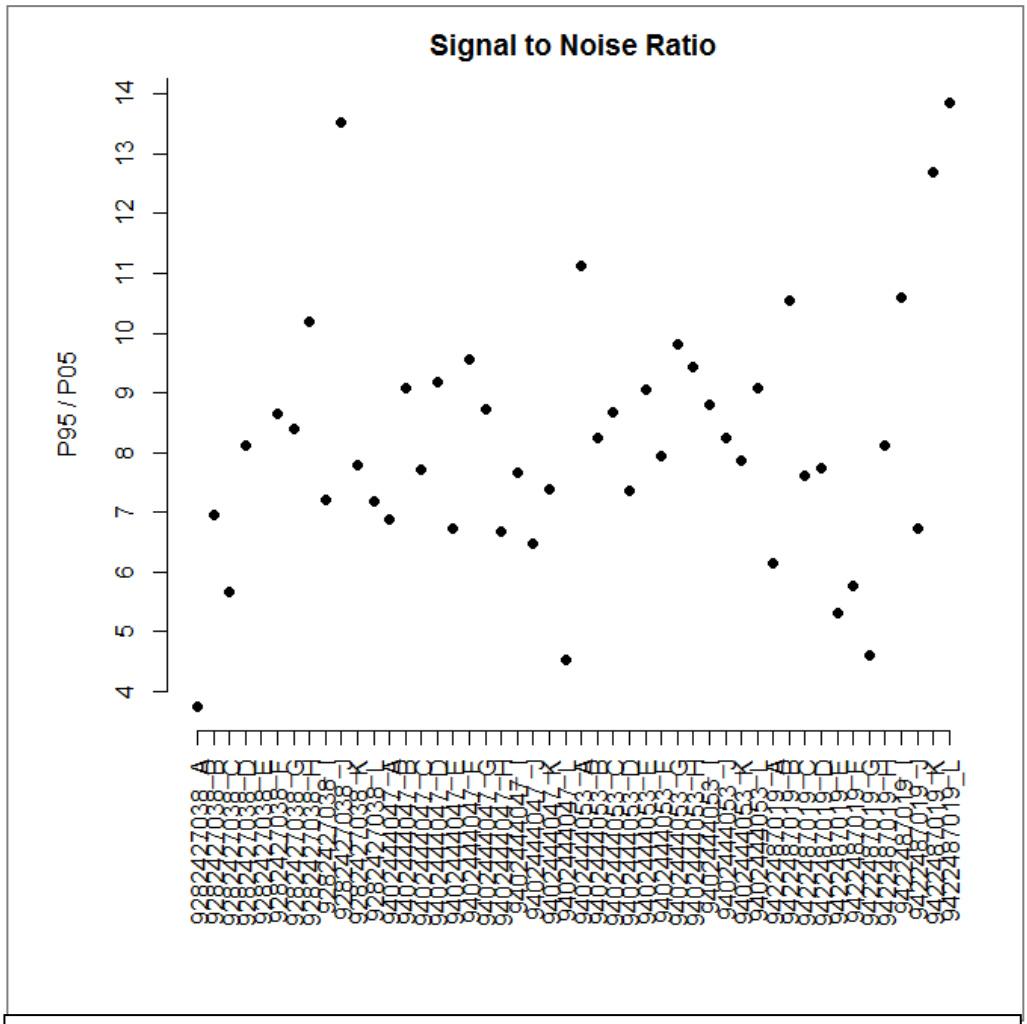


Figure 59 shows the signal to noise ratio calculated for all 45 samples. In this plot each sample/array is represented by a dot with the x-axis representing the array ID whilst the y-axis represents the ratio is signal to noise for each. The higher the dot on the y-axis the higher the signal:noise ratio.

3.4.2.17 Calculation of the intensity signals across probes

Average signal intensities were calculated for regular and negative control probes across all arrays. Figure 58 shows the $\log^2$ intensity signal of regular and negative control probes across all arrays. Background correction was then performed on the regular probes using the values obtained from the negative control probes before NEQC quantile normalization and $\log^2$ transformation. Figure 58 shows the neqc normalised $\log^2$ intensities across all arrays for regular probes.

Figure 60– Average signal intensity in raw and normalized data

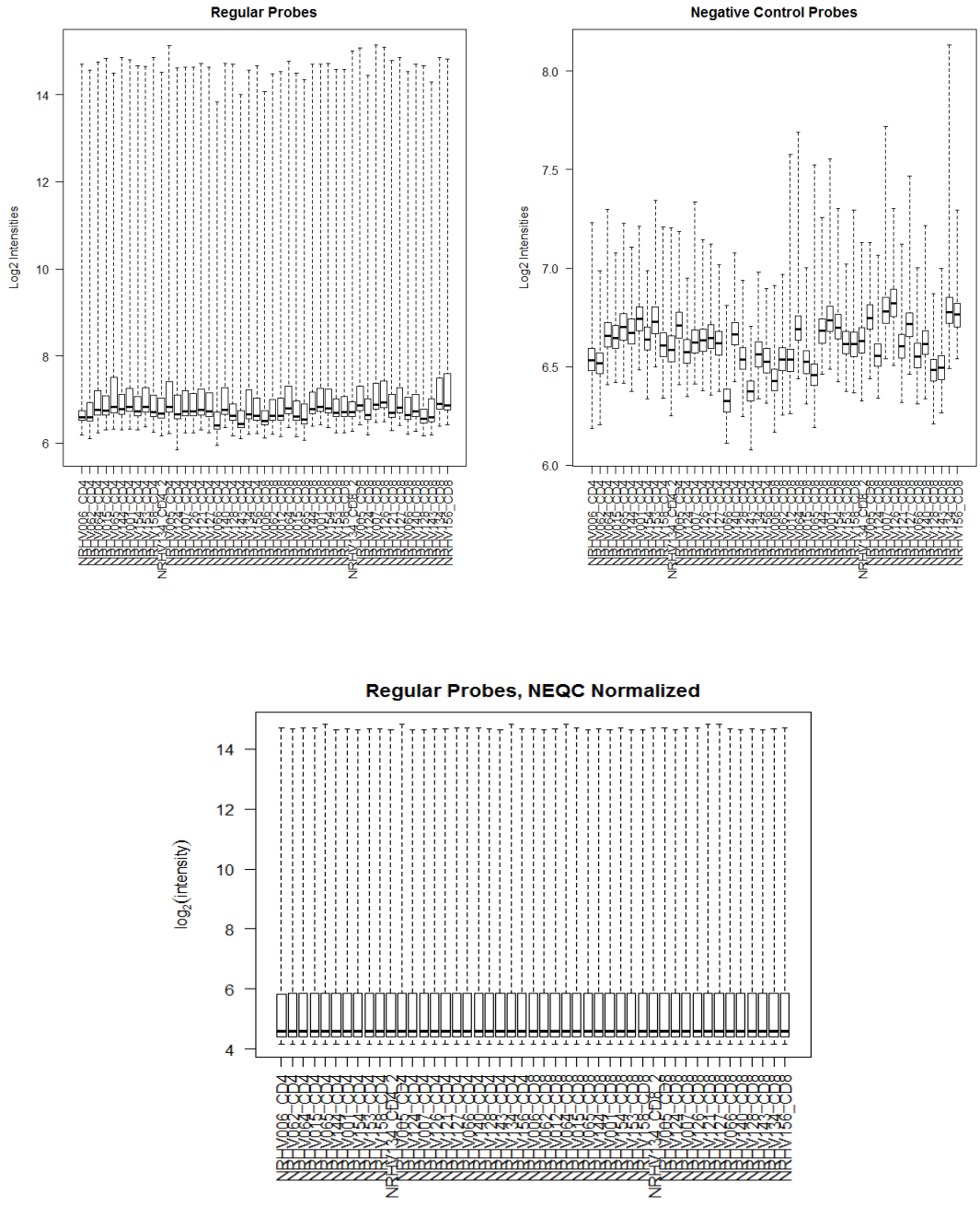


Figure 60 shows the average signal intensity in raw and normalized data. In each plot the x-axis represents the sample ID whilst in they-axis the $\log_2$ signal is shown. In the top plots the un-normalized data for the regular (L) and negative control (R) probes is shown. In the bottom plot the normalized data is shown.

3.4.2.18 Calculation of the proportion of probes expressed by each sample

The average proportion of probes expressed by the CD8+ and CD4+ samples was calculated successfully. In CD8+ cells the proportion of genes expressed was 0.46 whilst in CD4+ cells it was 0.47, indicating that slightly more of the probes are expressed in CD4+ cells. When compared, it was found that these proportions were not significantly different between cell types (p=0.3). Additionally 8940 probes were not expressed in any sample and therefore were removed from further analysis.

3.4.2.19 Matching probes to hg19 transcripts

In total 28404 probes mapped either good or perfectly to transcripts from hg19. All others were removed from further analysis.

3.4.2.20 Identification of sample outliers

To assess how similar the samples within sample groups were, MDS plots were generated. Figure 59 shows the MDS plots showing clustering by sample group, showing that with the exception of a single CD8+ sample, the samples cluster tightly with others of the same sample group. The CD8+ sample was removed from further analysis, as it may result in skewing of the results.

Figure 61– MDS plot showing clustering by sample type

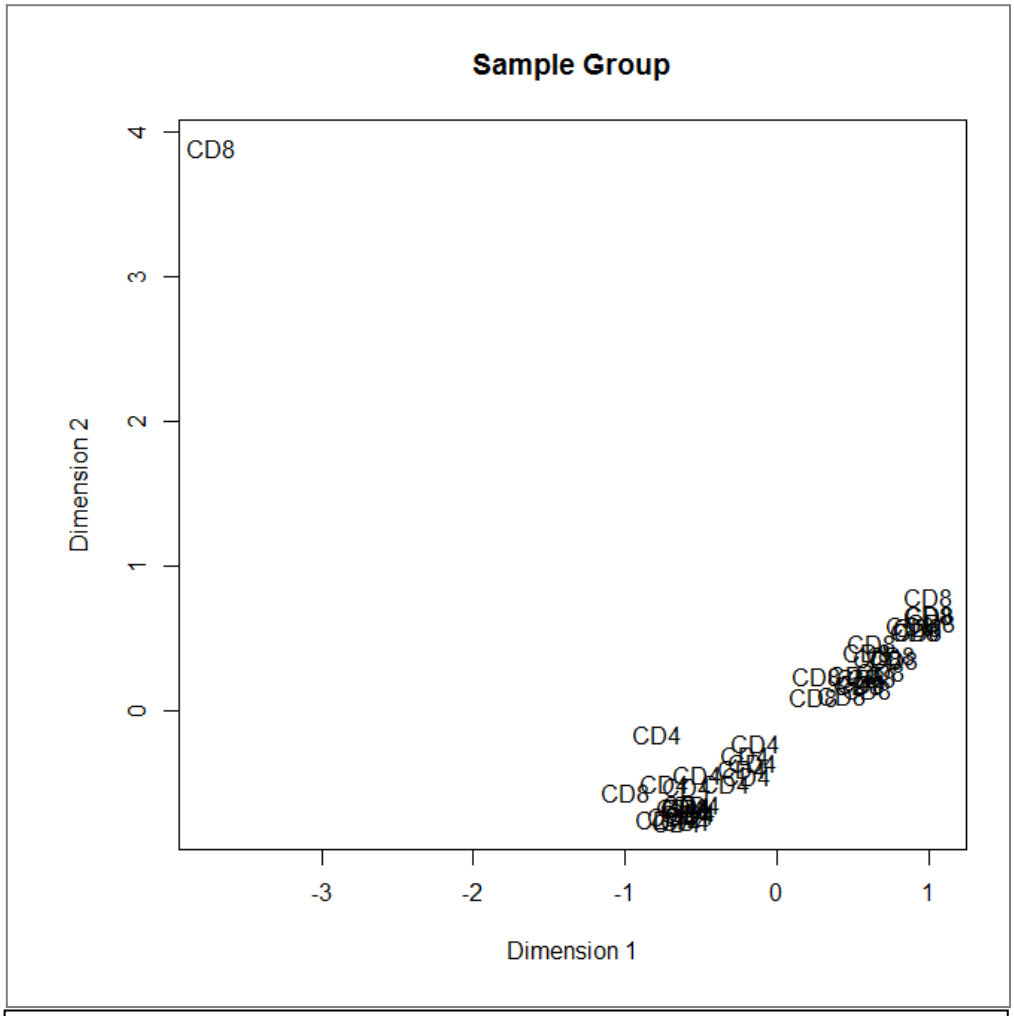


Figure 61 shows an MDS plot which indicates sample outliers in data based on similarities between probe expression. In this plot 1 outlier was detected and the remaining samples clustered well into CD4+ and CD8+ groups indicating they represent different cell types.

3.4.2.21 Principal components analysis

To identify factors which may be contributing to sample variance and therefore could contribute to batch effects, PCA was performed**.** Figure 62 shows the contribution of each PC to variance within the dataset. In total there were 2 PCs contributing to more than 10% of sample variance, which were both identified as being sample grouping. This indicates that the difference in CD8+ and CD4+ samples represents more than 10% of variance in the dataset and therefore these samples should be analysed separately.

Figure 62– Contribution of principal components to sample variance

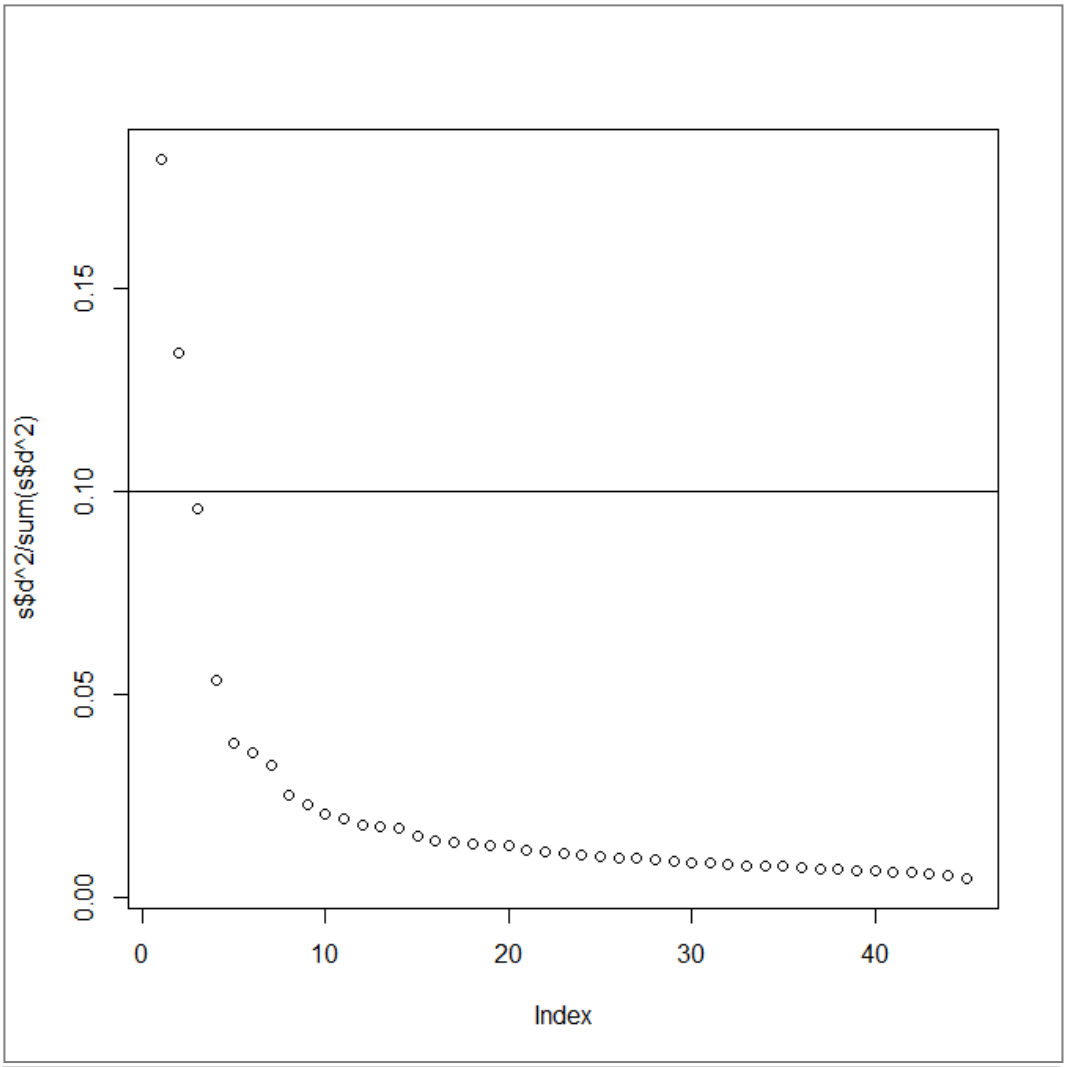


Figure 62 shows the contribution of the top 2 principal components generated in the PCA analysis. In this plot is can be seen that 2 principal components are contributing significantly to data structure. These were identified as both being cell type.

3.4.2.22 Array weighting

Array weighting was performed successfully for the remaining 44 samples.

3.4.2.23 Cell specific eQTL analysis

In the 22 CD8+ and 22 CD4+ samples available for analysis, gene expression values for the 4 probes in the *RUNX1* region were successfully calculated. Of these probes 2 mapped to the *RUNX1* gene whilst 2 were mapped to the non-coding RNA *LOC100506403* which is located 300kb upstream of the *RUNX1* region. Figure 61 shows a QC summary for the eQTL analysis when combining gene expression and genotype data from section 3.4.2.1.

When correlated with genotype at rs9979383, no significant correlations were found with expression with any of the 4 probes (all $p > 0.05$). Table 42 shows the p value obtained by linear regression, showing no significant results ($p < 0.05$) across the probes. For the result shown in bold, there does seem to be a trend towards significance ($p = 0.1$). This may be the result of limited power as only 1 sample in this cohort had 2 copies of the minor allele at rs9979383. This will require further investigation in a larger cohort. Figure 62 and Figure 63 show the expression for each of the individual probes stratified by genotype at rs9979393.

Figure 63- QC summary for eQTL analysis

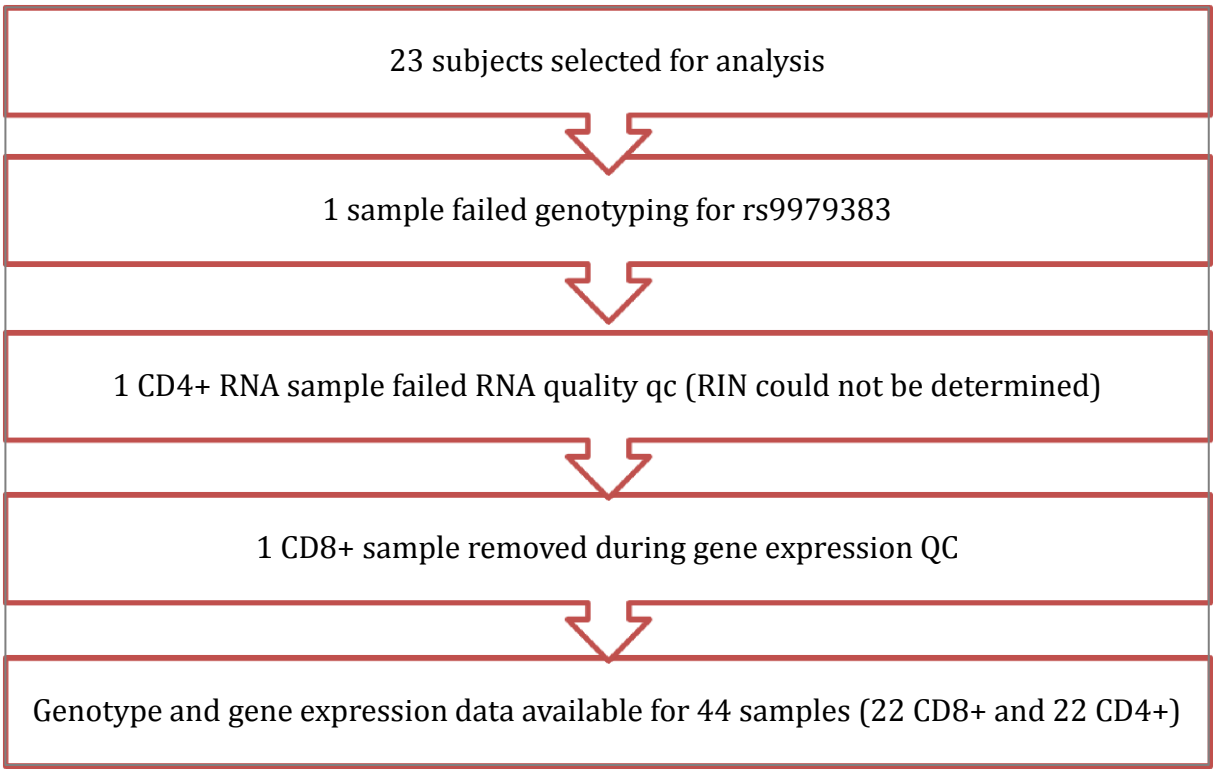


Figure 63 shows the QC stages employed for the cell specific eQTL analysis. Samples were removed sequentially based on the following parameters: failure to genotype; low RIN; failure to pass gene expression QC. Only samples which had both genotyping and gene expression data could be analysed. RIN = RNA intergrity number.

Table 44– eQTL results for *RUNX1* region

Cell type	Probe target	P
CD8	RUNX1	0.933
CD8	RUNX1	0.745
CD8	LOC100506403	0.747
CD8	LOC100506403	0.838
CD4	RUNX1	0.58
CD4	**RUNX1**	**0.1**
CD4	LOC100506403	0.492
CD4	LOC100506403	0.552

Table 44 shows eQTL analysis results in the RUNX1 region in both CD8 and CD4 lymphocytes. Across the columns cell type, the probe target and p value are shown. Although no significant eQTLs were identified in this data, there was a trend towards significance between rs9979383 and RUNX1 in CD4 lymphocytes. This is highlighted in bold.

Figure 64– *RUNX1* region eQTL analysis in CD8+ lymphocytes

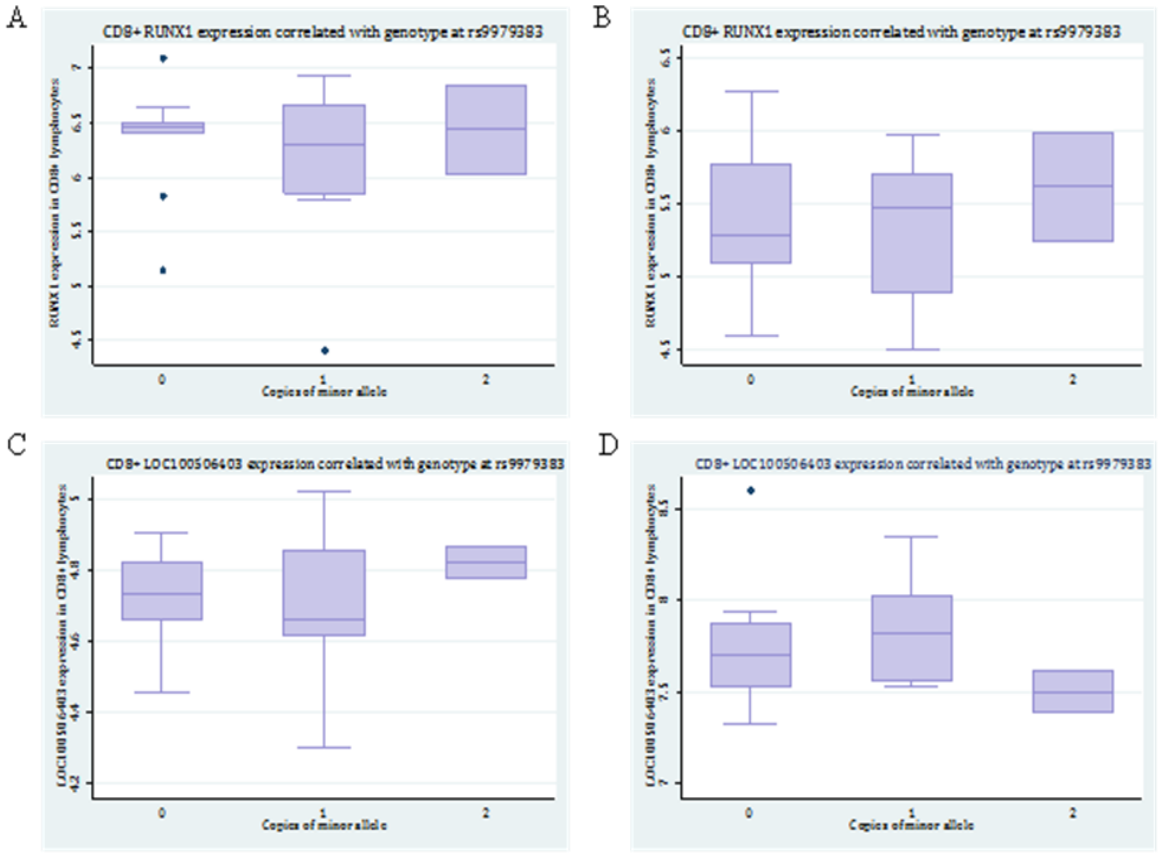


Figure 64 shows the eQTL results for the RUNX1 region in CD8 lymphocytes. In each plot the x-axis represents number of copies of the minor allele at rs9979383 whilst in the y-axis represents gene expression of the corresponding gene in the RUNX1 region. A+B) *RUNX1* gene expression in CD8+ lymphocytes correlated with genotype at rs9979383. C-D) *LOC100506403* gene expression in CD8+ lymphocytes correlated with genotype at rs9979383.

Figure 65– *RUNX1* region eQTL analysis in CD4+ lymphocytes

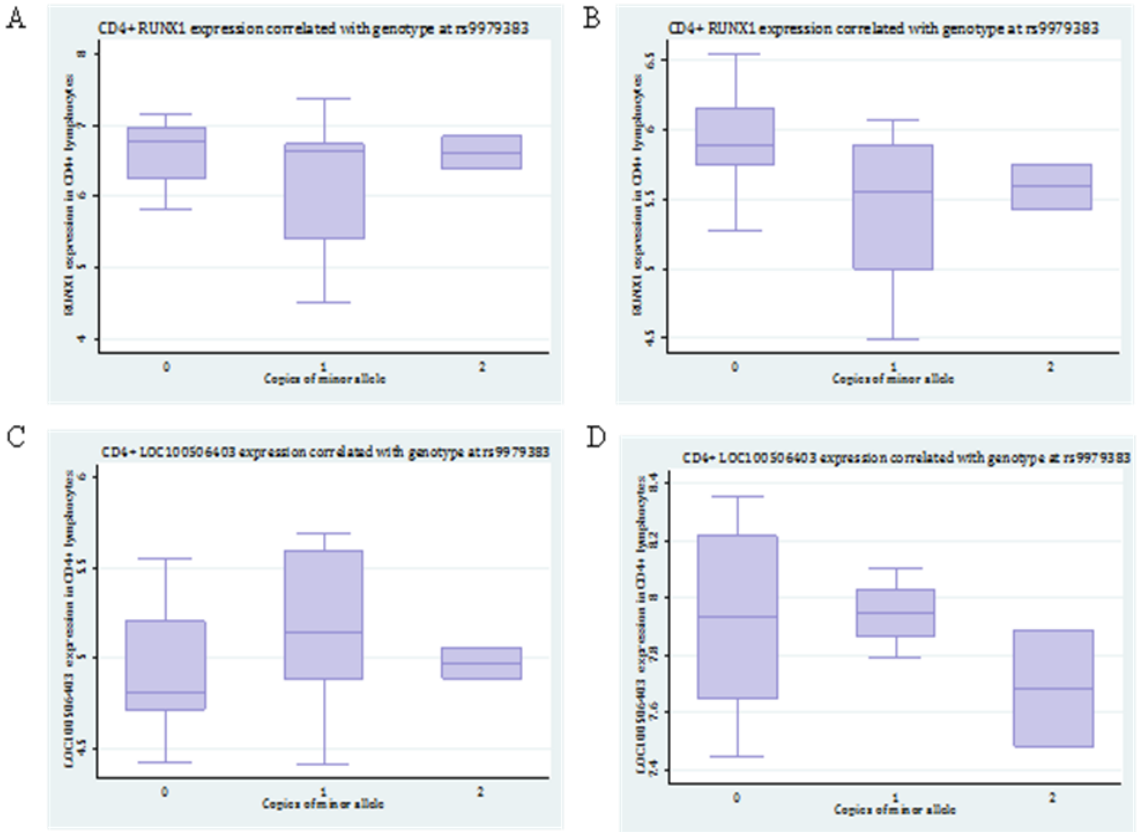


Figure 65 shows the eQTL results for the RUNX1 region in CD4 lymphocytes. In each plot the x-axis represents number of copies of the minor allele at rs9979383 whilst in the y-axis represents gene expression of the corresponding gene in the RUNX1 region. A+B) *RUNX1* gene expression in CD4+ lymphocytes correlated with genotype at rs9979383. C-D) *LOC100506403* gene expression in CD4+ lymphocytes correlated with genotype at rs9979383.

4.0 Discussion

4.0 Discussion

4.1 Summary of findings

The aim of this study was to assess the genetic overlap between the 3 types of IA (RA, JIA and PsA) using genotype data generated on the Immunochip array and to select a biologically promising overlapping region for further genetic and functional investigation using a variety of techniques.

50 genetic regions have now been identified as common to more than 1 type of IA ($p<1x10^{-3}$). Many of these overlapping regions represented novel disease associations and therefore required replication in an independent sample cohort. Of the 12 variants selected for the replication study, 2 variants have now been successfully replicated in a modestly sized independent RA cohort at $p<0.05$.

Bioinformatics analysis of the 50 overlapping regions indicated that many regions represent strong biological candidates. One particularly promising region, *RUNX*1, was selected for further investigation. In this region, the same variant (rs9979383) is associated across the 3 diseases, with similar odds ratios (0.8-0.9) observed in each disease. Fine mapping of this region in an independent RA cohort was performed, which resulted in replication of the association at $p=0.02$. No additional significant genetic effects were detected indicating the association signal is localized to this variant (or variants in high LD with it), at least in RA.

As rs9979383 lies ~280kb upstream of the *RUNX1* gene, a cis-eQTL analysis was performed to identify if the variant acts by regulation of *RUNX1* gene expression. This was performed in whole blood, CD4+ and CD8+ lymphocytes from healthy volunteers. Whole blood was selected due to sample availability and the T lymphocyte subsets were selected due to their importance in RA and PsA pathogenesis, respectively (Menon et al. 2014;Trynka et al. 2013). Although no significant cis-eQTL was detected in any of these tissues with either *RUNX1* or the nearby lnc-RNA *LOC100506403*, in cells from healthy volunteers under unstimulated conditions, further work is required to determine whether an eQTL exists when cell subsets are stimulated, as would be expected in inflammatory

conditions. These findings will direct future functional investigations into the role of this overlapping region in the susceptibility of IA.

4.2. Findings, strengths and weaknesses of the study

4.2.1 Immunochip overlap

My work is underpinned by the observation that immune mediated diseases often share common genetic factors. This has been shown for a number of immune mediated diseases such as RA, T1D, IBD and SLE, which have been shown to share multiple overlapping regions (Cotsapas and Hafler 2013;Eyre et al. 2010;Orozco et al. 2011). As RA, JIA and PsA are in some aspects clinically similar, for example they collectively involve articular disease and are often treated with similar anti-inflammatory therapies, it is also expected that some of their genetic component will also be shared.

Prior to my study, GWA and candidate gene genetic studies had identified over 50 genetic regions associated with a single type of IA but only 7 of these were identified as being associated with more than 1 type of IA, with only 1 region (*IL2/IL21*) being implicated across all 3 diseases (Hinks et al. 2010a;Stahl et al. 2010). At the start of my project, a large number of samples from patients with the 3 types of IA and healthy control samples were being genotyped on the Immunochip array for disease-specific genetic susceptibility studies.

The Immunochip array, which was used to genotype the samples included in my analysis, represents the ideal platform for identifying genetic overlap between diseases. This is because the content of the array was designed by a consortium of investigators studying 12 immune mediated diseases and the aim was to fine-map regions previously associated with immune mediated disease (Cortes and Brown 2011). The fact that the array could be tested in multiple immune diseases meant that the cost of the chip was much less than the cost of GWAS arrays at the time; in turn, this meant that investigators could test larger numbers of samples in multiple diseases, thereby increasing the power to detect modest effects. The array includes

many genes in essential immune pathways and hypothetically, these may also be associated with IA, so it enabled identification of association to novel regions. In many regions, there was dense coverage of genetic variants allowing fine mapping to be performed. This allows association signals to be refined and supports the identification of causal variants and putative multiple genetic effects within a region. The Immunochip provided a unique opportunity to genotype samples for 3 types of IA and perform overlap analysis in a large sample cohort. Although this was an effective strategy for many regions, it should be noted that not all regions on the Immunochip array were fine mapped. For example in the *RUNX1* region, only 7% of common variants from the 1000 genomes (July 2010 release; MAF>0.05) were captured, meaning that in order to identify causal variants and multiple genetic effects, further independent fine mapping studies were required.

Once association testing was performed for each disease separately, the results were compared to identify genetic regions associated with more than 1 type of IA. In total 50 genetic regions were associated with more than 1 disease at an association p value of less than $1x10^{-3}$. Of these, 14 regions showed association with more than 1 disease at p $<1x10^{-5}$. These p value thresholds were selected in order to detect all suggestive overlapping genetic effects, as the JIA and PsA cohorts had much lower power than the RA cohort due to the lower number of samples tested. Although several regions identified had been previously associated with 1 or more types of IA, several regions including *RUNX1, IL6R* and *RASGRP1* represented novel associations for 1 or more of the diseases. As novel associations, these regions were prioritized for replication in an independent cohort. The identification of novel overlapping regions in my study provides evidence that genetic overlap exists between these diseases, which is becoming apparent, as sample sizes get larger and power increases.

In total 10 regions were associated across the 3 diseases, which is a substantial increase from the pre-study findings, in which only 1 region was associated across the 3 diseases. Ideally, further analysis will be required to identify whether these overlapping regions fall into distinctive gene pathways but for now these findings provide novel insight into the genetic overlap between these 3 types of IA.

It was also observed that RA and JIA had the greatest genetic overlap overall with 31 of the 50 regions associated across the both diseases, which indicates they may share a considerable genetic component. Although an interesting finding and likely to be plausible, it is also important to consider the effect of differential power on the overlap results. With a large sample size available for RA, it is more likely smaller genetic effects were detected, which may not be detected in the lower powered JIA and PsA cohorts. This is often an issue with genetic association studies which lack power to detect genetic effects (de Bakker et al. 2005). It is therefore important that, as sample sizes increase, the question about the extent of shared genetic susceptibility between the 3 types of IA is revisited to identify whether this represents the true genetic overlap between the diseases.

The key strength of this study is that it represents the largest analysis of overlap between different types of IA to date, with genotyping data accessible for over 30,000 samples collectively. The availability of these samples was a major strength of this study and was only possible through the extensive international collaborations of the Arthritis Research UK Centre for Genetics and Genomics in Manchester. Despite this, the majority of the cases contributing to this number were derived from the RA cohort, which had 11,475 cases, compared to the JIA and PsA cohorts, which had 2816 and 929 PsA cases, respectively. This is reflected in the power calculations in section 2.1.5 which showed that the RA cohort had much greater power than the JIA and PsA cohorts and therefore was more likely to detect genetic effects. This factor has to be taken in account when examining the overlap between these diseases, as a well-powered study is much more likely to detect genetic associations. In addition, genuine genetic effects in the JIA and PsA cohorts may be missed as a consequence of lower power (type II error, false negative) in this analysis. An additional strength in this study was the large number of control samples, further increasing the power to detect genetic effects. This is advantageous but it is worth noting that in each individual disease association test, some of the controls were shared. It could be argued, therefore, that some associations may be driven due to deviance from expected allele frequencies in the controls, which may appear as an association (type 1 error, false positive). If this were the case then associations would be detected across all datasets and therefore considered overlapping in nature. Using shared or

overlapping controls is a standard procedure to increase power for GWA studies (Wellcome Trust Case Control Consortium 2007) and large-scale genetic studies, such as Immunochip. The collaborative nature of Immunochip meant that genotyping of shared controls was more cost-effective option. Stringent QC procedures are needed and in situations where there is uncertainty, replication in additional cohorts can give reassurance that the original observed associations were not due to type 1 error.

Once the 50 overlapping regions were identified, the LD between SNPs in the region was assessed to determine whether association at overlapping regions was conferred by identical or different genetic variants. In 31 regions, there was limited or no LD observed between the associated SNPs in comparisons between diseases ($r^2<0.4$). In many cases, the observed ORs were also opposing between diseases, suggesting that the overlapping region harbours differential genetic effects in each disease. One example of this is in the *IL2RA* and *IL2RB* regions where each of the index SNPs in the three types of IA show no or limited correlation by LD ($r^2<0.4$). In addition, the direction of effects is differential between RA and JIA with all RA index variants showing a risk effect for the minor allele (OR>1) whilst the JIA index SNPs appear to be protective (OR<1). This could consequently result in several potential outcomes including, first that the SNPs are regulating two different genes in the same region. Secondly, it is possible that the SNPs are both affecting the same gene but having an opposite effect on the gene, possibly through either upregulation or downregulation of gene expression. It is a particularly interesting finding as the regions contain genes of the IL2 cytokine pathway. The pathway is crucial for the activation and growth of immune cells, especially T lymphocytes (Cantrell and Smith 1984). Both CD4+ and CD8+ T lymphocytes have been shown to be important in IA pathogenesis (Berner et al. 2000) and, therefore, further investigation is required to identify how these differential effects between diseases contribute to disease susceptibility.

In 14 regions the SNP associated was either identical or highly correlated ($r^2>0.8$), whilst in 9 regions the SNPs associated with each disease are different but moderately correlated by LD ($r^2>0.4<0.8$). In many cases the direction of effects observed was similar between diseases and therefore might represent the same

genetic effect or the presence of multiple genetic effects in a region. One particularly interesting region is the *RUNX1* region in which the index SNP is identical across RA, JIA and PsA. Furthermore a similar OR of 0.83-0.9 was observed across the diseases, which makes this a particularly interesting association when exploring shared genetic susceptibility to IA. Given the fact that shared controls were used, however, replication of the association in an independent data set (including an independent set of controls) was essential to confirm that the association had not simply been observed because of an erroneous estimate of allele frequencies in the control population studied.

A key strength of my work, therefore, is that a separate replication in an independent cohort of RA cases and controls was undertaken, confirming association at the locus with a similar effect size to that observed in the original RA cohort. A limitation of the study is that independent JIA and PsA cases and further independent controls were not available to confirm the association in these diseases. However, the findings provided confidence that the original associations had not arisen due to type 1 error.

The analysis performed in this study was sufficient to obtain both an overview of the regions associated with more than 1 type of IA and select a promising region for follow up. Despite this, there are many more complex statistical techniques, which could be employed to detect further genetic overlap. For example, a pan-meta study analysis could be performed by combining all the IA (RA, JIA and PsA) cases and analysing against all healthy controls. This method would provide the greatest power and should hypothetically strengthen any common associations. The technique, however, does not account for different variants within the same region exhibiting different effects. For example associations in the *IL2RA* and *IL2RB* regions may not be detected if this method was utilized; therefore it is potentially not suitable for the current data. Another method, which could be employed, is cross phenotype meta-analysis, which has been previously used to identify overlap between 8 immune mediated traits (Cotsapas et al. 2011). This technique would be ideal for identifying different clusters of genes, which are associated commonly, and differentially with different types of IA. In future studies, such techniques are likely to be used to assess genetic overlap in more

detail.

Once *RUNX1* was selected as the region for overlap using a number of parameters, bioinformatics analysis was performed for rs9979383 (and its complete proxy, r^2=1, rs8129030). Further evidence was then identified to support it as a strong choice for further investigation. The variants associated with IA are located in a long non-coding RNA (lncRNA) on chromosome 21, upstream of the *RUNX1* gene. These variants are also located in a region of open chromatin, which moderately binds the p65 (REL) transcription factor, an essential component of the NFKB pathway, in lymphoblastoid cell lines treated with TNF (Karolchik et al. 2014).

Furthermore eQTL analysis using the SCAN database (http://www.scandb.org; (Nicolae et al. 2010) indicated that the variants correlated with increased expression of the *RBP5* gene on chromosome 12 and therefore could represent a long-range trans-eQTL. The bioinformatics analysis therefore provided a wide variety of hypothetical functional consequences for the overlapping region, which could be explored experimentally.

Once a literature search was performed it was identified that *RUNX1* is a member of the *RUNX1* family of transcription factors, which also includes *RUNX2* and *RUNX3*. These regions have also been associated with immune mediated disease (Alarcon-Riquelme 2003). *RUNX1* itself is highly expressed across a number of cells and has a large number of transcripts (Levanon et al. 2001). The principal function of *RUNX1* is the regulation of a number of genes crucial for processes such as haematopoiesis, lymphocyte differentiation and chondrogenesis (Ichikawa et al. 2008) making it a strong candidate for follow up in an inflammatory arthritis setting.

In particular *RUNX1* has been shown to be important in chondrogenesis by promoting cell maturation whilst regulating production of MMPs in the joint. *RUNX1* expression has been shown to be dysregulated in osteoarthritis, indicating downregulation of *RUNX1* could potentially result in joint hypertrophy characteristic of this disease (Yano et al. 2013). Alteration of this mechanism is currently being investigated as a potential drug target. Furthermore early results indicate that dysregulation of *RUNX1* expression may be linked to differential DNA

methylation between OA patients and healthy controls (Jeffries et al, personal communication).
RUNX1 also appears to be particularly important for lymphocyte cell development, which have been shown to be essential mediators in driving inflammation. By driving cell polarization, *RUNX1* has been shown to regulate differentiation of both CD4+ and CD8+ lymphocytes inclusively (Komine et al. 2003;Lazarevic et al. 2011). In Treg cells *RUNX1* also plays an important role, interacting with FoxP3+ to suppress transcription of inflammatory cytokines such as IL2 and IFNγ (Ono et al. 2007) which means it could potentially be a key player in IA pathogenesis.

RUNX1 binding sites have been shown to be associated with several autoimmune diseases including RA, PsV and SLE (Alarcon-Riquelme 2003). This observes that it is associated with more than 1 type of inflammatory arthritis particularly interesting. In addition SNPs in this region have been shown to interact with a number of IA susceptibility loci such as *AFF3* and *IL2RA* . Collectively these findings made *RUNX1* a strong choice for follow up using further genetic and functional techniques.

The availability of these databases allowed data to be collated on eQTL studies in different cell types, gene splicing and TF binding potential of regions which housed overlapping variants. Although an excellent resource, it is worth noting that the databases are not always extensively populated and may not include sufficient data for a region. The *RUNX1* region is not extensively covered by genotyping and expression arrays and is often not selected for ChIP and 3C experiments. That means that searches may be returned as null when really they have just not been performed yet. In the data which is actually available, experiments are often performed in specific cell subsets, which may not be applicable to specific diseases, and in very small sizes, which may not have been replicated in more than one sample. In the functional data which was obtained for *RUNX1*, both trans-eQTL and TF binding evidence was obtained. Although providing a good functional prediction of the region, lack of reliability in trans-eQTL studies mean that a finding has to be independently replicated before being confirmed (Westra et al. 2013). Furthermore the low number of replicates and variability in antibody binding used in ChIP experiments mean that these results also need to be replicated independently before they can be considered genuine functional

predictors.

Overall for this overlap study, the utility of Immunochip data, cohort control sizes, extensive genome coverage and availability of functional data were the main strengths for identifying genetic overlap between different types of IA. Limitations including low power in the more modestly sized JIA/PsA cohorts, the use of basic methods to define overlap and incomplete bioinformatics data within the region have been identified and potential solutions to these will be addressed in more detail in section 4.4.

4.2.2 Immunochip replication

Once a novel association is identified by a GWA or candidate gene study, it is desirable that is replicated in an independent cohort to provide further evidence that it is a true association. This method has been successful in confirming many susceptibility loci for IA including *STAT4* and *IL2RA* (Kurreeman et al. 2009;Orozco et al. 2008). As many of the 50 overlapping regions identified in this study contained variants which had not been previously associated with IA, a replication study was performed in an independent RA cohort. 7 overlapping regions were selected for replication based on their significance; however some regions were excluded, as they were being genotyped in additional studies, taking place within the department at the time. Duplicating genotyping of SNPs would represent an unnecessary waste of cost and resources. The index SNPs from each disease for the 7 regions selected were used to design a multiplex Sequenom assay, which allows multiple SNPs to be genotyped simultaneously in each sample. The method involves designing specific PCR and single base extension primers for each SNP, which have to be tolerated together as a SNP multiplex. In several cases the SNPs selected could not be tolerated together in a multiplex so high LD proxies (r^2>0.9) had to be included instead. Although SNPs were selected to capture the 7 overlapping regions, genotyping only the index SNP from each disease meant that only a limited amount of variation was captured in each region. Ideally a number of tag SNP variants would be genotyped in each region, allowing analysis to identify multiple effects in each region; however, this is a costly option, which may not be justified as many of the associations were not strongly associated in the

Immunochip analysis and the power to replicate the associations in the smaller RA replication cohort available was limited. Further fine mapping could be performed once the SNP association was replicated.

In total 12 SNPs from 7 overlapping regions were included in the multiplex assay. RA samples were chosen as the replication cohort as it is the most common of the 3 diseases and has been studied most extensively; therefore, it is the disease with the largest number of independent samples available for genotyping. Initially a total of 3879 cases and 2561 controls were selected for genotyping, which provided sufficient power to detect common SNPs (MAF>0.05) with larger effect sizes (OR >1.3).

Once genotyping was performed and SNP calls were assigned, it was observed that there was a low call rate and high assay failure across several sample plates. It was hypothesized the most likely source of this was low DNA concentration and/or quality. When low concentration or quality DNA is present, the PCR amplification and single base extension reactions in the Sequenom MassARRAY genotyping protocol do not work efficiently. This can lead to incorrect or null genotype calls, resulting in low call rates. However, as each set of PCR and single base extension primers are designed uniquely for each SNP, there may be differences in assay efficiency. This may result in some assays being able to tolerate low quality DNA whilst others cannot. As DNA QC was not performed for the samples prior to genotyping the results from the gel electrophoresis performed after the initial PCR reaction were consulted for each SNP. This was performed across a number of plates and the results compared to the genotype calls for the same plate. In plates, which had a large number of 'no call' samples or poorly defined clusters, it was found that there was a correlation with the presence of weak or absent PCR product bands on the agarose gel. Furthermore these sample plates had a higher prevalence of complete assay failures. In contrast samples with high call rates and minimum assay failure correlated with the presence of bright bands on the agarose gel. It was concluded that the large number of low call rates across the sample plates was due to the presence of low quality and concentration samples on those plates, which were removed from subsequent analysis. Hence, a strength of the study is the stringent QC that was performed so that only robust SNP assays were included in the final analysis, reducing the risk of false positive findings.

The availability of an independent cohort of RA cases and healthy controls, which were not genotyped as part of the Immunochip study, is a major strength of this study. However, the loss of many samples due to low quality DNA meant that the final number of samples used in the study was much lower than expected. This resulted in a decrease in the power to detect genetic effects. Low DNA quality may have been a consequence of poor extraction, storage or handling of samples. As the samples were extracted at a number of different sites and were stored long term without QC, it is challenging to determine which factor was responsible. To account for these issues stringent QC was performed and post quality control association testing was performed on genotyping data from 9 SNPs in 2595 cases and 1636 controls. This resulted in a significant decrease in power to detect even the larger effect sizes described previously. Despite limited power, association with 2 SNPs in the *CTLA4* and *MTMR3* regions was replicated in the independent RA cohort at a $p<0.05$, although it should be noted that no correction for multiple testing of 9 SNPs was applied. As the aim of this study was to replicate overlapping regions, a further limitation is that only samples from RA were tested. In future association with the region will need to be replicated in independent JIA and PsA cohorts to provide evidence that true disease overlap exists.

Interestingly, for both the *CTLA4* and *MTMR3* regions it was the JIA index SNP from the Immunochip study, which was most significantly associated in this RA cohort. In the *CTLA4* region, this may be because this was the only SNP which was genotyped in this independent RA cohort and if the RA SNP was genotyped it may have shown stronger association but in the *MTMR3* region both were genotyped. This provides further evidence that the effects may be the same between these diseases despite different index SNPs being identified in the Immunochip study.

Although the *RUNX1* region represented a novel association for RA, JIA and PsA, it was not included in the Immunochip replication study described in this section due to a number of factors. Through bioinformatics and literature searching, this region was selected as a strong functional candidate and therefore was selected for further genetic analysis. As this region had low SNP coverage on the Immunochip

array, it was decided that both replication and fine mapping of the region would be performed.

4.2.3 RUNX1 fine mapping and replication

Although a large number of regions were extensively fine-mapped on the Immunochip array, the *RUNX1* region itself was not well covered. The Immunochip array capture of common variation (MAF>0.05) from the 1000 genomes (July 2010 release) in the *RUNX1* region was calculated to be 7% using an r^2>0.8. This represents very low tag SNP coverage; therefore it could not be determined from the Immunochip alone that the index SNP rs9979383 was the strongest or the only associated genetic effect in the region. In order to replicate the association in an independent cohort and provide sufficient coverage of this region, a combined fine mapping and replication study was performed. Fine mapping is a method, which uses dense SNP genotyping to capture as much variation in a region as possible. The aim is to allow localization of association signals and identify if there are multiple genetic effects within a region. It can be performed on a region-specific basis or across multiple regions, such as the strategy taken in the Immunochip project, where many regions did have a dense SNP coverage. Fine mapping itself has been very successful in localizing association signals and identifying multiple effects such as those identified in the *STAT4* and *TNFAIP3* region (Orozco et al. 2009;Remmers et al. 2007). An advantage of my study is that it allowed region-specific fine mapping to be performed in an independent RA cohort. This allowed variation in the region, which was not included on the Immunochip array, to be captured and analysis to be undertaken to explore whether multiple effects existed in the region.

As the index SNP rs9979383 lies between 2 peaks of high recombination, it was hypothesised that the causal variant was also likely to be located between these 2 points. In order to capture as much variation as possible, a tag SNP approach was adopted. In total 2 multiplex genotyping assays including 51 SNPs were designed, capturing 75% of the total common variation (MAF>0.05) at the locus. One limitation, therefore, is that not all known variation was captured but the

percentage coverage represented a major improvement from that included on the Immunochip array.

As with the Immunochip overlap replication, the cohort selected for this study was an independent RA cohort, which had not been genotyped on the Immunochip. Initially the 51 SNPs were genotyped in 3491 RA cases and 2359 controls but, as with the Immunochip overlap replication, issues with DNA quantity and quality resulted in a large number of failing samples. In total genotype data was available for 42 SNPs in 2359 cases and 1877 controls which resulted in limited power to detect common associations with smaller effect sizes (OR<1.3). Furthermore although the study was designed to replicate rs9979383 in an independent RA cohort, it has yet to be replicated in JIA and PsA cohorts, which is necessary to confirm it as a true overlapping locus.

Association with the Immunochip index SNP rs9979383 was replicated at p=0.02. Although this is a much less significant p value than the Immunochip study, the sample size tested (and hence the power) was much lower. Furthermore, no correction for multiple testing has been applied so the association may be the result of type I error (false positive). However, the OR for this study was almost identical to that of the RA Immunochip study at OR=0.9, which provides support that it may represent a true genetic association. Although rs9979383 is the most promising candidate for a causal variant in this region, it cannot be excluded that the causal variant lies in the uncaptured variation, due to incomplete coverage. To capture all variation, additional genotyping of SNPs or target DNA sequencing of the region will be required, which is discussed in more detail in section 4.4.

To identify if there were multiple genetic effects, conditional logistic regression was performed conditioning on the lead SNP. Once conditioning was performed, no other SNPs remained significantly associated, indicating that no other genetic effects were present in the region. This may be a true reflection on the genetic architecture of the region but it is worth noting that the sample size used in this study was modest and thus power was limited to detect additional effects and a

larger sample size may identify multiple genetic effects.

As rs9979383 lies ~270kb upstream of the *RUNX1* gene itself and was identified as having a histone modification profile indicative of an enhancer region, it was hypothesized that it may affect expression of genes in the *RUNX1* region. To investigate this, online eQTL databases were consulted but due to poor coverage of the *RUNX1* region on arrays, minimal data was found. With this is mind, an eQTL study was designed to assess whether the variant acted as a cis-eQTL in a variety of relevant tissues.

4.2.4 Whole blood eQTL analysis

An eQTL is a single base change in DNA, which results in an alteration of expression of a nearby gene. This may occur at close range affecting a nearby gene (cis-eQTL) or long range across the genome (trans-eQTL). In several studies eQTLs have been shown to be important to disease susceptibility, through their alteration of gene expression (Anon 2013;Yang et al. 2010) . In order to perform an eQTL study, both genotype and gene expression values must be obtained for a sample. The availability of samples from the NHRV study to perform both genotyping and gene expression represented a major strength of the current study.

In order to identify whether rs9979383 altered expression of the *RUNX1* gene, a cis-eQTL study was performed using whole blood from healthy volunteers. This involved the collection of 2 peripheral blood samples; 1 for DNA extraction and another for total RNA. Whole blood is often selected as the tissue of choice for expression studies as it is the easiest to access and store long term (Emilsson et al. 2008;Westra et al. 2013). These advantages allow the collection of large numbers of samples for analysis, which would not be possible with other less accessible tissues. Given that eQTL SNPs have large effects on expression levels, smaller sample sizes are often able demonstrate significant effects and therefore this sample size was potentially sufficient to detect this eQTL (Hunt et al. 2008;Stranger et al. 2007).

Healthy volunteers were chosen for this study rather than disease cases. This is because patients with these diseases are often on medication and may have comorbidities that could interfere with the interpretation of findings. The initial hypothesis was that the functional consequence of genetic variation at rs9979383 was regulation of the *RUNX1* gene, so initially; a Taqman allelic discrimination assay and a Taqman gene expression assay were designed. As 19 splice variants of the *RUNX1* gene have been identified, to date, a gene expression assay was designed to capture as many of these isoforms as possible. By targeting exons 5 and 6, the assay captured 14 of the 19 splice variants. Limitations are, first that not all isoforms were captured and, second, relative expression of the different isoforms could not be tested because the assay cannot discriminate between different splice variants. It has been shown recently that gene splicing holds an important role in disease susceptibility and therefore capture of all splice variants is desirable (Wang et al. 2012). Although designing more Taqman gene expression assays would be one solution, the most practical approach would be to perform RNA re-sequencing of the region in samples of different genotypes at rs9979383. Such an experiment would be able to determine whether the genotype has any effect on the levels of *RUNX1* splice variants and is described in detail in section 4.4.

A strength of the study was the use of two endogenous controls, *GAPDH* and *ACTNB*, for *RUNX1* normalization, as they were recommended by the manufacturer. Endogenous controls are desirable for Taqman gene expression assays as they allow normalization of gene expression values to account for differential expression values between samples as a result of both the technology and the differential expression between cells. This is particularly important as whole blood represents a very heterogeneous tissue and therefore expression values may differ between samples with different cell composition. The use of an averaged value from two controls provides an extra level of quality assurance. However, no association of the variant with *RUNX1* gene expression was detected in whole blood. This finding has several possible interpretations. Firstly, this may indicate that the variant does not represent an eQTL and acts in a different way to contribute to IA susceptibility. Some of these potential roles are discussed in more

detail in section 4.4. Secondly, the variant may be affecting another gene in the region or across the genome as a trans-eQTL. Finally, the heterogeneous cell composition of whole blood means that weaker eQTL signals from low frequency cell types are masked by the most prevalent cell types. This is particularly an issue in whole blood, due to its heterogeneous cellular composition. To determine whether this is the case a cell specific eQTL study can be performed such as that described in the next section.

One of the strengths of the study is the sample size tested, 70, which is regarded as reasonable for this type of work. However, the study was limited by the distribution of different genotypes between the samples, with only 1 sample being homozygous for the minor allele. This lack of distribution was unexpected, as rs9979383 had been shown to have a MAF of 30% in the 1000 genomes CEU population, the majority of whom are healthy individuals (July 2010 release). This means that a lack of significant correlation with genotype may be a consequence of limited sample size and therefore collection of more samples is required to investigate this further. Using healthy controls in this study allowed investigation of the role of an overlapping variant in regulating *RUNX1*. By using healthy controls the effect of the variant can be examined without interference by disease mechanisms or treatment. On the other hand it would have been desirable to examine this affect in the disease cases, as an eQTL might only be present as a consequence of disease or response to treatment. Therefore this analysis should be repeated using case samples in the future.

As *RUNX1* represented the most likely candidate gene in the region to be affected by rs9979383, it was the gene selected for this eQTL analysis. Therefore other genes in the region were not tested. It may be the case that rs9979383 is not altering gene expression of *RUNX1* but is actually influencing expression of another gene through a cis or trans-eQTL. In particular, it would be interesting to investigate the trans-eQTL with the *RBP5* identified through the bioinformatics analysis in section 3.1.6.1. It may also be that this SNP does not represent an eQTL and is contributing to disease susceptibility via another mechanism. Further investigation to address this is described in section 4.4.

Using whole blood as the tissue of choice in this study has both its strengths and weaknesses. Whole blood is one of the most easily accessed tissues, with only a small volume required to obtain DNA and RNA. However, whole blood is also a very heterogeneous tissue, containing a large number of very different cell populations. These populations all vary in gene expression and frequency, therefore gene expression signatures from low frequency populations are often masked by the other more frequent cell types. The presence of globin transcripts present in whole blood can also contribute to low detection of gene expression by techniques such as Taqman gene expression assays (Wright et al. 2008). In recent years it has become apparent that the most effective method for analysis eQTLs is using highly pure single cell populations . The method allows determination of gene expression in a group of more homogenous cells, which can then be correlated with genotype from whole blood.

4.2.5 eQTL analysis in T lymphocytes

To investigate whether rs9979383 regulates expression of *RUNX1* in a population of T lymphocytes, a cell specific eQTL study was performed using samples from 23 healthy volunteers. To determine gene expression in the *RUNX1* region in CD4+ and CD8+ lymphocytes, a series of cell separations followed by gene expression analysis using a whole transcriptome array was performed. T lymphocytes were selected for analysis as a consequence of their involvement in IA pathogenesis. CD4+ T lymphocytes have been shown to be a critical cell type in RA, as analysis of cell specific chromatin marks in RA susceptibility regions have been shown to be enriched in this cell type (Trynka et al. 2013). More recently, it has been shown that the frequency of CD8+ T lymphocytes is increased in the synovial joint of PsA patients compared to RA, indicating that these cells have a role to play in disease pathogenesis (Menon et al. 2014). Furthermore *RUNX1* has been shown to be an important transcription factor in T lymphocyte lineage differentiation which makes it an excellent cell choice for this functional experiment (Komine et al. 2003) . It has also been shown that *RUNX1* is important for chondrogenesis and cartilage production during OA pathogenesis (Blanco and Ruiz-Romero 2013; Yano 2013). The region has been shown to be differentially methylated in chondrocytes

from OA and non-OA tissues, indicating that the region may play a role in disease pathogenesis at a tissue specific level (Jeffries et al, personal communication). As OA also involves articular disease, it is possible that the finding may also be true in IA.

One of the major strengths of the experiment performed is that primary cells were tested rather than immortalized cell lines. Although cell lines are useful for providing a large, completely homogenous cell population, they have been altered in such a way that the gene expression profiling from these cells may not reliably represent a primary human cell (Komine et al. 2003;Putz et al. 2012) . Cell lines are often only derived from a single donor, which makes performing large high-powered studies challenging.

A strength of the study was the high viability and numbers of PBMCs obtained. To allow for collection of the 23 samples tested and to minimize sample batch effects, samples were cryopreserved in liquid nitrogen vapour phase prior to separation. The technique is widely used to allow collection of a large number of cell samples but studies into the effect of cryopreservation on differential gene expression in PBMCs have yielded very inconsistent results (Chen et al. 2010). Although this is an issue to be considered, all samples were cryopreserved for a short time (<28 days total) and therefore were expected not be extensively affected by the process. To assess the effect of cryopreservation on the viability of the cells, a viability check was performed post-separation using flow cytometry.

Post-cryopreservation, CD4+ and CD8+ lymphocytes were separated from each PBMC sample using Miltenyi MACS magnetic cell separation. MACS separation was selected over FACS sorting and other magnetic separation methods, as it represented the best quality method for a number of reasons. Firstly, as this strategy uses an automated separator, 6 samples could be separated simultaneously therefore allowing large numbers of samples to be processed together, which in turn minimized the risk of batch effects being generated during the process. It also allowed the separation of several subsets of cells from the same blood sample. Although FACS would allow sorting of multiple populations from the same sample, this would have used a much larger quantity of antibody and would

take a much longer time as each sample needs to be processed individually at a slow rate, potentially adding to batch effects. Therefore MACS was selected to ensure the separation of high purity cells across a large sample size quickly and easily without excessive costs.

The use of automated magnetic separation allowed the option of either positive or negative selection strategies to be considered. Positive selection involves the direct targeting of a specific population of cells with antibodies conjugated to a magnetic bead. These can then be removed using a magnetic field, resulting in a highly pure homogenous cell population. In contrast, negative selection involves the targeting of every other cell population, except the population of interest, with a magnetically conjugated antibody. When these are passed through a magnetic field, all non-targeted cells are removed, leaving the cell population of interest. Positive selection generally results in significantly higher purity of cell populations than negative selection but there is an important factor to consider. Although the antibodies used to target cells during positive selection are designed to avoid aberrant activation of cells, cell markers such as CD4+ and CD8+ are often involved in cell stimulation. This means that when an antibody binds the marker it may result in changes to the cell. This in turn could potentially result in changes in the cell transcriptome which could potentially affect the results of a study. To date limited data has been produced showing that this is a concern, though it is certainly something worthy of consideration. The most important factor in deciding between these strategies was the ability to separate different cell populations from the same sample. This is only possible with positive selection as, during negative selection all other cells are removed from the sample and cannot be recovered. As the aim in this study was to separate both CD4+ and CD8+ T lymphocytes from the same sample, positive selection was, therefore used.

The choice of a positive selection strategy was also supported by the need for high purity cell populations when analysing the transcriptome of a homogenous population of cells. If contaminating cells were present, this could result in inaccurate gene expression levels being detected. Furthermore positive selection allowed a double column strategy to be used in which the cells were passed

through 2 magnetic fields, therefore providing a greater chance of a highly pure cell population. Consequently choosing this strategy resulted in very high average purities of 99.4% for CD8+ lymphocytes and 95.31% for CD4+ lymphocytes, which are sufficient for gene expression analysis. Although an insignificant amount of contamination was observed in the CD8+ lymphocyte populations (0.06%), slightly more was detected in the CD4+ lymphocytes (4.69%). This is likely to be due the fact that the CD4+ separation represented the third cycle in the technique, meaning that the cells had passed through the magnetic field and undergone centrifugation several times resulting in more damage than in the first cycle. This was supported by the findings in the viability screening, as the CD8+ population had an average viability of 80% compared to the CD4+ population at 67%. The drop in purity may also be due to a small number of monocytes, which were not removed by the CD14+ selection being picked up in the CD4+ separation, due to their low expression of the CD4 marker (Kazazi et al. 1989) . Overall, this small contamination was not considered a significant issue as any alteration of the gene expression signature caused by the contaminating cells would be masked by the high number of cells of interest. As all the samples obtained were highly pure, all were processed for RNA extraction and subsequent whole transcriptome gene expression analysis.

Once RNA was extracted from all samples, QC was performed using the Agilent Bioanalyzer and Nanodrop N-1000 to assess the concentration and calculate the RIN values for each sample. Post QC, it was noted that several of the samples had lower 260/230 ratios than expected. This appears to be due to minor Trizol carryover from the RNA extraction protocol as it absorbs at 230nm on the bioanalyzeN-1000 absorbance spectrum. Normally this small contamination would not be detected but the small quantities of RNA obtained from small cell populations mean that it becomes apparent. Although minor contamination is not likely to alter gene expression detection, a DNase treatment, which included an acid chloroform re-extraction and several clean up stages, was selected to optimize the purity of the sample, without compromising integrity. The DNase treatment was performed as RNA contamination by genomic DNA has been shown to actively alter gene expression analysis and therefore should be addressed during sample processing (Naderi et al. 2004). Once all samples had been processed, 45 (23 CD8+

and 22 CD4+) were converted to cRNA for analysis using the Illumina Human HT expression array.

Once all 45 samples were analysed using this array, quality control and normalization was performed. Normalisation is essential for whole transcriptome data as it is required to account for differences in samples and probes across the array so that comparison of gene expression values can be made across samples and probes. One of the strengths of the study was the extensive QC that was undertaken during analysis. This included, checking the signal to noise ratios for each sample and using this as a covariate in subsequent analysis; ensuring that the number of probes expressed was not significantly different between the CD8+ and CD4+ samples; ensuring that all probes included in analysis correlated with a transcript and removing those that did not; performing MDS analysis to identify sample outliers and performing PCA analysis. The PCA analysis reassuringly showed that the sample characteristic accounting for most of the variance between samples was cell type.

Once normalization and QC was complete, the gene expression data for all probes mapping within 400kb of the *RUNX1* gene were selected for eQTL analysis. This included data for 2 probes mapping to the *RUNX1* gene and 2 probes mapping to *LOC100506403*, a lnc-RNA which is upstream of *RUNX1*. To assess whether rs9979383 represented an eQTL, gene expression at these 4 probes was correlated with genotype using linear regression. Of the 23 selected for analysis, 22 had genotypes for rs9979383 which were generated in the previous whole blood eQTL study. As with the whole blood analysis, no significant eQTL was detected between genotype and any of the probes, with the smallest p value being observed with *RUNX1* in CD4+ cells (p=0.1). Again as these samples represented a subset of the samples used in the whole blood analysis, there was limited distribution of sample genotypes, with only 1 sample representing double carriage of the minor allele. Although the results represent a more convincing trend to that seen in the whole blood analysis, this does not provide enough evidence that rs9979383 represents an eQTL and therefore further analysis is required such as that described in section 4.4.

At 23 healthy volunteers, the sample size used in this study is significantly smaller than consortium driven eQTL projects such as the Immvar project (Towfique et al, personal communication). Due to the lack of large scale expression projects it is challenging to tell at exactly what point the study will be well powered enough to detect a genetic effect. As seen in the whole blood analysis, the study was also limited by lack of genotype distribution, as only 1 sample was used which carried the double minor allele. If this sample size was to be increased to include more samples from the heterozygous and double minor allele groups, it would have greater power to detect genetic eQTL effects. As this variant has a 30% frequency in the healthy population, this is potentially a strategy to consider.

CD8+ and CD4+ T lymphocyte subsets were selected as cells of interest in this study, due to their involvement in RA and PsA, respectively. Although a very hypothesis driven decision at the time, new evidence indicates that *RUNX1* may be acting as a mediator of inflammation via its effect on chondrogenesis in the damaged joint (Blanco and Ruiz-Romero 2013). This may be one of the reasons that an eQTL has not been detected in CD8+ and CD4+ cells as they are derived from PBMCs whilst this effect may only occur exclusively in the joints of patients. It would be desirable that chondrocytes are also included as a cell type of interest, in any further work.

To ensure that a large number of primary cells from donors could be collected during this study, cryopreservation techniques were adopted. Although this is a standardized technique across the primary cell field, some papers have shown that long-term cryopreservation can affect the overall gene expression profile of PBMCs. It was ensured in this study that samples were not stored in liquid nitrogen vapour phase for more than 26 days, therefore minimizing the risk of changes in gene expression. To fully assess whether this had an effect, a study comparing the effects of short-term cryopreservation in these samples would have to be performed.

Once the cells, were separated total RNA was extracted using Trizol chloroform, followed by a DNase treatment to remove genomic DNA. Overall this technique

resulted in the extraction of good quality RNA overall which could be used for gene expression analysis. For the test, the Illumina Human HT whole transcriptome array was used. This array contains probes for 47,000 gene transcripts at a genome wide level. Array technology is greatly advantageous as a large volume of data can be generated from each sample, though it does come with some issues. Despite having extensive coverage of the genome, the array does not always account for the large number of transcripts a gene may have. For example, the *RUNX1* gene investigated has 19 identified transcripts but the array is only capable of capturing 2. This means that others may not be detected, which may be regulated by an eQTL effect. In order to capture all variants a Taqman gene expression assay or RNA resequencing must be performed as described in section 4.4.

4.3. Implications of study

4.3.1 Immunochip overlap

Overall, the findings in this study have led to a greater understanding of the genetic overlap between different types of IA. In the Immunochip overlap analysis 50 genetic regions, were identified as being associated with more than one type of IA at $p<1x10^{-3}$, which is a significant addition to previous knowledge which had identified just 7 regions. Many of these regions also represented novel IA associations and therefore signify a leap in knowledge for these diseases. 32 of these 50 overlapping regions were associated with both RA and JIA inclusively, indicating that the greatest genetic overlap appears to be between these 2 diseases. This is further supported by the identification of overlap between these diseases in other independent studies (Hinks et al. 2010b;Hinks et al. 2012). Despite this it is also worth keeping in mind that this study was performed using samples from different subsets of RA and JIA and that further analysis stratified by disease subset will be required to fully explore this. Nonetheless, the results indicate that common pathways contribute to these diseases, which if identified could be used to re-classify diseases according to the primary pathway involved, direct the use of current therapies and identify targets against which to develop

drugs that could be used to treat all three types of IA. The development of these multi disease therapies could potentially target a common pathway, which may avoid the blanket immunosuppressive effects which are an issue for current IA therapies.

When the most associated/index variants within the 50 overlapping regions were compared between different diseases, extensive similarities and differences were identified, indicating that genetic overlap is not always clear-cut. In 14 regions the SNP was found to be identical or highly correlated by LD (r^2>0.8). This included regions such as the *PTPN22*, *UBE2L3* and *TYK2* regions. In many cases the index SNP associated with more than 1 type of IA is also identical to SNPs which are associated with other immune mediated diseases. For example in the *TYK2* region, the SNP which is associated with RA and JIA (rs3453663) is also associated with MS and PsV (Beecham et al. 2013;Tsoi et al. 2012) . Another example is in the *PTPN22* region, in which the same SNP rs6679677 is associated with RA, JIA, and T1D (Barrett et al. 2009). This indicates that these diseases share common genetics and therefore potentially pathways which contribute to pathogenesis. Further investigation is required to identify the pathway in which these variants contribute to and learn more about the common processes driving these diseases.

In contrast, in the majority of regions (31 regions), the index SNPs were completely different between diseases and were not correlated by LD (r^2<0.4). This is perhaps the most interesting finding as it shows that although a particular region is implicated in disease pathogenesis, there may be different effects present in the region which could be potentially responsible for the differences between the diseases. Alternatively, these different variants could be very independent but act on the same redundant pathway and therefore produce identical effects. Two regions which are associated with more than 1 type of IA but contain variants which are not correlated are the *IL2RA* and *IL2RB* regions. In the *IL2RB* region, completely different variants are associated with RA compared to JIA. Furthermore the direction of the genetic effects appears different, with the minor allele of the RA SNP conferring disease risk, whilst the minor allele of the JIA SNP confers disease protection. A similar situation has been observed in the *IL2RA* region, where two uncorrelated SNPs within the same region confer different directions of

effect in each disease. It has since been shown through conditional analysis that 2 independent genetic effects exist for JIA, whilst in RA there is only 1 (Eyre et al. 2010;Hinks et al. 2013). . When subjected to further analysis, it appears that secondary genetic effect in the region in JIA is identical to that identified in RA, which indicates there is in fact a shared association within this region. This is a particularly interesting finding as *IL2RA* and *IL2RB* genes both encode subunits of the IL2 cytokine receptor. The receptor is essential for the activation and proliferation of T lymphocytes and other essential immune cells. The finding shows that although the association in these regions appeared to be very different in the initial analysis, undetected overlap may be found by further investigation of the genetic architecture within the region.

In total 9 regions were found to be associated with all 3 types of IA which is an exciting finding, as only 1 pan-IA region had been identified prior to this study. Although this could represent very different associations within the same genetic region, these regions are of particular interest as they indicate that the same genes could be driving the 3 diseases. Furthermore, several of these regions have been associated with other immune mediated diseases, indicating that these variants may be causing a significant dysregulation of the immune system which is capable of manifesting itself as very different pathologies. In two regions (*RUNX1* and *TYK2*), an identical or highly correlated SNP was found to be associated with all 3 types of IA studied but only in the *RUNX1* region was the SNP found to be exactly the same between diseases. This unique region was selected for further investigation as the positioning of the SNP and the previously characterized functional role of the gene made this an excellent candidate to be contributing the IA pathogenesis. It also represented a completely novel association for all 3 diseases and therefore further analysis of the region was required.

4.3.2 Immunochip replication

The SNPs which were replicated in this study were located in the *CTLA4* and *MTMR3* regions which are particularly interesting themselves. The *CTLA4* region has been associated with RA in a number of different cohorts and has been

suggestively associated with JIA . It has also been associated with a number of immune mediated diseases such as T1D and ATD (Barrett et al. 2009;Cooper et al. 2012). Therefore, this study adds to the existing body of evidence that the gene is associated with disease susceptibility to RA and JIA (Fairfax et al. 2012). On the other hand *MTMR3* had not previously been associated with any type of IA, although it has been associated with both T1D and IBD (Hoefkens et al. 2013) This therefore represents a novel and exciting region which should be subjected to further investigation in the future.

Although it was disappointing that none of the SNPs included in this study was associated at a genome wide significant level with RA, they did show a trend towards association which is promising. The most likely explanation for this is the lower power as a consequence of sample size which was unavoidable in this case. This analysis ideally this should be repeated in larger JIA and PsA cohorts which could potentially be collected through collaboration with other investigators. This is discussed in more detail in the section 4.4.

4.3.3 RUNX1 replication and fine mapping

The SNP rs9979383 replicated successfully at p=0.02, which in a study of this size is an acceptable indication of association. Furthermore with an odds ratio of 0.903 in this cohort, this was identical to the odds ratio of 0.91 seen in the RA Immunochip cohort. This provides further evidence that the association may represent a true RA susceptibility variant. Unfortunately this could not also be confirmed for JIA and PsA, due to lack of sample availability. In future this panel of SNPs should be genotyped in JIA and PsA samples in order to replicate it in these disease cohorts but is subject to limitations as described previously.

To identify if any additional effects were present within the region, conditional analysis was performed. By conditioning on rs9979383, the aim was to identify all SNPs which are strongly correlated with rs9979383 and therefore detect if any additional independent effects exist within this region. This strategy has been essential in identifying regions which contain multiple genetic effects such as the *STAT4*, *TNFAIP3* and *PTPN2* regions which are all IA susceptibility regions. No

additional effects were identified however the limited power of the study is worth considering. This leads to the idea that rs9979383 (or markers in linkage disequilibrium with it) represents the true causal variant within the region but this will not be confirmed until complete coverage of variance in the region is captured. Regardless, the study shows that fine mapping is an effective strategy for capturing a large proportion of the genetic variants in a region as possible.

Given that the most associated variant in the region, rs9979383 is located approximately 280kb upstream of the *RUNX1* gene and is located in a region with a histone H3K4Me1 mark of an enhancer, it was hypothesized that it may regulate gene expression by a cis-eQTL effect. This was further supported by evidence that the *RUNX1* gene has 2 promoters, distal and proximal, which means that this SNP could potentially lie directly in the distal promoter (Sroczynska et al. 2009) and therefore effect the regulation of the *RUNX1* gene.

4.3.4 RUNX1 eQTL analysis in whole blood and T lymphocytes

Although the current study did not identify an eQTL with rs9979383 and *RUNX1* in whole blood or T lymphocyte subsets this has been a valuable contribution to the *RUNX1* story and has shaped the studies which will be performed in this region in the future. Furthermore although it looks likely that this SNP or its proxies may not represent an eQTL in these tissues, this may have been a consequence of several limiting factors. Firstly as the studies were only performed in 75 and 23 healthy volunteers respectively, the power to detect an eQTL effect may not have been adequate. Since then a number of larger eQTL studies have been initiated by international consortiums such as the Immvar consortium, with data anticipated to be released very soon (Towfique et al, unpublished). Although limited in size it was expected that any larger cell-specific eQTL effects would be picked up in this sample size and since then several interesting eQTLs have been picked up using this dataset, indicating there was sufficient power to detect genetic effects (Bowes et.al, Manuscript in preparation). For example, a SNP mapping the 5q chromosomal region associated with PsA has been shown to be an eQTL with the

SLC22A5 gene in CD8+ T cells.

Although the cell-specific eQTL study was designed to investigate cell types of importance in IA, as whole blood represents a very heterogeneous tissue, an eQTL may be present in a cell type which was not captured in this study. This is particularly important as the study was performed in blood whereas *RUNX1* has been shown be important for chondrogenesis; therefore, an eQTL may only be observed in joint tissue (Wang et al. 2013b). Other cell types such as B cells have been shown to be particularly important in RA and JIA as both effector cells and producers of autoantibodies (Mauri and Bosma 2012;Prakken et al. 2011). Most recently a particular FcRL4-expressing set of B cells have been shown to be present in the RA synovial joint and therefore signify a cell type which should be investigate in a future eQTL study (Yeo et al. 2014) . Another possibility is that although the CD4+ and CD8+ T lymphocyte compartments were investigated in this study, a signal may be present in a low frequency subset of these cells such as T regulatory (Treg) or naïve T cells. Again, as with the whole blood study this may lead to signals being masked in a heterogeneous cell population and therefore a more refined study is required to investigate this further.

Another very important issue with this study is that it was performed in both healthy controls and using unstimulated primary cells. This means that eQTL signals which are a consequence of the inflammatory process driving IA would not be picked up. In a recent study comparing cis-eQTLs in interferon gamma stimulated (IFNγ) and unstimulated monocytes, a large proportion of the eQTLs identified were exclusive to the stimulated cells indicating that in disease studies it may be essential to stimulate cells with an inflammatory mediator such as IFNγ or TNF-α (Fairfax et al. 2014).

Although the *RUNX1* gene seemed like the most likely candidate for an eQTL effect, it may be the case that it is a totally different gene which is regulated by rs9979383. This will require further investigation through gene expression studies which capture all the genes in the surrounding region, with *LOC100506403* long coding-RNA (lncRNA) representing a particularly interesting candidate for future investigation. lnc-RNA's represent non-protein coding transcripts which are

greater than 200 nucleotides, making them significantly longer than other non-coding variation (Perkel 2013). Although they represent a novel type of variation identified in recent years, they are believed to regulate gene expression and splicing, which can consequently confer susceptibility to disease.

Most importantly this study only examined the presence of a cis-eQTL between rs9979383 and *RUNX1*/*LOC10050640*; however the SNP may be conferring disease susceptibility by a different mechanism. By examination of studies in other diseases a likely candidate mechanism for this SNP is either a differential effect on gene splice variation or TF binding. Investigation of this effect requires application of further functional techniques which are described in the following section.

4.4 Future Work

This study has assessed the genetic overlap between 3 times of IA. To take this work forward it would be desirable to repeat this analysis in a much larger sample size to increase the power to detect smaller genetic effects. As RA had the largest sample size of the 3 diseases, it is crucially important that more samples are added to the JIA and PsA cohorts, which are significantly less well powered in the current analysis. This may be challenging as these represent rarer diseases for which sample cohorts have not been collected for as long as RA but already as the result of international collaborations, samples are becoming available. Since my analysis additional genotyping has been performed in additional JIA and PsA cases, for which data is currently available.

Although the strategy used in this study was successful in identifying genetic overlap, the application of more powerful statistical methods to assess overlap may reveal more regions in the future. These include cross phenotype meta-analysis and pan-meta GWA study techniques described in section 1.7.4.2. By combining the cases for the 3 diseases, an IA cohort of nearly 20,000 samples would be generated, which would provide extensive power to detect shared genetic effects. Indeed, another researcher in the department is already exploring this approach (personal communication). Analysis will be performed using a

simple case-control model to identify which regions remain significant and therefore represent the most likely to be true overlapping regions. Limitations include the fact that as the RA cases form the largest numbers, the known RA associated regions may drive associations observed. Furthermore, different SNP associations in the different diseases may be acting on the same gene but this will not be detected using a simple case-control approach. Nonetheless, the approach does provide the greatest power to detect associations that may not reach statistical significance in individual data sets but are associated with IA in general if the same SNPs are responsible across diseases.

As a number of different RA and JIA subtypes exist, it would be ideal to perform further subtype specific analysis. It has been shown previously that some associations are only observed in specific disease subtypes; for example the *AFF3* and *CD28* regions are associated with ACPA+ but not ACPA- RA whilst the *IL23R* region is associated with juvenile psoriatic arthritis but no other JIA subtypes (Hinks et al. 2011;Viatte et al. 2012) Therefore, future analysis would include a more detailed exploration of which JIA subtypes are genetically more similar to PsA or to ACPA positive / negative RA.

The current analysis of overlap identified a number of regions with suggestive evidence for association with more than one type of IA. However, only the *RUNX1* region was investigated further in my study. If further convincing evidence of true association is found for the other regions of overlap identified, it will be important that further studies, to functionally characterise the effect of the associated variants, are performed so that the fundamental pathways underpinning IA are identified. Two particularly interesting regions which should be prioritized for follow up are the *IL2RA* and *IL2RB* regions. These regions appear to exhibit different associations in RA and JIA and, therefore, might provide more information about the mechanism which makes the diseases two separate clinical entities. Firstly, it would be essential to use fine mapping, sequencing and haplotype analysis to identify the true causal variant in the regions. As association with both regions has already been replicated in several independent cohorts for both RA and JIA, they represent regions for which functional analysis could be

undertaken immediately. By contrast, several of the overlap regions identified represent novel associations (e.g. *RUNX1*, *MTMR3*, and *EOMES*) and, therefore, the association should be replicated in independent disease cohorts before any functional work is undertaken.

The aim of the Immunochip replication was to provide further evidence that 9 of the 50 overlapping regions identified were associated with RA using an independent disease cohort. If this study was to be enhanced it would be essential that it was performed in a larger sample size, as the current study was underpowered to detect moderate genetic effects. It is also necessary to replicate these associations in JIA and PsA. As described previously, this may be challenging due to the lower prevalence of disease and limited sample cohorts which are available. Furthermore, as only the index SNP from each disease was selected for replication, it did not provide sufficient representation of the genetic variation in any of the regions. As the SNP tested is unlikely to represent the causal variant, because of the low prior probability that a causal SNP would be included on an array, associations may be observed because of the LD with a causal SNP. In the lower powered replication cohort, if the SNP was not strongly correlated with the causal SNP but in modest LD, then the chances of replication are reduced. Therefore in future it will be desirable to genotype several tag SNPs in each region, to maximize the chance of detecting the causal variant. Genotyping in an independent RA cohort did reveal replicated association of variants in 2 overlapping regions at a $p<0.05$ (*CTLA4* and *MTMR3*) but, in order to explore further, the results will be combined with the Immunochip data in a meta-analysis, as this will provide results from the greatest number of samples and therefore will maximise the power.

One region which did provide further convincing evidence of association was the *RUNX1* region, which was associated at $p=0.02$ in an independent RA cohort. It is essential that the association is now tested in independent JIA and PsA cohort to confirm that the region truly represents a pan-IA locus. Given that the association observed in the RA replication cohort remained with the same SNP associated in the original RA Immunochip cohort, rs9979383, it may be possible to just test that

variant in the independent JIA and PsA samples, rather than performing fine mapping as was done for RA. If this is the case, this data could then be combined with the Immunochip data to provide an even greater powered analysis to estimate the true effect size.

Although at 75% of common (MAF<0.05) variation, the tag SNP coverage of the region in the *RUNX1* fine mapping was significantly higher than that in the Immunochip analysis, it is essential that complete coverage of the region is achieved in future. This means that all variants within the region will be captured and therefore it is more likely that the SNP identified either tags or is the causal variant within that region. Although 75% of variation was captured using the Sequenom MassARRAY SNP genotyping of 51 SNPs, it was calculated that to capture 100% of variation, 202 additional SNPs would have to be genotyped, just to account for common variation in the region, without taking into account any low frequency variation (MAF<0.05) which may be present in the region and therefore contributing to disease susceptibility. As many arrays such as the Immunochip, do not appear to capture this type of variation particularly well, it may be advantageous to perform imputation or sequence the DNA instead. DNA sequencing involves the determination of the complete sequence of DNA and can be performed at a regional or genome wide level with several next generation sequencing platforms now available. Both approaches have been used successfully in other immune mediated diseases to determine differences in variation between individuals, which may be indicators of disease susceptibility (Nejentsev et al. 2009;Rivas et al. 2011) . The technique can also be used to increase the likelihood of identifying a causal variant as a complete analysis of complete variation within a region can be performed. This has been shown to be effective in a recent study targeting the *IKBKE* and *IFIH1* regions which had previously been associated with SLE. In this case, DNA resequencing was utilised to successfully identify the most likely causal variants within these regions and help formation of a hypothesis to explain the regions' functional contribution to disease susceptibility (Wang et al. 2013a) . As the variants identified in that study represented low frequency variants, they may not have been identified by GWA studies alone, highlighting the importance of resequencing to identifying novel associations.

Additionally DNA resequencing is used in a number of experiments as a means of determining the sequence of end products such as in ChIP experiments where it is used to determine the sequence bound to the TFs under investigation. This has been successful in identifying the genes which are regulated by *IRF5* and *STAT4* in SLE (Wang et al. 2013a). DNA sequencing of the *RUNX1* region in RA, JIA and PsA cases plus healthy controls would allow an analysis comparing the sequences to be performed, which could be used to determine the most likely causal variant in the region. This variant can then be prioritised for further functional studies. Although DNA resequencing is a great resource for identifying causal variants, identifying low frequency variants and sequencing ChIP end products, it does have its disadvantages. For sequencing to be performed, a large volume of good quality DNA is required, which may not be possible in many cases. Furthermore, in order to be effective a large number of samples are required to be sequenced. As this technique is both very low throughput and expensive this can mean that experiments take up an extensive volume of time and cost, though arguably the data quantity and quality obtained makes it a worthwhile investment.

As the most likely causal variant in the *RUNX1* region identified to date, rs9979383 was selected as a candidate for functional studies. When assessed bioinformatically the presence of the H3K4Me1 histone methylation mark combined with the position of the SNP meant that this SNP was hypothesized to lie in a *RUNX1* regulatory region. Therefore a series of eQTL studies were designed to assess the effect of rs9979383 on gene expression in the region.

In the *RUNX1* whole blood eQTL, analysis was performed correlating rs9979383 genotype with *RUNX1* gene expression in 70 healthy controls using a Taqman gene expression assay designed to capture the majority of the known *RUNX1* splice transcripts. In the future this experiment should be repeated in a larger number of samples, as the eQTL effect may not have been identified as a consequence of limited power. Furthermore as rs9979383 was found to be associated with RA, JIA and PsA, the study should be extended to include DNA and RNA samples from cases with each of these 3 diseases. As explained previously due to low prevalence of JIA and PsA combined with ethical implications of taking samples from children, many studies have not collected the numbers of RNA samples early stage of studies

required for a well-powered study.

The study design itself involved the genotyping of rs9979383 and a gene expression assay to capture 15 of the 19 known splice variants for *RUNX1*. As only rs9979383 was genotyped this experiment did not take into account additional variation in the region. Furthermore although the gene expression assay was designed to capture as many *RUNX1* splice variants as possible for a single assay, not all transcript variation was captured. The nature of Taqman gene expression assays also means that the technology cannot discriminate between splice variants; therefore if the SNP is regulating one of the less common transcripts, it could be masked by the presence of the common transcripts. To explore this in the future, the experiment would have to be repeated either using multiple gene expression assays specifically designed for the 19 variants, using a whole transcriptome array as used in the current study or by RNA resequencing. Of these, RNA resequencing is the only one likely to capture all transcripts in a region.

RNA resequencing is similar to DNA sequencing, except it is used to sequence mRNA which is produced when genes are transcribed in the cell. Again it can be used to explore specific genomic regions or at a whole transcriptome level (Costa et al. 2013) . Unlike microarray technology, RNA resequencing does not require transcripts to be mapped to a genome build. This means it can be used to identify novel genes, splice variants and can be used to detect allele specific expression. Furthermore the technique can be used to detect non-coding RNAs and micro-RNA, which are thought to contribute to the development of many diseases including IA (Ceribelli et al. 2011a;Ceribelli et al. 2011b). Overall RNA resequencing is believed to be subject to much less background noise than microarrays, therefore potentially can produce more accurate findings (Wang et al. 2009).

RNA resequencing could be used to follow up this study as it would allow all splice transcripts to be detected and different splice variants to be discriminated from each other. The technique could then be used to determine which splice transcripts are produced when different alleles / genotypes of rs9979383 are present and therefore if this SNP regulates expression in the *RUNX1* region. Currently a study is underway within the department to perform RNA resequencing of the *RUNX1*

region in RA cases with differential genotype at rs9979383. These results may provide definitive confirmation of whether rs9979383 does regulate expression of RUNX1 through differential gene splicing or not. The data could also be combined with DNA sequencing data in the region to provide an insight into complete DNA and transcript variation in the region in a small group of RA cases.

Although RNA resequencing is likely to be the most utilised method for investigating gene expression in future, the technology is in its infancy and can therefore be particularly low throughput whilst being very costly. Furthermore it requires very high quality high concentration total RNA samples, which for many studies might be challenging to obtain.

As with the *RUNX1* whole blood eQTL study, the *RUNX1* eQTL study in T lymphocytes could be followed up with a number of strategies to investigate how rs9979383 contributes to disease susceptibility. As this study was only performed in 23 healthy volunteers and therefore was limited by a lack of genotype distribution at rs9979383, it is desirable to perform this study in a larger number of samples to achieve sufficient genotype distribution. Additionally as any potential eQTL may only be present in IA disease cases, it is important that this study is also repeated in RA, JIA and PsA cases. This may be challenging as for separation strategies to be employed, fresh blood samples are required. As IA patients only visit a clinician a couple of times a year, this limits the number of samples which can be taken and may restrict any potential future studies. Another factor which should be considered is the activation status. In a recent paper monocytes treated with IFN-γ had a very different eQTL profile compared to unstimulated cells (Fairfax et al. 2014). This means that the lack of eQTL in this study could potentially be the result of the resting state of the cell. Further investigation is required, by repeating the process in this study in cells which have been treated with inflammatory mediators such as IFN-γ or TNF-α.

At the time this study was initiated, work by Raychaudhuri et al. exploring the overlap of RA associations with epigenetic marks in different cell subsets had indicated that CD4+ EM T cells were likely to be the key cell type responsible for RA (Trynka et al. 2013). A similar approach in PsA by John Bowes had identified CD8+ T cells as the key cell type for PsA (Bowes et al, personal communication).

Hence, these two immune cell types were prioritised for cell-specific eQTL analysis. Since then evidence suggests that *RUNX1* may be regulating chondrogenesis in OA at a joint specific level. This may also be the case for IA types and therefore it would be ideal to repeat the eQTL study in chondrocytes derived from joints of IA patients. However, there are very few opportunities to sample joint tissue from patients and it is ethically not acceptable to sample healthy volunteers. Cartilage can be harvested from patients with IA at the time of joint replacement but that tissue may not be representative of the situation in early arthritis when inflammation is more active. If these samples cannot be obtained, chondrocytes could also be derived from primary cell blood monocytes but this may potentially produce a gene expression profile different to those seen in joint chondrocytes.

In order to collect enough peripheral blood for multiple cell separations, PBMC cryopreservation techniques were utilized. Although it was not known whether this has a major effect on the gene expression profile of the separated cells, it is a widely used technique across other studies, although not always mentioned in publications. To assess the effect of cryopreservation, the gene expression profiles of the CD8+ and CD4+ samples were plotted on an MDS plot to identify any aberrant differences in expression between samples (). Although the profiles of the CD8+ and CD4+samples looked significantly different, all samples within the same group clustered very strongly together, indicating that, although the samples were cryopreserved for different times, it did not have any large effects on the gene expression profiles. Although this is a strong indicator that cryopreservation is suitable for use in future studies within the department, it is desirable that a comprehensive study be performed comparing the gene expression profiles of cryopreserved and non-cryopreserved cells. This is described in more detail in section 4.4.

The methods used in this study to determine genotype and gene expression could also be improved. As described previously only genotype data from rs9979383 was analysed, therefore all other SNPs in the region were not taken into account. This is important as they could potentially be an eQTL but have not been picked up

in this study. Since completion of the whole blood eQTL study, the samples have been genotyped on the Illumina core exome array which includes a number of tag SNPs in the *RUNX1* region. In that study, a genotyping array approach was chosen due to a number of factors. Genotyping arrays are fairly high throughput and have a defined analysis pipeline for interpretation of data. This allowed generation of a large volume of high throughput genetic data from 23 samples at low cost across a short period of time. Further work will involve correlating these SNPs with expression in the region, to determine the presence of any eQTLs. However, unless the eQTL SNP was more strongly associated with IA, it would not explain how the associated variant for IA acts.

Although a "whole transcriptome," array was used in this study, it did not contain probes which captured all 19 splice variants of *RUNX1* described previously. To address this in future studies, a gene expression assay would have to be designed to capture all splice transcripts or RNA resequencing would have to be performed. Although this has not been performed using RNA from individual cell types in our department, the development of single cell gene expression techniques such as the Nanostring technology will allow this to be performed using a small amount of sample, allowing for discrimination of different splice variants. The approach would be ideal for following up the current study due to availability of leftover RNA, which has been extracted but is available for further testing.

Although an eQTL was not found in the *RUNX1* region in whole blood, CD8+ or CD4+ T lymphocytes this may mean that rs9979383 confers disease susceptibility in the *RUNX1* region by a completely different mechanism. One possibility is that the region in which rs9979383 is located represents a TFBS or DNA binding site. The presence of the SNP could change the sequence in such a way that binding cannot occur. This mechanism can be investigated using 3C or ChIP analysis as described in section 1.6.3 and would allow determination of which sequence or transcription factor binding is altered in the presence of the different SNP alleles. The analysis could be performed in a number of cell lines such as immortal T lymphocyte or chondrocyte cells with known genotype, making it much easier to obtain genetic material than the primary cell strategy used in this study.

Another potential role of the rs9979383 is the epigenetic regulation within the *RUNX1* region. Epigenetic regulation can result in changes to how transcription factors bind or how genes are transcribed. This is particularly interesting with regards to DNA methylation, as DNA within the *RUNX1* region has been shown to be differentially methylated in OA and non-OA joint tissue (Jeffries et al, personal communication). Further investigation of this potential mechanism of action would involve assessing methylation in DNA samples with differential genotypes at rs9979383, and examining the changes in methylation levels across the region. The experiment should ideally be performed in a cell specific study, as methylation has been shown to be differential between different cell types (Reinius et al. 2012). This could be achieved on a small scale using Pyrosequencing or as part of a larger study using an Illumina whole genome methylation array. The arrays have probes to detect over 485,000 methylation sites across the genome and could ideally be combined with gene expression data to give a comprehensive insight into gene regulation in the *RUNX1* region.

4.5 Conclusion

This study has provided a comprehensive profile of the genetic overlap between different types of IA. Using data generated using the Immunochip array, 50 regions were identified as being associated with more than 1 type of IA at $p<1x10^{-3}$. In many regions these genetic effects appeared to be shared between diseases, with either an identical or a highly correlated SNP associated with each disease. In contrast, in a number of regions the SNP appeared to completely different, with either low or no LD detected between the SNPs associated with each disease. Many interesting findings were identified in this study; with the IL2 pathway genes, *IL2RA* and *IL2RB* showing very different effects between diseases whilst in the *RUNX1* region an identical SNP was associated across RA, JIA and PsA collectively.

Two novel disease associations were replicated in an independent RA cohort, confirming association with the *CTLA4* and *MTMR3* genes. These associations will require further follow up in JIA and PsA cohorts, to be confirmed as true overlapping associations.

Association with a third novel locus, the *RUNX*1 region, was also replicated in an independent RA cohort. Fine-mapping revealed that the original SNP remained the most associated in the replication study, indicating that the SNP, or one in high LD with it, is likely to be the causal variant.

The SNP, rs9979383 maps ~280kb upstream of the *RUNX1* gene and appears to have the H3K4Me1 histone methylation, which is usually indicative of enhancers. However, initial studies to investigate whether it is involved in gene regulation showed no significant eQTLs in whole blood, CD8+ or CD4+ lymphocytes, in unstimulated samples from healthy control volunteers.

Overall this study provides an insight into the genetic similarities and differences between RA, JIA and PsA, which is crucial in learning more about the susceptibility and pathogenesis of these diseases. It also shows how a pipeline can be generated to follow up genetic loci associated with disease for subsequent replication of association and characterisation of SNP function. Although no eQTLs were detected in the *RUNX1* region in this study, these findings have guided future investigations in this region, which will be essential in identifying how this region is contributing to disease across these 3 types of IA.

5. Appendix

5.1 Tempus spin RNA isolation kit Tempus Spin RNA Isolation Kit

The Tempus Spin RNA Isolation Kit (Applied Biosystems) is for extraction using Tempus RNA Blood Tubes. These tubes contain stabilizing reagent (Applied Biosystems) which is activated upon shaking with whole blood for 10 seconds. This reagent induces immediate cell lysis, inactivation of RNases and selective precipitation of RNA whilst genomic DNA and proteins remain in solution. Collected whole blood can be stored for 5 days at 27°C, 4 days at 4°C or indefinitely at -20/-80°C. The RNA spin isolation kit is for the purification of total RNA from 3 ml whole blood. The protocol involves two stages: processing of stabilized blood and purification of RNA. All stages are performed where possible on ice.

5.1.1 Processing of stabilized blood

The sample was transferred to a 50ml conical tube and 3ml phosphate buffered saline (PBS) was added. The sample was mixed by vortexing for 30 seconds and centrifuged 3000xg at 4°C for 30 minutes. The supernatant was discarded and excess supernatant removed by inverting on absorbent paper for 2 minutes. 400µl RNA Purification Suspension Solution was added and vortexed briefly to mix contents.

5.1.2 Purification of RNA

The glass fibre purification filter cartridge (Ambion) was inserted into a waste collection tube (Ambion). Pre-treatment of the membrane was performed by adding 100µl of RNA Purification Wash Solution 1 (Applied Biosystems). The sample (approx. 400µl) was transferred to the filter and centrifuged for 16,000xg for 1 minute. The filter cartridge was removed and the supernatant discarded. The

filter was re-inserted into the waste tube, 500µl RNA Purification Wash Solution 2 (Applied Biosystems) added and centrifuged at 16,000xg for 1 minute.

A DNase treatment was performed by adding 100µl Absolute RNA Wash Solution (Applied Biosystems) and incubating for 15 minutes at room temperature. 500µl Wash Solution 2 (Applied Biosystems) was added, incubated for 5 minutes at room temperature and centrifuged at 16,000xg for 30 seconds. The filter cartridge was removed and the supernatant discarded. The filter was re-inserted into the waste tube, 500µl Wash Solution 2 added and centrifuged at 16,000xg for 30 seconds. The filter cartridge was removed and the supernatant discarded. The filter was re-inserted into the waste tube and centrifuged at 16,000xg for 30 seconds to dry the membrane.

The filter cartridge was transferred to a new collection tube. 100µl Nucleic Acid Purification Elution Solution (Applied Biosystems) was added, incubated at 70°C for 2 minutes and centrifuged at 16,000xg for 30 seconds. The collected RNA was re-added and centrifuged at 16,000xg for 2 minutes to ensure optimum elution of RNA. The filter cartridge was discarded and the top 90µl of elute transferred to a new collection tube. The samples were then transferred to ice for quality control analysis.

5.1.3 RNA QC

Quality control was performed on isolated RNA to assess both the quantity and the quality of RNA obtained using the two different protocols.

5.1.3.1 RNA quality control using Nanodrop N-1000

Quality of RNA obtained by both methods was analysed using a Nanodrop ND-1000 (Thermo Scientific). This spectrophotometer uses UV light absorbance ranging (approx. 200-350nm) to quantify nucleic acids present in a sample and determine the presence of any contaminants. As maximum RNA absorption takes

place at A280, absorption at this wavelength is analysed and converted to a quantifiable concentration of ng/µl using the Beer-Lambert Law.

Absorption values are also obtained at A230 and A260, which are used to generate 260/230 and 260/280 ratios. These ratios can be used to determine the purity of RNA obtained. Pure RNA has a 260/230 ratio of 2 and a 260/280 ratio of 2.1 but values ranging 1.8-2.1 are acceptable. Large deviations from expected values indicate sample contamination. Possible contaminants include DNA, protein and any reagents used during extraction such as ethanol or guanidine isothiocyanate.

After each extraction the yield and quality of RNA obtained was assessed using the Nanodrop ND-1000. Values for concentration, 260/280 and 260/230 ratios were recorded to ensure all samples values were within an acceptable range. This ensures that the RNA is suitable for use in downstream processes and is not affected by sample contamination.

6.0 References

2013. The Genotype-Tissue Expression (GTEx) project. *Nat.Genet.*, 45, (6) 580-585 available from: PM:23715323

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., & McVean, G.A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, (7422) 56-65 available from: PM:23128226

Adib, N., Silman, A., & Thomson, W. 2005. Outcome following onset of juvenile idiopathic inflammatory arthritis: I. frequency of different outcomes. *Rheumatology.(Oxford)*, 44, (8) 995-1001 available from: PM:15827045

Alarcon-Riquelme, M.E. 2003. A RUNX trio with a taste for autoimmunity. *Nat.Genet.*, 35, (4) 299-300 available from: PM:14647282

Albers, H.M., Kurreeman, F.A., Houwing-Duistermaat, J.J., Brinkman, D.M., Kamphuis, S.S., Girschick, H.J., Wouters, C., Van Rossum, M.A., Verduijn, W., Toes, R.E., Huizinga, T.W., Schilham, M.W., & ten, C.R. 2008. The TRAF1/C5 region is a risk factor for polyarthritis in juvenile idiopathic arthritis. *Ann.Rheum.Dis.*, 67, (11) 1578-1580 available from: PM:18593758

Aletaha, D., Neogi, T., Silman, A.J., Funovits, J., Felson, D.T., Bingham, C.O., III, Birnbaum, N.S., Burmester, G.R., Bykerk, V.P., Cohen, M.D., Combe, B., Costenbader, K.H., Dougados, M., Emery, P., Ferraccioli, G., Hazes, J.M., Hobbs, K., Huizinga, T.W., Kavanaugh, A., Kay, J., Kvien, T.K., Laing, T., Mease, P., Menard, H.A., Moreland, L.W., Naden, R.L., Pincus, T., Smolen, J.S., Stanislawska-Biernat, E., Symmons, D., Tak, P.P., Upchurch, K.S., Vencovsky, J., Wolfe, F., & Hawker, G. 2010. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann.Rheum.Dis.*, 69, (9) 1580-1588 available from: PM:20699241

Apel, M., Uebe, S., Bowes, J., Giardina, E., Korendowych, E., Juneblad, K., Pasutto, F., Ekici, A.B., McManus, R., Ho, P., Bruce, I.N., Ryan, A.W., Behrens, F., Bohm, B.,

Traupe, H., Lohmann, J., Gieger, C., Wichmann, H.E., Padyukov, L., Fitzgerald, O., Alenius, G.M., McHugh, N.J., Novelli, G., Burkhardt, H., Barton, A., Reis, A., & Huffmeier, U. 2013. Variants in RUNX3 contribute to susceptibility to psoriatic arthritis, exhibiting further common ground with ankylosing spondylitis. *Arthritis Rheum.*, 65, (5) 1224-1231 available from: PM:23401011

Arnett, F.C., Edworthy, S.M., Bloch, D.A., McShane, D.J., Fries, J.F., Cooper, N.S., Healey, L.A., Kaplan, S.R., Liang, M.H., Luthra, H.S., & . 1988. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.*, 31, (3) 315-324 available from: PM:3358796

Arts, E.E., Fransen, J., den Broeder, A.A., Popa, C.D., & van Riel, P.L. 2014. The effect of disease duration and disease activity on the risk of cardiovascular disease in rheumatoid arthritis patients. *Ann.Rheum.Dis.* available from: PM:24458537

Balding, D.J. 2006. A tutorial on statistical methods for population association studies. *Nat.Rev.Genet.*, 7, (10) 781-791 available from: PM:16983374

Baltimore, D. 2011. NF-kappaB is 25. *Nat.Immunol.*, 12, (8) 683-685 available from: PM:21772275

Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D.J., Stevens, H., Todd, J.A., Walker, N.M., & Rich, S.S. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat.Genet.*, 41, (6) 703-707 available from: PM:19430480

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., Bitton, A., Dassopoulos, T., Datta, L.W., Green, T., Griffiths, A.M., Kistner, E.O., Murtha, M.T., Regueiro, M.D., Rotter, J.I., Schumm, L.P., Steinhart, A.H., Targan, S.R., Xavier, R.J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van, G.A., Zelenika, D., Franchimont, D., Hugot, J.P., de, V.M., Vermeire, S., Louis, E., Cardon, L.R., Anderson, C.A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N.J., Onnie, C.M., Fisher, S.A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R.,

Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C.G., Parkes, M., Georges, M., & Daly, M.J. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat.Genet.*, 40, (8) 955-962 available from: PM:18587394

Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Gibbons, L., Plant, D., Wilson, A.G., Marinou, I., Morgan, A., Emery, P., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P., & Worthington, J. 2008a. Re-evaluation of putative rheumatoid arthritis susceptibility genes in the post-genome wide association study era and hypothesis of a key pathway underlying susceptibility. *Hum.Mol.Genet.*, 17, (15) 2274-2279 available from: PM:18434327

Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Plant, D., Gibbons, L.J., Wilson, A.G., Bax, D.E., Morgan, A.W., Emery, P., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P., & Worthington, J. 2008b. Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat.Genet.*, 40, (10) 1156-1159 available from: PM:18794857

Bayley, R., Kite, K.A., McGettrick, H.M., Smith, J.P., Kitas, G.D., Buckley, C.D., & Young, S.P. 2014. The autoimmune-associated genetic variant PTPN22 R620W enhances neutrophil activation and function in patients with rheumatoid arthritis and healthy individuals. *Ann.Rheum.Dis.* available from: PM:24665115

Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kemppinen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D., Goris, A., Oturai, A., Saarela, J., Fontaine, B., Hemmer, B., Martin, C., Zipp, F., D'Alfonso, S., Martinelli-Boneschi, F., Taylor, B., Harbo, H.F., Kockum, I., Hillert, J., Olsson, T., Ban, M., Oksenberg, J.R., Hintzen, R., Barcellos, L.F., Agliardi, C., Alfredsson, L., Alizadeh, M., Anderson, C., Andrews, R., Sondergaard, H.B., Baker, A., Band, G., Baranzini, S.E., Barizzone, N., Barrett, J., Bellenguez, C., Bergamaschi, L., Bernardinelli, L., Berthele, A., Biberacher, V., Binder, T.M., Blackburn, H., Bomfim, I.L., Brambilla, P., Broadley, S., Brochet, B., Brundin, L., Buck, D., Butzkueven, H., Caillier, S.J., Camu, W., Carpentier, W., Cavalla, P., Celius, E.G., Coman, I., Comi, G., Corrado, L., Cosemans, L., Cournu-Rebeix, I., Cree, B.A., Cusi, D., Damotte, V., Defer, G., Delgado, S.R., Deloukas, P., di, S.A., Dilthey, A.T., Donnelly, P., Dubois, B., Duddy, M., Edkins, S., Elovaara, I., Esposito, F., Evangelou,

N., Fiddes, B., Field, J., Franke, A., Freeman, C., Frohlich, I.Y., Galimberti, D., Gieger, C., Gourraud, P.A., Graetz, C., Graham, A., Grummel, V., Guaschino, C., Hadjixenofontos, A., Hakonarson, H., Halfpenny, C., Hall, G., Hall, P., Hamsten, A., Harley, J., Harrower, T., Hawkins, C., Hellenthal, G., Hillier, C., Hobart, J., Hoshi, M., Hunt, S.E., Jagodic, M., Jelcic, I., Jochim, A., Kendall, B., Kermode, A., Kilpatrick, T., Koivisto, K., Konidari, I., Korn, T., Kronsbein, H., Langford, C., Larsson, M., Lathrop, M., Lebrun-Frenay, C., Lechner-Scott, J., Lee, M.H., Leone, M.A., Leppa, V., Liberatore, G., Lie, B.A., Lill, C.M., Linden, M., Link, J., Luessi, F., Lycke, J., Macciardi, F., Mannisto, S., Manrique, C.P., Martin, R., Martinelli, V., Mason, D., Mazibrada, G., McCabe, C., Mero, I.L., Mescheriakova, J., Moutsianas, L., Myhr, K.M., Nagels, G., Nicholas, R., Nilsson, P., Piehl, F., Pirinen, M., Price, S.E., Quach, H., Reunanen, M., Robberecht, W., Robertson, N.P., Rodegher, M., Rog, D., Salvetti, M., Schnetz-Boutaud, N.C., Sellebjerg, F., Selter, R.C., Schaefer, C., Shaunak, S., Shen, L., Shields, S., Siffrin, V., Slee, M., Sorensen, P.S., Sorosina, M., Sospedra, M., Spurkland, A., Strange, A., Sundqvist, E., Thijs, V., Thorpe, J., Ticca, A., Tienari, P., van, D.C., Visser, E.M., Vucic, S., Westerlind, H., Wiley, J.S., Wilkins, A., Wilson, J.F., Winkelmann, J., Zajicek, J., Zindler, E., Haines, J.L., Pericak-Vance, M.A., Ivinson, A.J., Stewart, G., Hafler, D., Hauser, S.L., Compston, A., McVean, G., De, J.P., Sawcer, S.J., & McCauley, J.L. 2013. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat.Genet.*, 45, (11) 1353-1360 available from: PM:24076602

Behrens, E.M., Finkel, T.H., Bradfield, J.P., Kim, C.E., Linton, L., Casalunovo, T., Frackelton, E.C., Santa, E., Otieno, F.G., Glessner, J.T., Chiavacci, R.M., Grant, S.F., & Hakonarson, H. 2008. Association of the TRAF1-C5 locus on chromosome 9 with juvenile idiopathic arthritis. *Arthritis Rheum.*, 58, (7) 2206-2207 available from: PM:18576341

Berner, B., Akca, D., Jung, T., Muller, G.A., & Reuss-Borst, M.A. 2000. Analysis of Th1 and Th2 cytokines expressing CD4+ and CD8+ T cells in rheumatoid arthritis by flow cytometry. *J.Rheumatol.*, 27, (5) 1128-1135 available from: PM:10813277

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., & Snyder, M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, (7414) 57-74 available from: PM:22955616

Blanco, F.J. & Ruiz-Romero, C. 2013. New targets for disease modifying osteoarthritis drugs: chondrogenesis and Runx1. *Ann.Rheum.Dis.*, 72, (5) 631-634 available from: PM:23444194

Bodmer, W. & Bonilla, C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat.Genet.*, 40, (6) 695-701 available from: PM:18509313

Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M., MacMurray, J., Meloni, G.F., Lucarelli, P., Pellecchia, M., Eisenbarth, G.S., Comings, D., & Mustelin, T. 2004. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat.Genet.*, 36, (4) 337-338 available from: PM:15004560

Brennan, F.M., Chantry, D., Jackson, A.M., Maini, R.N., & Feldmann, M. 1989. Cytokine production in culture by cells isolated from the synovial membrane. *J.Autoimmun.*, 2 Suppl, 177-186 available from: PM:2505790

Buckley, C.D., Filer, A., Haworth, O., Parsonage, G., & Salmon, M. 2004. Defining a role for fibroblasts in the persistence of chronic inflammatory joint disease. *Ann.Rheum.Dis.*, 63 Suppl 2, ii92-ii95 available from: PM:15479882

Buckley, C.D., Pilling, D., Lord, J.M., Akbar, A.N., Scheel-Toellner, D., & Salmon, M. 2001. Fibroblasts regulate the switch from acute resolving to chronic persistent inflammation. *Trends Immunol.*, 22, (4) 199-204 available from: PM:11274925

Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., & Hirschhorn, J.N. 2005. Demonstrating stratification in a European American population. *Nat.Genet.*, 37, (8) 868-872 available from: PM:16041375

Campbell, R.C., Batley, M., Hammond, A., Ibrahim, F., Kingsley, G., & Scott, D.L. 2012. The impact of disease activity, pain, disability and treatments on fatigue in established rheumatoid arthritis. *Clin.Rheumatol.*, 31, (4) 717-722 available from: PM:22124789

Cantrell, D.A. & Smith, K.A. 1984. The interleukin-2 T-cell system: a new cell growth model. *Science*, 224, (4655) 1312-1316 available from: PM:6427923

Carson, D.A., Bayer, A.S., Eisenberg, R.A., Lawrance, S., & Theofilopoulos, A. 1978. IgG rheumatoid factor in subacute bacterial endocarditis: relationship to IgM rheumatoid factor and circulating immune complexes. *Clin.Exp.Immunol.*, 31, (1) 100-103 available from: PM:639341

Carson, D.A., Chen, P.P., Kipps, T.J., Radoux, V., Jirik, F., Goldfien, R.D., Fox, R.I., Silverman, G.J., & Fong, S. 1987. Molecular basis for the cross-reactive idiotypes on human anti-IgG autoantibodies (rheumatoid factors). *Ciba Found.Symp.*, 129, 123-134 available from: PM:3315499

Ceribelli, A., Nahid, M.A., Satoh, M., & Chan, E.K. 2011a. MicroRNAs in rheumatoid arthritis. *FEBS Lett.*, 585, (23) 3667-3674 available from: PM:21600203

Ceribelli, A., Yao, B., Dominguez-Gutierrez, P.R., Nahid, M.A., Satoh, M., & Chan, E.K. 2011b. MicroRNAs in systemic rheumatic diseases. *Arthritis Res.Ther.*, 13, (4) 229 available from: PM:21787439

Cespedes-Cruz, A., Gutierrez-Suarez, R., Pistorio, A., Ravelli, A., Loy, A., Murray, K.J., Gerloni, V., Wulffraat, N., Oliveira, S., Walsh, J., Penades, I.C., Alpigiani, M.G., Lahdenne, P., Saad-Magalhaes, C., Cortis, E., Lepore, L., Kimura, Y., Wouters, C., Martini, A., & Ruperto, N. 2008. Methotrexate improves the health-related quality of life of children with juvenile idiopathic arthritis. *Ann.Rheum.Dis.*, 67, (3) 309-314 available from: PM:17875547

Chandran, V., Schentag, C.T., Brockbank, J.E., Pellett, F.J., Shanmugarajah, S., Toloza, S.M., Rahman, P., & Gladman, D.D. 2009. Familial aggregation of psoriatic arthritis. *Ann.Rheum.Dis.*, 68, (5) 664-667 available from: PM:18524791

Chen, J., Bruns, A.H., Donnelly, H.K., & Wunderink, R.G. 2010. Comparative in vitro stimulation with lipopolysaccharide to study TNFalpha gene expression in fresh whole blood, fresh and frozen peripheral blood mononuclear cells. *J.Immunol.Methods*, 357, (1-2) 33-37 available from: PM:20307542

Ciechomska, M., Wilson, C.L., Floudas, A., Hui, W., Rowan, A.D., van, E.W., Robinson, J.H., & Knight, A.M. 2014. Antigen-specific B lymphocytes acquire proteoglycan aggrecan from cartilage extracellular matrix resulting in antigen presentation and CD4+ T-cell activation. *Immunology*, 141, (1) 70-78 available from: PM:24032649

Cinek, O., Hradsky, O., Ahmedov, G., Slavcev, A., Kolouskova, S., Kulich, M., & Sumnik, Z. 2007. No independent role of the -1123 G>C and+2740 A>G variants in the association of PTPN22 with type 1 diabetes and juvenile idiopathic arthritis in two Caucasian populations. *Diabetes Res.Clin.Pract.*, 76, (2) 297-303 available from: PM:17000021

Coenen, M.J., Trynka, G., Heskamp, S., Franke, B., van Diemen, C.C., Smolonska, J., van, L.M., Brouwer, E., Boezen, M.H., Postma, D.S., Platteel, M., Zanen, P., Lammers, J.W., Groen, H.J., Mali, W.P., Mulder, C.J., Tack, G.J., Verbeek, W.H., Wolters, V.M., Houwen, R.H., Mearin, M.L., van Heel, D.A., Radstake, T.R., van Riel, P.L., Wijmenga, C., Barrera, P., & Zhernakova, A. 2009. Common and different genetic background for rheumatoid arthritis and coeliac disease. *Hum.Mol.Genet.*, 18, (21) 4195-4203 available from: PM:19648290

Cohen, S.B., Dore, R.K., Lane, N.E., Ory, P.A., Peterfy, C.G., Sharp, J.T., van der, H.D., Zhou, L., Tsuji, W., & Newmark, R. 2008. Denosumab treatment effects on structural damage, bone mineral density, and bone turnover in rheumatoid arthritis: a twelve-month, multicenter, randomized, double-blind, placebo-controlled, phase II clinical trial. *Arthritis Rheum.*, 58, (5) 1299-1309 available from: PM:18438830

Conaway, H.H., Pirhayati, A., Persson, E., Pettersson, U., Svensson, O., Lindholm, C., Henning, P., Tuckermann, J., & Lerner, U.H. 2011. Retinoids stimulate periosteal bone resorption by enhancing the protein RANKL, a response inhibited by monomeric glucocorticoid receptor. *J.Biol.Chem.*, 286, (36) 31425-31436 available from: PM:21715325

Cooper, J.D., Simmonds, M.J., Walker, N.M., Burren, O., Brand, O.J., Guo, H., Wallace, C., Stevens, H., Coleman, G., Franklyn, J.A., Todd, J.A., & Gough, S.C. 2012. Seven

newly identified loci for autoimmune thyroid disease. *Hum.Mol.Genet.*, 21, (23) 5202-5208 available from: PM:22922229

Cope, A.P. 2008. T cells in rheumatoid arthritis. *Arthritis Res.Ther.*, 10 Suppl 1, S1 available from: PM:19007421

Cordova, K.N., Willis, V.C., Haskins, K., & Holers, V.M. 2013. A citrullinated fibrinogen-specific T cell line enhances autoimmune arthritis in a mouse model of rheumatoid arthritis. *J.Immunol.*, 190, (4) 1457-1465 available from: PM:23319740

Cortes, A. & Brown, M.A. 2011. Promise and pitfalls of the Immunochip. *Arthritis Res.Ther.*, 13, (1) 101 available from: PM:21345260

Costa, V., Aprile, M., Esposito, R., & Ciccodicola, A. 2013. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur.J.Hum.Genet.*, 21, (2) 134-142 available from: PM:22739340

Cotsapas, C. & Hafler, D.A. 2013. Immune-mediated disease genetics: the shared basis of pathogenesis. *Trends Immunol.*, 34, (1) 22-26 available from: PM:23031829

Cotsapas, C., Voight, B.F., Rossin, E., Lage, K., Neale, B.M., Wallace, C., Abecasis, G.R., Barrett, J.C., Behrens, T., Cho, J., De Jager, P.L., Elder, J.T., Graham, R.R., Gregersen, P., Klareskog, L., Siminovitch, K.A., van Heel, D.A., Wijmenga, C., Worthington, J., Todd, J.A., Hafler, D.A., Rich, S.S., & Daly, M.J. 2011. Pervasive sharing of genetic effects in autoimmune disease. *PLoS.Genet.*, 7, (8) e1002254 available from: PM:21852963

Courtenay, J.S., Dallman, M.J., Dayan, A.D., Martin, A., & Mosedale, B. 1980. Immunisation against heterologous type II collagen induces arthritis in mice. *Nature*, 283, (5748) 666-668 available from: PM:6153460

Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., Holmes, C., Marchini, J.L., Stirrups, K., Tobin, M.D., Wain, L.V., Yau, C., Aerts, J., Ahmad, T., Andrews, T.D., Arbury, H., Attwood, A., Auton, A., Ball, S.G., Balmforth, A.J., Barrett, J.C., Barroso, I.,

Barton, A., Bennett, A.J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O.J., Braund, P.S., Bredin, F., Breen, G., Brown, M.J., Bruce, I.N., Bull, J., Burren, O.S., Burton, J., Byrnes, J., Caesar, S., Clee, C.M., Coffey, A.J., Connell, J.M., Cooper, J.D., Dominiczak, A.F., Downes, K., Drummond, H.E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D.M., Evans, G., Eyre, S., Farmer, A., Ferrier, I.N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J.A., Freathy, R.M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C.J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G.A., Hocking, L., Howard, E., Howard, P., Howson, J.M., Hughes, D., Hunt, S., Isaacs, J.D., Jain, M., Jewell, D.P., Johnson, T., Jolley, J.D., Jones, I.R., Jones, L.A., Kirov, G., Langford, C.F., Lango-Allen, H., Lathrop, G.M., Lee, J., Lee, K.L., Lees, C., Lewis, K., Lindgren, C.M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin, P., Massey, D.C., McArdle, W.L., McGuffin, P., McLay, K.E., Mentzer, A., Mimmack, M.L., Morgan, A.E., Morris, A.P., Mowat, C., Myers, S., Newman, W., Nimmo, E.R., O'Donovan, M.C., Onipinla, A., Onyiah, I., Ovington, N.R., Owen, M.J., Palin, K., Parnell, K., Pernet, D., Perry, J.R., Phillips, A., Pinto, D., Prescott, N.J., Prokopenko, I., Quail, M.A., Rafelt, S., Rayner, N.W., Redon, R., Reid, D.M., Renwick, Ring, S.M., Robertson, N., Russell, E., St, C.D., Sambrook, J.G., Sanderson, J.D., Schuilenburg, H., Scott, C.E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B.M., Simmonds, M.J., Smyth, D.J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H.E., Stone, M.A., Su, Z., Symmons, D.P., Thompson, J.R., Thomson, W., Travers, M.E., Turnbull, C., Valsesia, A., Walker, M., Walker, N.M., Wallace, C., Warren-Perry, M., Watkins, N.A., Webster, J., Weedon, M.N., Wilson, A.G., Woodburn, M., Wordsworth, B.P., Young, A.H., Zeggini, E., Carter, N.P., Frayling, T.M., Lee, C., McVean, G., Munroe, P.B., Palotie, A., Sawcer, S.J., Scherer, S.W., Strachan, D.P., Tyler-Smith, C., Brown, M.A., Burton, P.R., Caulfield, M.J., Compston, A., Farrall, M., Gough, S.C., Hall, A.S., Hattersley, A.T., Hill, A.V., Mathew, C.G., Pembrey, M., Satsangi, J., Stratton, M.R., Worthington, J., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W., Parkes, M., Rahman, N., Todd, J.A., Samani, N.J., & Donnelly, P. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464, (7289) 713-720 available from: PM:20360734

Crawford, D.C. & Nickerson, D.A. 2005. Definition and clinical importance of haplotypes. *Annu.Rev.Med.*, 56, 303-320 available from: PM:15660514

Cui, J., Taylor, K.E., Lee, Y.C., Kallberg, H., Weinblatt, M.E., Coblyn, J.S., Klareskog, L., Criswell, L.A., Gregersen, P.K., Shadick, N.A., Plenge, R.M., & Karlson, E.W. 2014. The influence of polygenic risk scores on heritability of anti-CCP level in RA. *Genes Immun.*, 15, (2) 107-114 available from: PM:24385024

Cutolo, M. & Straub, R.H. 2008. Circadian rhythms in arthritis: hormonal effects on the immune/inflammatory reaction. *Autoimmun.Rev.*, 7, (3) 223-228 available from: PM:18190882

Daridon, C., Burmester, G.R., & Dorner, T. 2009. Anticytokine therapy impacting on B cells in autoimmune diseases. *Curr.Opin.Rheumatol.*, 21, (3) 205-210 available from: PM:19346949

Dayer, J.M. 2003. The pivotal role of interleukin-1 in the clinical manifestations of rheumatoid arthritis. *Rheumatology.(Oxford)*, 42 Suppl 2, ii3-10 available from: PM:12817089

de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., & Altshuler, D. 2005. Efficiency and power in genetic association studies. *Nat.Genet.*, 37, (11) 1217-1223 available from: PM:16244653

De, B.F., Massa, M., Pignatti, P., Albani, S., Novick, D., & Martini, A. 1994. Serum soluble interleukin 6 (IL-6) receptor and IL-6/soluble IL-6 receptor complex in systemic juvenile rheumatoid arthritis. *J.Clin.Invest*, 93, (5) 2114-2119 available from: PM:8182142

Deane, K.D. & El-Gabalawy, H. 2014. Pathogenesis and prevention of rheumatic disease: focus on preclinical RA and SLE. *Nat.Rev.Rheumatol.*, 10, (4) 212-228 available from: PM:24514912

Deighton, C.M., Walker, D.J., Griffiths, I.D., & Roberts, D.F. 1989. The contribution of HLA to rheumatoid arthritis. *Clin.Genet.*, 36, (3) 178-182 available from: PM:2676268

Diogo, D., Kurreeman, F., Stahl, E.A., Liao, K.P., Gupta, N., Greenberg, J.D., Rivas, M.A., Hickey, B., Flannick, J., Thomson, B., Guiducci, C., Ripke, S., Adzhubey, I., Barton, A.,

Kremer, J.M., Alfredsson, L., Sunyaev, S., Martin, J., Zhernakova, A., Bowes, J., Eyre, S., Siminovitch, K.A., Gregersen, P.K., Worthington, J., Klareskog, L., Padyukov, L., Raychaudhuri, S., & Plenge, R.M. 2013. Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. *Am.J.Hum.Genet.*, 92, (1) 15-27 available from: PM:23261300

Doorenspleet, M.E., Klarenbeek, P.L., de Hair, M.J., van Schaik, B.D., Esveldt, R.E., van Kampen, A.H., Gerlag, D.M., Musters, A., Baas, F., Tak, P.P., & de, V.N. 2014. Rheumatoid arthritis synovial tissue harbours dominant B-cell and plasma-cell clones associated with autoreactivity. *Ann.Rheum.Dis.*, 73, (4) 756-762 available from: PM:23606709

Dougados, M., van der, L.S., Juhlin, R., Huitfeldt, B., Amor, B., Calin, A., Cats, A., Dijkmans, B., Olivieri, I., Pasero, G., & . 1991. The European Spondylarthropathy Study Group preliminary criteria for the classification of spondylarthropathy. *Arthritis Rheum.*, 34, (10) 1218-1227 available from: PM:1930310

Dunning, M.J., Smith, M.L., Ritchie, M.E., & Tavare, S. 2007. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics.*, 23, (16) 2183-2184 available from: PM:17586828

Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., & McVean, G.A. 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467, (7319) 1061-1073 available from: PM:20981092

Edwards, J.C., Cambridge, G., & Abrahams, V.M. 1999. Do self-perpetuating B lymphocytes drive human autoimmune disease? *Immunology*, 97, (2) 188-196 available from: PM:10447731

Ehrenstein, M.R., Evans, J.G., Singh, A., Moore, S., Warnes, G., Isenberg, D.A., & Mauri, C. 2004. Compromised function of regulatory T cells in rheumatoid arthritis and reversal by anti-TNFalpha therapy. *J.Exp.Med.*, 200, (3) 277-285 available from: PM:15280421

Emery, P., Deodhar, A., Rigby, W.F., Isaacs, J.D., Combe, B., Racewicz, A.J., Latinis, K., Abud-Mendoza, C., Szczepanski, L.J., Roschmann, R.A., Chen, A., Armstrong, G.K., Douglass, W., & Tyrrell, H. 2010. Efficacy and safety of different doses and retreatment of rituximab: a randomised, placebo-controlled trial in patients who are biological naive with active rheumatoid arthritis and an inadequate response to methotrexate (Study Evaluating Rituximab's Efficacy in MTX iNadequate rEsponders (SERENE)). *Ann.Rheum.Dis.*, 69, (9) 1629-1635 available from: PM:20488885

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G.H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K.P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H.G., Stefansson, T., Leifsson, B.G., Thorsteinsdottir, U., Lamb, J.R., Gulcher, J.R., Reitman, M.L., Kong, A., Schadt, E.E., & Stefansson, K. 2008. Genetics of gene expression and its effect on disease. *Nature*, 452, (7186) 423-428 available from: PM:18344981

Evans, D.M. & Purcell, S. 2012. Power calculations in genetic studies. *Cold Spring Harb.Protoc.*, 2012, (6) 664-674 available from: PM:22661434

Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., Amos, C.I., Padyukov, L., Toes, R.E., Huizinga, T.W., Wijmenga, C., Trynka, G., Franke, L., Westra, H.J., Alfredsson, L., Hu, X., Sandor, C., de Bakker, P.I., Davila, S., Khor, C.C., Heng, K.K., Andrews, R., Edkins, S., Hunt, S.E., Langford, C., Symmons, D., Concannon, P., Onengut-Gumuscu, S., Rich, S.S., Deloukas, P., Gonzalez-Gay, M.A., Rodriguez-Rodriguez, L., Arlsetig, L., Martin, J., Rantapaa-Dahlqvist, S., Plenge, R.M., Raychaudhuri, S., Klareskog, L., Gregersen, P.K., & Worthington, J. 2012. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat.Genet.*, 44, (12) 1336-1340 available from: PM:23143596

Eyre, S., Hinks, A., Bowes, J., Flynn, E., Martin, P., Wilson, A.G., Morgan, A.W., Emery, P., Steer, S., Hocking, L.J., Reid, D.M., Harrison, P., Wordsworth, P., Thomson, W., Worthington, J., & Barton, A. 2010. Overlapping genetic susceptibility variants

between three autoimmune disorders: rheumatoid arthritis, type 1 diabetes and coeliac disease. *Arthritis Res.Ther.*, 12, (5) R175 available from: PM:20854658

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., & Knight, J.C. 2014. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343, (6175) 1246949 available from: PM:24604202

Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., & Knight, J.C. 2012. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat.Genet.*, 44, (5) 502-510 available from: PM:22446964

Feldmann, M. 1996. What is the mechanism of action of anti-tumour necrosis factor-alpha antibody in rheumatoid arthritis? *Int.Arch.Allergy Immunol.*, 111, (4) 362-365 available from: PM:8957109

Fernandez-Vina, M., Fink, C.W., & Stastny, P. 1994. HLA associations in juvenile arthritis. *Clin.Exp.Rheumatol.*, 12, (2) 205-214 available from: PM:8039292

Filer, C., Ho, P., Smith, R.L., Griffiths, C., Young, H.S., Worthington, J., Bruce, I.N., & Barton, A. 2008. Investigation of association of the IL12B and IL23R genes with psoriatic arthritis. *Arthritis Rheum.*, 58, (12) 3705-3709 available from: PM:19035472

Fitau, J., Boulday, G., Coulon, F., Quillard, T., & Charreau, B. 2006. The adaptor molecule Lnk negatively regulates tumor necrosis factor-alpha-dependent VCAM-1 expression in endothelial cells through inhibition of the ERK1 and -2 pathways. *J.Biol.Chem.*, 281, (29) 20148-20159 available from: PM:16644735

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H.S., Rios, D., Ritchie, G.R., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y.A., Trevanion, S.,

Vandrovcova, J., Vilella, A.J., White, S., Wilder, S.P., Zadissa, A., Zamora, J., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J., Parker, A., Proctor, G., Vogel, J., & Searle, S.M. 2011. Ensembl 2011. *Nucleic Acids Res.*, 39, (Database issue) D800-D806 available from: PM:21045057

Gamazon, E.R., Huang, R.S., & Cox, N.J. 2013. SCAN: a systems biology approach to pharmacogenomic discovery. *Methods Mol.Biol.*, 1015, 213-224 available from: PM:23824859

Gamazon, E.R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E.O., Nicolae, D.L., Dolan, M.E., & Cox, N.J. 2010. SCAN: SNP and copy number annotation. *Bioinformatics.*, 26, (2) 259-262 available from: PM:19933162

Giannini, E.H., Ilowite, N.T., Lovell, D.J., Wallace, C.A., Rabinovich, C.E., Reiff, A., Higgins, G., Gottlieb, B., Singer, N.G., Chon, Y., Lin, S.L., & Baumgartner, S.W. 2009. Long-term safety and effectiveness of etanercept in children with selected categories of juvenile idiopathic arthritis. *Arthritis Rheum.*, 60, (9) 2794-2804 available from: PM:19714630

Gladman, D.D., Antoni, C., Mease, P., Clegg, D.O., & Nash, P. 2005. Psoriatic arthritis: epidemiology, clinical features, course, and outcome. *Ann.Rheum.Dis.*, 64 Suppl 2, ii14-ii17 available from: PM:15708927

Gladman, D.D., Shuckett, R., Russell, M.L., Thorne, J.C., & Schachter, R.K. 1987. Psoriatic arthritis (PSA)--an analysis of 220 patients. *Q.J.Med.*, 62, (238) 127-141 available from: PM:3659255

Glant, T.T., Mikecz, K., & Rauch, T.A. 2014. Epigenetics in the pathogenesis of rheumatoid arthritis. *BMC.Med.*, 12, 35 available from: PM:24568138

Glas, J., Stallhofer, J., Ripke, S., Wetzke, M., Pfennig, S., Klein, W., Epplen, J.T., Griga, T., Schiemann, U., Lacher, M., Koletzko, S., Folwaczny, M., Lohse, P., Goke, B., Ochsenkuhn, T., Muller-Myhsok, B., & Brand, S. 2009. Novel genetic risk markers for ulcerative colitis in the IL2/IL21 region are in epistasis with IL23R and suggest

a common genetic background for ulcerative colitis and celiac disease. *Am.J.Gastroenterol.*, 104, (7) 1737-1744 available from: PM:19455118

Glass, D.N. & Giannini, E.H. 1999. Juvenile rheumatoid arthritis as a complex genetic trait. *Arthritis Rheum.*, 42, (11) 2261-2268 available from: PM:10555018

Goldschmidt, T.J., Andersson, M., Malmstrom, V., & Holmdahl, R. 1992. Activated type II collagen reactive T cells are not eliminated by in vivo anti-CD4 treatment. Implications for therapeutic approaches on autoimmune arthritis. *Immunobiology*, 184, (4-5) 359-371 available from: PM:1350565

Gossec, L., Smolen, J.S., Gaujoux-Viala, C., Ash, Z., Marzo-Ortega, H., van der Heijde, D., FitzGerald, O., Aletaha, D., Balint, P., Boumpas, D., Braun, J., Breedveld, F.C., Burmester, G., Canete, J.D., de, W.M., Dagfinrud, H., de, V.K., Dougados, M., Helliwell, P., Kavanaugh, A., Kvien, T.K., Landewe, R., Luger, T., Maccarone, M., McGonagle, D., McHugh, N., McInnes, I.B., Ritchlin, C., Sieper, J., Tak, P.P., Valesini, G., Vencovsky, J., Winthrop, K.L., Zink, A., & Emery, P. 2012. European League Against Rheumatism recommendations for the management of psoriatic arthritis with pharmacological therapies. *Ann.Rheum.Dis.*, 71, (1) 4-12 available from: PM:21953336

Gottlieb, A., Menter, A., Mendelsohn, A., Shen, Y.K., Li, S., Guzzo, C., Fretzin, S., Kunynetz, R., & Kavanaugh, A. 2009. Ustekinumab, a human interleukin 12/23 monoclonal antibody, for psoriatic arthritis: randomised, double-blind, placebo-controlled, crossover trial. *Lancet*, 373, (9664) 633-640 available from: PM:19217154

Gravallese, E.M., Manning, C., Tsay, A., Naito, A., Pan, C., Amento, E., & Goldring, S.R. 2000. Synovial tissue in rheumatoid arthritis is a source of osteoclast differentiation factor. *Arthritis Rheum.*, 43, (2) 250-258 available from: PM:10693863

Gregersen, P.K., Silver, J., & Winchester, R.J. 1987. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum.*, 30, (11) 1205-1213 available from: PM:2446635

Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B.J., Diogo, D., Stahl, E.A., Gregersen, P.K., Worthington, J., Klareskog, L., Raychaudhuri, S., Plenge, R.M., Pasaniuc, B., & Price, A.L. 2013. Quantifying missing heritability at known GWAS loci. *PLoS.Genet.*, 9, (12) e1003993 available from: PM:24385918

Hafler, D.A., Compston, A., Sawcer, S., Lander, E.S., Daly, M.J., De Jager, P.L., de Bakker, P.I., Gabriel, S.B., Mirel, D.B., Ivinson, A.J., Pericak-Vance, M.A., Gregory, S.G., Rioux, J.D., McCauley, J.L., Haines, J.L., Barcellos, L.F., Cree, B., Oksenberg, J.R., & Hauser, S.L. 2007. Risk alleles for multiple sclerosis identified by a genomewide study. *N.Engl.J.Med.*, 357, (9) 851-862 available from: PM:17660530

Hamel, K.M., Cao, Y., Ashaye, S., Wang, Y., Dunn, R., Kehry, M.R., Glant, T.T., & Finnegan, A. 2011. B cell depletion enhances T regulatory cell activity essential in the suppression of arthritis. *J.Immunol.*, 187, (9) 4900-4906 available from: PM:21948985

Hammer, G.E., Turer, E.E., Taylor, K.E., Fang, C.J., Advincula, R., Oshima, S., Barrera, J., Huang, E.J., Hou, B., Malynn, B.A., Reizis, B., DeFranco, A., Criswell, L.A., Nakamura, M.C., & Ma, A. 2011. Expression of A20 by dendritic cells preserves immune homeostasis and prevents colitis and spondyloarthritis. *Nat.Immunol.*, 12, (12) 1184-1193 available from: PM:22019834

Hamshere, M.L., Holmans, P.A., McCarthy, G.M., Jones, L.A., Murphy, K.C., Sanders, R.D., Gray, M.Y., Zammit, S., Williams, N.M., Norton, N., Williams, H.J., McGuffin, P., O'Donovan, M.C., Craddock, N., Owen, M.J., & Cardno, A.G. 2011a. Phenotype evaluation and genomewide linkage study of clinical variables in schizophrenia. *Am.J.Med.Genet.B Neuropsychiatr.Genet.*, 156B, (8) 929-940 available from: PM:21960518

Hamshere, M.L., O'Donovan, M.C., Jones, I.R., Jones, L., Kirov, G., Green, E.K., Moskvina, V., Grozeva, D., Bass, N., McQuillin, A., Gurling, H., St, C.D., Young, A.H., Ferrier, I.N., Farmer, A., McGuffin, P., Sklar, P., Purcell, S., Holmans, P.A., Owen, M.J., & Craddock, N. 2011b. Polygenic dissection of the bipolar phenotype. *Br.J.Psychiatry*, 198, (4) 284-288 available from: PM:21972277

Han, C., Robinson, D.W., Jr., Hackett, M.V., Paramore, L.C., Fraeman, K.H., & Bala, M.V. 2006. Cardiovascular disease and risk factors in patients with rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis. *J.Rheumatol.*, 33, (11) 2167-2172 available from: PM:16981296

Heath, S.C., Gut, I.G., Brennan, P., McKay, J.D., Bencko, V., Fabianova, E., Foretova, L., Georges, M., Janout, V., Kabesch, M., Krokan, H.E., Elvestad, M.B., Lissowska, J., Mates, D., Rudnai, P., Skorpen, F., Schreiber, S., Soria, J.M., Syvanen, A.C., Meneton, P., Hercberg, S., Galan, P., Szeszenia-Dabrowska, N., Zaridze, D., Genin, E., Cardon, L.R., & Lathrop, M. 2008. Investigation of the fine structure of European populations with applications to disease association studies. *Eur.J.Hum.Genet.*, 16, (12) 1413-1429 available from: PM:19020537

Herrmann, M., Scholmerich, J., & Straub, R.H. 2000. Stress and rheumatic diseases. *Rheum.Dis.Clin.North Am.*, 26, (4) 737-63, viii available from: PM:11084942

Hinks, A., Barton, A., John, S., Bruce, I., Hawkins, C., Griffiths, C.E., Donn, R., Thomson, W., Silman, A., & Worthington, J. 2005. Association between the PTPN22 gene and rheumatoid arthritis and juvenile idiopathic arthritis in a UK population: further support that PTPN22 is an autoimmunity gene. *Arthritis Rheum.*, 52, (6) 1694-1699 available from: PM:15934099

Hinks, A., Barton, A., Shephard, N., Eyre, S., Bowes, J., Cargill, M., Wang, E., Ke, X., Kennedy, G.C., John, S., Worthington, J., & Thomson, W. 2009a. Identification of a novel susceptibility locus for juvenile idiopathic arthritis by genome-wide association analysis. *Arthritis Rheum.*, 60, (1) 258-263 available from: PM:19116933

Hinks, A., Cobb, J., Marion, M.C., Prahalad, S., Sudman, M., Bowes, J., Martin, P., Comeau, M.E., Sajuthi, S., Andrews, R., Brown, M., Chen, W.M., Concannon, P., Deloukas, P., Edkins, S., Eyre, S., Gaffney, P.M., Guthery, S.L., Guthridge, J.M., Hunt, S.E., James, J.A., Keddache, M., Moser, K.L., Nigrovic, P.A., Onengut-Gumuscu, S., Onslow, M.L., Rose, C.D., Rich, S.S., Steel, K.J., Wakeland, E.K., Wallace, C.A., Wedderburn, L.R., Woo, P., Bohnsack, J.F., Haas, J.P., Glass, D.N., Langefeld, C.D., Thomson, W., & Thompson, S.D. 2013. Dense genotyping of immune-related

disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat.Genet.*, 45, (6) 664-669 available from: PM:23603761

Hinks, A., Cobb, J., Sudman, M., Eyre, S., Martin, P., Flynn, E., Packham, J., Barton, A., Worthington, J., Langefeld, C.D., Glass, D.N., Thompson, S.D., & Thomson, W. 2012. Investigation of rheumatoid arthritis susceptibility loci in juvenile idiopathic arthritis confirms high degree of overlap. *Ann.Rheum.Dis.*, 71, (7) 1117-1121 available from: PM:22294642

Hinks, A., Eyre, S., Ke, X., Barton, A., Martin, P., Flynn, E., Packham, J., Worthington, J., & Thomson, W. 2010a. Association of the AFF3 gene and IL2/IL21 gene region with juvenile idiopathic arthritis. *Genes Immun.*, 11, (2) 194-198 available from: PM:20072139

Hinks, A., Eyre, S., Ke, X., Barton, A., Martin, P., Flynn, E., Packham, J., Worthington, J., & Thomson, W. 2010b. Overlap of disease susceptibility loci for rheumatoid arthritis and juvenile idiopathic arthritis. *Ann.Rheum.Dis.*, 69, (6) 1049-1053 available from: PM:19674979

Hinks, A., Ke, X., Barton, A., Eyre, S., Bowes, J., Worthington, J., Thompson, S.D., Langefeld, C.D., Glass, D.N., & Thomson, W. 2009b. Association of the IL2RA/CD25 gene with juvenile idiopathic arthritis. *Arthritis Rheum.*, 60, (1) 251-257 available from: PM:19116909

Hinks, A., Martin, P., Flynn, E., Eyre, S., Packham, J., Barton, A., Worthington, J., & Thomson, W. 2010c. Association of the CCR5 gene with juvenile idiopathic arthritis. *Genes Immun.* available from: PM:20463745

Hinks, A., Martin, P., Flynn, E., Eyre, S., Packham, J., Barton, A., Worthington, J., & Thomson, W. 2011. Subtype specific genetic associations for juvenile idiopathic arthritis: ERAP1 with the enthesitis related arthritis subtype and IL23R with juvenile psoriatic arthritis. *Arthritis Res.Ther.*, 13, (1) R12 available from: PM:21281511

Hirano, T., Matsuda, T., Turner, M., Miyasaka, N., Buchan, G., Tang, B., Sato, K., Shimizu, M., Maini, R., Feldmann, M., & . 1988. Excessive production of interleukin 6/B cell stimulatory factor-2 in rheumatoid arthritis. *Eur.J.Immunol.*, 18, (11) 1797-1801 available from: PM:2462501

Ho, P.Y., Barton, A., Worthington, J., Plant, D., Griffiths, C.E., Young, H.S., Bradburn, P., Thomson, W., Silman, A.J., & Bruce, I.N. 2008. Investigating the role of the HLA-Cw*06 and HLA-DRB1 genes in susceptibility to psoriatic arthritis: comparison with psoriasis and undifferentiated inflammatory arthritis. *Ann.Rheum.Dis.*, 67, (5) 677-682 available from: PM:17728335

Hoefkens, E., Nys, K., John, J.M., Van, S.K., Arijs, I., Van der Goten, J., Van, A.G., Agostinis, P., Rutgeerts, P., Vermeire, S., & Cleynen, I. 2013. Genetic association and functional role of Crohn disease risk alleles involved in microbial sensing, autophagy, and endoplasmic reticulum (ER) stress. *Autophagy.*, 9, (12) 2046-2055 available from: PM:24247223

Holloway, S.M., Porteous, M.E., Fitzpatrick, D.R., Crosbie, A.E., Cetnarskyj, R., Warner, J., & Barron, L. 1998. Presymptomatic testing for Huntington's disease by linkage and by direct mutation analysis: comparison of uptake of testing and characteristics of test applicants. *Genet.Couns.*, 9, (2) 103-111 available from: PM:9664206

Hou, S., Du, L., Lei, B., Pang, C.P., Zhang, M., Zhuang, W., Zhang, M., Huang, L., Gong, B., Wang, M., Zhang, Q., Hu, K., Zhou, Q., Qi, J., Wang, C., Tian, Y., Ye, Z., Liang, L., Yu, H., Li, H., Zhou, Y., Cao, Q., Liu, Y., Bai, L., Liao, D., Kijlstra, A., Xu, J., Yang, Z., & Yang, P. 2014. Genome-wide association analysis of Vogt-Koyanagi-Harada syndrome identifies two new susceptibility loci at 1p31.2 and 10q21.3. *Nat.Genet.*, 46, (9) 1007-1011 available from: PM:25108386

Howcroft, T.K., Weissman, J.D., Gegonne, A., & Singer, D.S. 2005. A T lymphocyte-specific transcription complex containing RUNX1 activates MHC class I expression. *J.Immunol.*, 174, (4) 2106-2115 available from: PM:15699141

Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A., & Scheet, P. 2009. Genotype-imputation accuracy across worldwide human populations. *Am.J.Hum.Genet.*, 84, (2) 235-250 available from: PM:19215730

Huang, Q.Q. & Pope, R.M. 2009. The role of toll-like receptors in rheumatoid arthritis. *Curr.Rheumatol.Rep.*, 11, (5) 357-364 available from: PM:19772831

Huffmeier, U., Lascorz, J., Bohm, B., Lohmann, J., Wendler, J., Mossner, R., Reich, K., Traupe, H., Kurrat, W., Burkhardt, H., & Reis, A. 2009. Genetic variants of the IL-23R pathway: association with psoriatic arthritis and psoriasis vulgaris, but no specific risk factor for arthritis. *J.Invest Dermatol.*, 129, (2) 355-358 available from: PM:18800148

Huffmeier, U., Uebe, S., Ekici, A.B., Bowes, J., Giardina, E., Korendowych, E., Juneblad, K., Apel, M., McManus, R., Ho, P., Bruce, I.N., Ryan, A.W., Behrens, F., Lascorz, J., Bohm, B., Traupe, H., Lohmann, J., Gieger, C., Wichmann, H.E., Herold, C., Steffens, M., Klareskog, L., Wienker, T.F., FitzGerald, O., Alenius, G.M., McHugh, N.J., Novelli, G., Burkhardt, H., Barton, A., & Reis, A. 2010. Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis. *Nat.Genet.*, 42, (11) 996-999 available from: PM:20953186

Hughes, T., Kim-Howard, X., Kelly, J.A., Kaufman, K.M., Langefeld, C.D., Ziegler, J., Sanchez, E., Kimberly, R.P., Edberg, J.C., Ramsey-Goldman, R., Petri, M., Reveille, J.D., Martin, J., Brown, E.E., Vila, L.M., Alarcon, G.S., James, J.A., Gilkeson, G.S., Moser, K.L., Gaffney, P.M., Merrill, J.T., Vyse, T.J., Alarcon-Riquelme, M.E., Nath, S.K., Harley, J.B., & Sawalha, A.H. 2011. Fine-mapping and transethnic genotyping establish IL2/IL21 genetic association with lupus and localize this genetic effect to IL21. *Arthritis Rheum.*, 63, (6) 1689-1697 available from: PM:21425124

Humby, F., Bombardieri, M., Manzo, A., Kelly, S., Blades, M.C., Kirkham, B., Spencer, J., & Pitzalis, C. 2009. Ectopic lymphoid structures support ongoing production of class-switched autoantibodies in rheumatoid synovium. *PLoS.Med.*, 6, (1) e1 available from: PM:19143467

Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D., Gwilliam, R., Takeuchi, F., McLaren, W.M., Holmes, G.K., Howdle, P.D., Walters, J.R., Sanders, D.S., Playford, R.J., Trynka, G., Mulder, C.J., Mearin, M.L., Verbeek, W.H., Trimble, V., Stevens, F.M., O'Morain, C., Kennedy, N.P., Kelleher, D., Pennington, D.J., Strachan, D.P., McArdle, W.L., Mein, C.A., Wapenaar, M.C., Deloukas, P., McGinnis, R., McManus, R., Wijmenga, C., & van Heel, D.A. 2008. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat.Genet.*, 40, (4) 395-402 available from: PM:18311140

Ichikawa, M., Goyama, S., Asai, T., Kawazu, M., Nakagawa, M., Takeshita, M., Chiba, S., Ogawa, S., & Kurokawa, M. 2008. AML1/Runx1 negatively regulates quiescent hematopoietic stem cells in adult hematopoiesis. *J.Immunol.*, 180, (7) 4402-4408 available from: PM:18354160

Jaakkola, J.J. & Gissler, M. 2005. Maternal smoking in pregnancy as a determinant of rheumatoid arthritis and other inflammatory polyarthropathies during the first 7 years of life. *Int.J.Epidemiol.*, 34, (3) 664-671 available from: PM:15649961

Kallberg, H., Jacobsen, S., Bengtsson, C., Pedersen, M., Padyukov, L., Garred, P., Frisch, M., Karlson, E.W., Klareskog, L., & Alfredsson, L. 2009. Alcohol consumption is associated with decreased risk of rheumatoid arthritis: results from two Scandinavian case-control studies. *Ann.Rheum.Dis.*, 68, (2) 222-227 available from: PM:18535114

Kallberg, H., Padyukov, L., Plenge, R.M., Ronnelid, J., Gregersen, P.K., van der Helm-van Mil AH, Toes, R.E., Huizinga, T.W., Klareskog, L., & Alfredsson, L. 2007. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. *Am.J.Hum.Genet.*, 80, (5) 867-875 available from: PM:17436241

Karlson, E.W., Lee, I.M., Cook, N.R., Manson, J.E., Buring, J.E., & Hennekens, C.H. 1999. A retrospective cohort study of cigarette smoking and risk of rheumatoid arthritis in female health professionals. *Arthritis Rheum.*, 42, (5) 910-917 available from: PM:10323446

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hinrichs, A.S., Learned, K., Lee, B.T., Li, C.H., Raney, B.J., Rhead, B., Rosenbloom, K.R., Sloan, C.A., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., & Kent, W.J. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, 42, (Database issue) D764-D770 available from: PM:24270787

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., & Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, 32, (Database issue) D493-D496 available from: PM:14681465

Karouzakis, E., Rengel, Y., Jungel, A., Kolling, C., Gay, R.E., Michel, B.A., Tak, P.P., Gay, S., Neidhart, M., & Ospelt, C. 2011. DNA methylation regulates the expression of CXCL12 in rheumatoid arthritis synovial fibroblasts. *Genes Immun.*, 12, (8) 643-652 available from: PM:21753787

Kavanaugh, A.F. & Ritchlin, C.T. 2006. Systematic review of treatments for psoriatic arthritis: an evidence based approach and basis for treatment guidelines. *J.Rheumatol.*, 33, (7) 1417-1421 available from: PM:16724373

Kazazi, F., Mathijs, J.M., Foley, P., & Cunningham, A.L. 1989. Variations in CD4 expression by human monocytes and macrophages and their relationships to infection with the human immunodeficiency virus. *J.Gen.Virol.*, 70 (Pt 10), 2661-2672 available from: PM:2677236

Keystone, E.C. 2003. Abandoned therapies and unpublished trials in rheumatoid arthritis. *Curr.Opin.Rheumatol.*, 15, (3) 253-258 available from: PM:12707578

Kidd, B.A., Ho, P.P., Sharpe, O., Zhao, X., Tomooka, B.H., Kanter, J.L., Steinman, L., & Robinson, W.H. 2008. Epitope spreading to citrullinated antigens in mouse models of autoimmune arthritis and demyelination. *Arthritis Res.Ther.*, 10, (5) R119 available from: PM:18826638

Kim, K., Bang, S.Y., Lee, H.S., Cho, S.K., Choi, C.B., Sung, Y.K., Kim, T.H., Jun, J.B., Yoo, D.H., Kang, Y.M., Kim, S.K., Suh, C.H., Shim, S.C., Lee, S.S., Lee, J., Chung, W.T., Choe,

J.Y., Shin, H.D., Lee, J.Y., Han, B.G., Nath, S.K., Eyre, S., Bowes, J., Pappas, D.A., Kremer, J.M., Gonzalez-Gay, M.A., Rodriguez-Rodriguez, L., Arlestig, L., Okada, Y., Diogo, D., Liao, K.P., Karlson, E.W., Raychaudhuri, S., Rantapaa-Dahlqvist, S., Martin, J., Klareskog, L., Padyukov, L., Gregersen, P.K., Worthington, J., Greenberg, J.D., Plenge, R.M., & Bae, S.C. 2014. High-density genotyping of immune loci in Koreans and Europeans identifies eight new rheumatoid arthritis risk loci. *Ann.Rheum.Dis.* available from: PM:24532676

Kirkham, B.W., Kavanaugh, A., & Reich, K. 2014. Interleukin-17A: a unique pathway in immune-mediated diseases: psoriasis, psoriatic arthritis and rheumatoid arthritis. *Immunology*, 141, (2) 133-142 available from: PM:23819583

Klareskog, L., Stolt, P., Lundberg, K., Kallberg, H., Bengtsson, C., Grunewald, J., Ronnelid, J., Harris, H.E., Ulfgren, A.K., Rantapaa-Dahlqvist, S., Eklund, A., Padyukov, L., & Alfredsson, L. 2006. A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. *Arthritis Rheum.*, 54, (1) 38-46 available from: PM:16385494

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., & Hoh, J. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308, (5720) 385-389 available from: PM:15761122

Knevel, R., de Rooy, D.P., Zhernakova, A., Grondal, G., Krabben, A., Steinsson, K., Wijmenga, C., Cavet, G., Toes, R.E., Huizinga, T.W., Gregersen, P.K., & van der Helm-van Mil AH 2013. Association of variants in IL2RA with progression of joint destruction in rheumatoid arthritis. *Arthritis Rheum.*, 65, (7) 1684-1693 available from: PM:23529819

Koczan, D., Drynda, S., Hecker, M., Drynda, A., Guthke, R., Kekow, J., & Thiesen, H.J. 2008. Molecular discrimination of responders and nonresponders to anti-TNF alpha therapy in rheumatoid arthritis by etanercept. *Arthritis Res.Ther.*, 10, (3) R50 available from: PM:18454843

Komine, O., Hayashi, K., Natsume, W., Watanabe, T., Seki, Y., Seki, N., Yagi, R., Sukzuki, W., Tamauchi, H., Hozumi, K., Habu, S., Kubo, M., & Satake, M. 2003. The Runx1 transcription factor inhibits the differentiation of naive CD4+ T cells into the Th2 lineage by repressing GATA3 expression. *J.Exp.Med.*, 198, (1) 51-61 available from: PM:12835475

Kurreeman, F.A., Daha, N.A., Chang, M., Catanese, J.J., Begovich, A.B., Huizinga, T.W., & Toes, R.E. 2009. Association of IL2RA and IL2RB with rheumatoid arthritis: a replication study in a Dutch population. *Ann.Rheum.Dis.*, 68, (11) 1789-1790 available from: PM:19822714

Lamana, A., Balsa, A., Rueda, B., Ortiz, A.M., Nuno, L., Miranda-Carus, M.E., Gonzalez-Escribano, M.F., Lopez-Nevot, M.A., Pascual-Salcedo, D., Martin, J., & Gonzalez-Alvaro, I. 2012. The TT genotype of the STAT4 rs7574865 polymorphism is associated with high disease activity and disability in patients with early arthritis. *PLoS.One.*, 7, (8) e43661 available from: PM:22937072

Lane, P.J., Gaspal, F.M., & Kim, M.Y. 2005. Two sides of a cellular coin: CD4(+)CD3- cells regulate memory responses and lymph-node organization. *Nat.Rev.Immunol.*, 5, (8) 655-660 available from: PM:16034364

Lazarevic, V., Chen, X., Shim, J.H., Hwang, E.S., Jang, E., Bolm, A.N., Oukka, M., Kuchroo, V.K., & Glimcher, L.H. 2011. T-bet represses T(H)17 differentiation by preventing Runx1-mediated activation of the gene encoding RORgammat. *Nat.Immunol.*, 12, (1) 96-104 available from: PM:21151104

Lee, S.K., Bridges, S.L., Jr., Koopman, W.J., & Schroeder, H.W., Jr. 1992. The immunoglobulin kappa light chain repertoire expressed in the synovium of a patient with rheumatoid arthritis. *Arthritis Rheum.*, 35, (8) 905-913 available from: PM:1642656

Leipe, J., Grunke, M., Dechant, C., Reindl, C., Kerzendorf, U., Schulze-Koops, H., & Skapenko, A. 2010. Role of Th17 cells in human autoimmune arthritis. *Arthritis Rheum.*, 62, (10) 2876-2885 available from: PM:20583102

Levanon, D., Glusman, G., Bangsow, T., Ben-Asher, E., Male, D.A., Avidan, N., Bangsow, C., Hattori, M., Taylor, T.D., Taudien, S., Blechschmidt, K., Shimizu, N., Rosenthal, A., Sakaki, Y., Lancet, D., & Groner, Y. 2001. Architecture and anatomy of the genomic locus encoding the human leukemia-associated transcription factor RUNX1/AML1. *Gene*, 262, (1-2) 23-33 available from: PM:11179664

Liu, D.J., Peloso, G.M., Zhan, X., Holmen, O.L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P.L., Goel, A., Zhang, H., Peters, U., Farrall, M., Orho-Melander, M., Kooperberg, C., McPherson, R., Watkins, H., Willer, C.J., Hveem, K., Melander, O., Kathiresan, S., & Abecasis, G.R. 2014. Meta-analysis of gene-level tests for rare variant association. *Nat.Genet.*, 46, (2) 200-204 available from: PM:24336170

Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekstrom, T.J., & Feinberg, A.P. 2013. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat.Biotechnol.*, 31, (2) 142-147 available from: PM:23334450

Liu, Y., Helms, C., Liao, W., Zaba, L.C., Duan, S., Gardner, J., Wise, C., Miner, A., Malloy, M.J., Pullinger, C.R., Kane, J.P., Saccone, S., Worthington, J., Bruce, I., Kwok, P.Y., Menter, A., Krueger, J., Barton, A., Saccone, N.L., & Bowcock, A.M. 2008. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS.Genet.*, 4, (3) e1000041 available from: PM:18369459

Love, T.J., Zhu, Y., Zhang, Y., Wall-Burns, L., Ogdie, A., Gelfand, J.M., & Choi, H.K. 2012. Obesity and the risk of psoriatic arthritis: a population-based study. *Ann.Rheum.Dis.*, 71, (8) 1273-1277 available from: PM:22586165

Lovell, D.J., Ruperto, N., Goodman, S., Reiff, A., Jung, L., Jarosova, K., Nemcova, D., Mouy, R., Sandborg, C., Bohnsack, J., Elewaut, D., Foeldvari, I., Gerloni, V., Rovensky, J., Minden, K., Vehe, R.K., Weiner, L.W., Horneff, G., Huppertz, H.I., Olson, N.Y., Medich, J.R., Carcereri-De-Prati, R., McIlraith, M.J., Giannini, E.H., & Martini, A. 2008. Adalimumab with or without methotrexate in juvenile rheumatoid arthritis. *N.Engl.J.Med.*, 359, (8) 810-820 available from: PM:18716298

Lowe, C.E., Cooper, J.D., Brusko, T., Walker, N.M., Smyth, D.J., Bailey, R., Bourget, K., Plagnol, V., Field, S., Atkinson, M., Clayton, D.G., Wicker, L.S., & Todd, J.A. 2007. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat.Genet.*, 39, (9) 1074-1082 available from: PM:17676041

MacGregor, A.J., Snieder, H., Rigby, A.S., Koskenvuo, M., Kaprio, J., Aho, K., & Silman, A.J. 2000. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.*, 43, (1) 30-37 available from: PM:10643697

Maeno, N., Takei, S., Nomura, Y., Imanaka, H., Hokonohara, M., & Miyata, K. 2002. Highly elevated serum levels of interleukin-18 in systemic juvenile idiopathic arthritis but not in other juvenile idiopathic arthritis subtypes or in Kawasaki disease: comment on the article by Kawashima et al. *Arthritis Rheum.*, 46, (9) 2539-2541 available from: PM:12355506

Maier, L.M., Lowe, C.E., Cooper, J., Downes, K., Anderson, D.E., Severson, C., Clark, P.M., Healy, B., Walker, N., Aubin, C., Oksenberg, J.R., Hauser, S.L., Compston, A., Sawcer, S., De Jager, P.L., Wicker, L.S., Todd, J.A., & Hafler, D.A. 2009. IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS.Genet.*, 5, (1) e1000322 available from: PM:19119414

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A., & Visscher, P.M. 2009. Finding the missing heritability of complex diseases. *Nature*, 461, (7265) 747-753 available from: PM:19812666

Martin, J.E., Assassi, S., Diaz-Gallo, L.M., Broen, J.C., Simeon, C.P., Castellvi, I., Vicente-Rabaneda, E., Fonollosa, V., Ortego-Centeno, N., Gonzalez-Gay, M.A., Espinosa, G., Carreira, P., Camps, M., Sabio, J.M., D'alfonso, S., Vonk, M.C., Voskuyl, A.E., Schuerwegh, A.J., Kreuter, A., Witte, T., Riemekasten, G., Hunzelmann, N., Airo,

P., Beretta, L., Scorza, R., Lunardi, C., Van, L.J., Chee, M.M., Worthington, J., Herrick, A., Denton, C., Fonseca, C., Tan, F.K., Arnett, F., Zhou, X., Reveille, J.D., Gorlova, O., Koeleman, B.P., Radstake, T.R., Vyse, T., Mayes, M.D., Alarcon-Riquelme, M.E., & Martin, J. 2013. A systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals new shared susceptibility loci. *Hum.Mol.Genet.*, 22, (19) 4021-4029 available from: PM:23740937

Martin, P., Barton, A., & Eyre, S. 2011. ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies. *Bioinformatics.*, 27, (1) 144-146 available from: PM:21177990

Mauri, C. & Bosma, A. 2012. Immune regulatory function of B cells. *Annu.Rev.Immunol.*, 30, 221-241 available from: PM:22224776

McAllister, K., Yarwood, A., Bowes, J., Orozco, G., Viatte, S., Diogo, D., Hocking, L.J., Steer, S., Wordsworth, P., Wilson, A.G., Morgan, A.W., Kremer, J.M., Pappas, D., Gregersen, P., Klareskog, L., Plenge, R., Barton, A., Greenberg, J., Worthington, J., & Eyre, S. 2013. Identification of BACH2 and RAD51B as rheumatoid arthritis susceptibility loci in a meta-analysis of genome-wide data. *Arthritis Rheum.*, 65, (12) 3058-3062 available from: PM:24022229

McClure, A., Lunt, M., Eyre, S., Ke, X., Thomson, W., Hinks, A., Bowes, J., Gibbons, L., Plant, D., Wilson, A.G., Marinou, I., Morgan, A.W., Emery, P., Steer, S., Hocking, L.J., Reid, D.M., Wordsworth, P., Harrison, P., Worthington, J., & Barton, A. 2009. Investigating the viability of genetic screening/testing for RA susceptibility using combinations of five confirmed risk loci. *Rheumatology.(Oxford)*, 48, (11) 1369-1374 available from: PM:19741008

McErlane, F., Foster, H.E., Davies, R., Lunt, M., Watson, K.D., Symmons, D.P., & Hyrich, K.L. 2013. Biologic treatment response among adults with juvenile idiopathic arthritis: results from the British Society for Rheumatology Biologics Register. *Rheumatology.(Oxford)*, 52, (10) 1905-1913 available from: PM:23873820

McGonagle, D., Conaghan, P.G., & Emery, P. 1999. Psoriatic arthritis: a unified concept twenty years on. *Arthritis Rheum.*, 42, (6) 1080-1086 available from: PM:10366099

McInnes, I.B., Liew, F.Y., & Gracie, J.A. 2005. Interleukin-18: a therapeutic target in rheumatoid arthritis? *Arthritis Res.Ther.*, 7, (1) 38-41 available from: PM:15642152

McInnes, I.B. & Schett, G. 2011. The pathogenesis of rheumatoid arthritis. *N.Engl.J.Med.*, 365, (23) 2205-2219 available from: PM:22150039

Menon, B., Gullick, N.J., Walter, G.J., Rajasekhar, M., Garrood, T., Evans, H.G., Taams, L.S., & Kirkham, B.W. 2014. IL-17+CD8+ T-cells are enriched in the joints of patients with psoriatic arthritis and correlate with disease activity and joint damage progression. *Arthritis Rheumatol.* available from: PM:24470327

Milicic, A., Lee, D., Brown, M.A., Darke, C., & Wordsworth, B.P. 2002. HLA-DR/DQ haplotype in rheumatoid arthritis: novel allelic associations in UK Caucasians. *J.Rheumatol.*, 29, (9) 1821-1826 available from: PM:12233873

Misjak, P., Bosze, S., Horvati, K., Pasztoi, M., Paloczi, K., Holub, M.C., Szakacs, F., Aradi, B., Gyorgy, B., Szabo, T.G., Nagy, G., Glant, T.T., Mikecz, K., Falus, A., & Buzas, E.I. 2013. The role of citrullination of an immunodominant proteoglycan (PG) aggrecan T cell epitope in BALB/c mice with PG-induced arthritis. *Immunol.Lett.*, 152, (1) 25-31 available from: PM:23578666

Moll, J.M. & Wright, V. 1973. Psoriatic arthritis. *Semin.Arthritis Rheum.*, 3, (1) 55-78 available from: PM:4581554

Morgan, A.W., Thomson, W., Martin, S.G., Carter, A.M., Erlich, H.A., Barton, A., Hocking, L., Reid, D.M., Harrison, P., Wordsworth, P., Steer, S., Worthington, J., Emery, P., Wilson, A.G., & Barrett, J.H. 2009. Reevaluation of the interaction between HLA-DRB1 shared epitope alleles, PTPN22, and smoking in determining susceptibility to autoantibody-positive and autoantibody-negative rheumatoid arthritis in a large UK Caucasian population. *Arthritis Rheum.*, 60, (9) 2565-2576 available from: PM:19714585

Morgan, M.E., Flierman, R., van Duivenvoorde, L.M., Witteveen, H.J., van, E.W., van Laar, J.M., de Vries, R.R., & Toes, R.E. 2005. Effective treatment of collagen-induced arthritis by adoptive transfer of CD25+ regulatory T cells. *Arthritis Rheum.*, 52, (7) 2212-2221 available from: PM:15986351

Morris, A.P. 2011. Transethnic meta-analysis of genomewide association studies. *Genet.Epidemiol.*, 35, (8) 809-822 available from: PM:22125221

Mottonen, T., Paimela, L., Leirisalo-Repo, M., Kautiainen, H., Ilonen, J., & Hannonen, P. 1998. Only high disease activity and positive rheumatoid factor indicate poor prognosis in patients with early rheumatoid arthritis treated with "sawtooth" strategy. *Ann.Rheum.Dis.*, 57, (9) 533-539 available from: PM:9849312

Murray, K.J., Luyrink, L., Grom, A.A., Passo, M.H., Emery, H., Witte, D., & Glass, D.N. 1996. Immunohistological characteristics of T cell infiltrates in different forms of childhood onset chronic arthritis. *J.Rheumatol.*, 23, (12) 2116-2124 available from: PM:8970050

Naderi, A., Ahmed, A.A., Barbosa-Morais, N.L., Aparicio, S., Brenton, J.D., & Caldas, C. 2004. Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling. *BMC.Genomics*, 5, (1) 9 available from: PM:15005798

Nair, R.P., Henseler, T., Jenisch, S., Stuart, P., Bichakjian, C.K., Lenk, W., Westphal, E., Guo, S.W., Christophers, E., Voorhees, J.J., & Elder, J.T. 1997. Evidence for two psoriasis susceptibility loci (HLA and 17q) and two novel candidate regions (16q and 20p) by genome-wide scan. *Hum.Mol.Genet.*, 6, (8) 1349-1356 available from: PM:9259283

Nejentsev, S., Walker, N., Riches, D., Egholm, M., & Todd, J.A. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324, (5925) 387-389 available from: PM:19264985

Newton, J.L., Harney, S.M., Wordsworth, B.P., & Brown, M.A. 2004. A review of the MHC genetics of rheumatoid arthritis. *Genes Immun.*, 5, (3) 151-157 available from: PM:14749714

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., & Cox, N.J. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS.Genet.*, 6, (4) e1000888 available from: PM:20369019

Nievergelt, C.M., Wineinger, N.E., Libiger, O., Pham, P., Zhang, G., Baker, D.G., & Schork, N.J. 2014. Chip-based direct genotyping of coding variants in genome wide association studies: Utility, issues and prospects. *Gene* available from: PM:24521671

Okada, Y., Diogo, D., Greenberg, J.D., Mouassess, F., Achkar, W.A., Fulton, R.S., Denny, J.C., Gupta, N., Mirel, D., Gabriel, S., Li, G., Kremer, J.M., Pappas, D.A., Carroll, R.J., Eyler, A.E., Trynka, G., Stahl, E.A., Cui, J., Saxena, R., Coenen, M.J., Guchelaar, H.J., Huizinga, T.W., Dieude, P., Mariette, X., Barton, A., Canhao, H., Fonseca, J.E., de, V.N., Tak, P.P., Moreland, L.W., Bridges, S.L., Jr., Miceli-Richard, C., Choi, H.K., Kamatani, Y., Galan, P., Lathrop, M., Raj, T., De Jager, P.L., Raychaudhuri, S., Worthington, J., Padyukov, L., Klareskog, L., Siminovitch, K.A., Gregersen, P.K., Mardis, E.R., Arayssi, T., Kazkaz, L.A., & Plenge, R.M. 2014a. Integration of sequence data from a Consanguineous family with genetic data from an outbred population identifies PLB1 as a candidate rheumatoid arthritis risk gene. *PLoS.One.*, 9, (2) e87645 available from: PM:24520335

Okada, Y., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Kawaguchi, T., Stahl, E.A., Kurreeman, F.A., Nishida, N., Ohmiya, H., Myouzen, K., Takahashi, M., Sawada, T., Nishioka, Y., Yukioka, M., Matsubara, T., Wakitani, S., Teshima, R., Tohma, S., Takasugi, K., Shimada, K., Murasawa, A., Honjo, S., Matsuo, K., Tanaka, H., Tajima, K., Suzuki, T., Iwamoto, T., Kawamura, Y., Tanii, H., Okazaki, Y., Sasaki, T., Gregersen, P.K., Padyukov, L., Worthington, J., Siminovitch, K.A., Lathrop, M., Taniguchi, A., Takahashi, A., Tokunaga, K., Kubo, M., Nakamura, Y., Kamatani, N., Mimori, T., Plenge, R.M., Yamanaka, H., Momohara, S., Yamada, R., Matsuda, F., & Yamamoto, K. 2012. Meta-analysis identifies nine new loci associated with rheumatoid arthritis

in the Japanese population. *Nat.Genet.*, 44, (5) 511-516 available from: PM:22446963

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R.R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J.C., Carroll, R.J., Eyler, A.E., Greenberg, J.D., Kremer, J.M., Pappas, D.A., Jiang, L., Yin, J., Ye, L., Su, D.F., Yang, J., Xie, G., Keystone, E., Westra, H.J., Esko, T., Metspalu, A., Zhou, X., Gupta, N., Mirel, D., Stahl, E.A., Diogo, D., Cui, J., Liao, K., Guo, M.H., Myouzen, K., Kawaguchi, T., Coenen, M.J., van Riel, P.L., van de Laar, M.A., Guchelaar, H.J., Huizinga, T.W., Dieude, P., Mariette, X., Bridges, S.L., Jr., Zhernakova, A., Toes, R.E., Tak, P.P., Miceli-Richard, C., Bang, S.Y., Lee, H.S., Martin, J., Gonzalez-Gay, M.A., Rodriguez-Rodriguez, L., Rantapaa-Dahlqvist, S., Arlestig, L., Choi, H.K., Kamatani, Y., Galan, P., Lathrop, M., Eyre, S., Bowes, J., Barton, A., de, V.N., Moreland, L.W., Criswell, L.A., Karlson, E.W., Taniguchi, A., Yamada, R., Kubo, M., Liu, J.S., Bae, S.C., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P.K., Raychaudhuri, S., Stranger, B.E., De Jager, P.L., Franke, L., Visscher, P.M., Brown, M.A., Yamanaka, H., Mimori, T., Takahashi, A., Xu, H., Behrens, T.W., Siminovitch, K.A., Momohara, S., Matsuda, F., Yamamoto, K., & Plenge, R.M. 2014b. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506, (7488) 376-381 available from: PM:24390342

Okuda, T., van, D.J., Hiebert, S.W., Grosveld, G., & Downing, J.R. 1996. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell*, 84, (2) 321-330 available from: PM:8565077

Omoyinmi, E., Hamaoui, R., Pesenacker, A., Nistala, K., Moncrieffe, H., Ursu, S., Wedderburn, L.R., & Woo, P. 2012. Th1 and Th17 cell subpopulations are enriched in the peripheral blood of patients with systemic juvenile idiopathic arthritis. *Rheumatology.(Oxford)*, 51, (10) 1881-1886 available from: PM:22772320

Ono, M., Yaguchi, H., Ohkura, N., Kitabayashi, I., Nagamura, Y., Nomura, T., Miyachi, Y., Tsukada, T., & Sakaguchi, S. 2007. Foxp3 controls regulatory T-cell function by interacting with AML1/Runx1. *Nature*, 446, (7136) 685-689 available from: PM:17377532

Orozco, G., Alizadeh, B.Z., Delgado-Vega, A.M., Gonzalez-Gay, M.A., Balsa, A., Pascual-Salcedo, D., Fernandez-Gutierrez, B., Gonzalez-Escribano, M.F., Petersson, I.F., van Riel, P.L., Barrera, P., Coenen, M.J., Radstake, T.R., van Leeuwen, M.A., Wijmenga, C., Koeleman, B.P., Alarcon-Riquelme, M., & Martin, J. 2008. Association of STAT4 with rheumatoid arthritis: a replication study in three European populations. *Arthritis Rheum.*, 58, (7) 1974-1980 available from: PM:18576336

Orozco, G., Eyre, S., Hinks, A., Bowes, J., Morgan, A.W., Wilson, A.G., Wordsworth, P., Steer, S., Hocking, L., Thomson, W., Worthington, J., & Barton, A. 2011. Study of the common genetic background for rheumatoid arthritis and systemic lupus erythematosus. *Ann.Rheum.Dis.*, 70, (3) 463-468 available from: PM:21068098

Orozco, G., Hinks, A., Eyre, S., Ke, X., Gibbons, L.J., Bowes, J., Flynn, E., Martin, P., Wilson, A.G., Bax, D.E., Morgan, A.W., Emery, P., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P., Thomson, W., Barton, A., & Worthington, J. 2009. Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23. *Hum.Mol.Genet.*, 18, (14) 2693-2699 available from: PM:19417005

Orozco, G., Rueda, B., & Martin, J. 2006. Genetic basis of rheumatoid arthritis. *Biomed.Pharmacother.*, 60, (10) 656-662 available from: PM:17055211

Palmer, L.J. & Cardon, L.R. 2005. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet*, 366, (9492) 1223-1234 available from: PM:16198771

Parkes, M., Cortes, A., van Heel, D.A., & Brown, M.A. 2013. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat.Rev.Genet.*, 14, (9) 661-673 available from: PM:23917628

Patterson, N., Price, A.L., & Reich, D. 2006. Population structure and eigenanalysis. *PLoS.Genet.*, 2, (12) e190 available from: PM:17194218

Perez-Alamino, R., Garcia-Valladares, I., Cuchacovich, R., Iglesias-Gamarra, A., & Espinoza, L.R. 2014. Are anti-CCP antibodies in psoriatic arthritis patients a biomarker of erosive disease? *Rheumatol.Int.* available from: PM:24515446

Perkel, J.M. 2013. Visiting "noncodarnia". *Biotechniques*, 54, (6) 301, 303-301, 304 available from: PM:23750541

Pesenacker, A.M., Bending, D., Ursu, S., Wu, Q., Nistala, K., & Wedderburn, L.R. 2013. CD161 defines the subset of FoxP3+ T cells capable of producing proinflammatory cytokines. *Blood*, 121, (14) 2647-2658 available from: PM:23355538

Petty, R.E., Southwood, T.R., Manners, P., Baum, J., Glass, D.N., Goldenberg, J., He, X., Maldonado-Cocco, J., Orozco-Alcala, J., Prieur, A.M., Suarez-Almazor, M.E., & Woo, P. 2004. International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *J.Rheumatol.*, 31, (2) 390-392 available from: PM:14760812

Pierer, M., Kaltenhauser, S., Arnold, S., Wahle, M., Baerwald, C., Hantzschel, H., & Wagner, U. 2006. Association of PTPN22 1858 single-nucleotide polymorphism with rheumatoid arthritis in a German cohort: higher frequency of the risk allele in male compared to female patients. *Arthritis Res.Ther.*, 8, (3) R75 available from: PM:16635271

Plenge, R.M., Padyukov, L., Remmers, E.F., Purcell, S., Lee, A.T., Karlson, E.W., Wolfe, F., Kastner, D.L., Alfredsson, L., Altshuler, D., Gregersen, P.K., Klareskog, L., & Rioux, J.D. 2005. Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am.J.Hum.Genet.*, 77, (6) 1044-1060 available from: PM:16380915

Plenge, R.M., Seielstad, M., Padyukov, L., Lee, A.T., Remmers, E.F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L.R., Li, W., Tan, A.K., Bonnard, C., Ong, R.T., Thalamuthu, A., Pettersson, S., Liu, C., Tian, C., Chen, W.V., Carulli, J.P., Beckman, E.M., Altshuler, D., Alfredsson, L., Criswell, L.A., Amos, C.I., Seldin, M.F., Kastner, D.L., Klareskog, L., & Gregersen, P.K. 2007. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *N.Engl.J.Med.*, 357, (12) 1199-1209 available from: PM:17804836

Polzer, K., Baeten, D., Soleiman, A., Distler, J., Gerlag, D.M., Tak, P.P., Schett, G., & Zwerina, J. 2008. Tumour necrosis factor blockade increases lymphangiogenesis in murine and human arthritic joints. *Ann.Rheum.Dis.*, 67, (11) 1610-1616 available from: PM:18174217

Prahalad, S. 2004. Genetics of juvenile idiopathic arthritis: an update. *Curr.Opin.Rheumatol.*, 16, (5) 588-594 available from: PM:15314499

Prahalad, S. 2006. Negative association between the chemokine receptor CCR5-Delta32 polymorphism and rheumatoid arthritis: a meta-analysis. *Genes Immun.*, 7, (3) 264-268 available from: PM:16541097

Prahalad, S., Hansen, S., Whiting, A., Guthery, S.L., Clifford, B., McNally, B., Zeft, A.S., Bohnsack, J.F., & Jorde, L.B. 2009. Variants in TNFAIP3, STAT4, and C12orf30 loci associated with multiple autoimmune diseases are also associated with juvenile idiopathic arthritis. *Arthritis Rheum.*, 60, (7) 2124-2130 available from: PM:19565500

Prakken, B., Albani, S., & Martini, A. 2011. Juvenile idiopathic arthritis. *Lancet*, 377, (9783) 2138-2149 available from: PM:21684384

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., & Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat.Genet.*, 38, (8) 904-909 available from: PM:16862161

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., & Willer, C.J. 2010. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.*, 26, (18) 2336-2337 available from: PM:20634204

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., & Sham, P.C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am.J.Hum.Genet.*, 81, (3) 559-575 available from: PM:17701901

Putz, S.M., Boehm, A.M., Stiewe, T., & Sickmann, A. 2012. iTRAQ analysis of a cell culture model for malignant transformation, including comparison with 2D-PAGE and SILAC. *J.Proteome.Res.*, 11, (4) 2140-2153 available from: PM:22313033

Quirke, A.M., Lugli, E.B., Wegner, N., Hamilton, B.C., Charles, P., Chowdhury, M., Ytterberg, A.J., Zubarev, R.A., Potempa, J., Culshaw, S., Guo, Y., Fisher, B.A., Thiele, G., Mikuls, T.R., & Venables, P.J. 2014. Heightened immune response to autocitrullinated Porphyromonas gingivalis peptidylarginine deiminase: a potential mechanism for breaching immunologic tolerance in rheumatoid arthritis. *Ann.Rheum.Dis.*, 73, (1) 263-269 available from: PM:23463691

Ravelli, A. & Martini, A. 2007. Juvenile idiopathic arthritis. *Lancet*, 369, (9563) 767-778 available from: PM:17336654

Raychaudhuri, S., Remmers, E.F., Lee, A.T., Hackett, R., Guiducci, C., Burtt, N.P., Gianniny, L., Korman, B.D., Padyukov, L., Kurreeman, F.A., Chang, M., Catanese, J.J., Ding, B., Wong, S., van der Helm-van Mil AH, Neale, B.M., Coblyn, J., Cui, J., Tak, P.P., Wolbink, G.J., Crusius, J.B., van der Horst-Bruinsma IE, Criswell, L.A., Amos, C.I., Seldin, M.F., Kastner, D.L., Ardlie, K.G., Alfredsson, L., Costenbader, K.H., Altshuler, D., Huizinga, T.W., Shadick, N.A., Weinblatt, M.E., de, V.N., Worthington, J., Seielstad, M., Toes, R.E., Karlson, E.W., Begovich, A.B., Klareskog, L., Gregersen, P.K., Daly, M.J., & Plenge, R.M. 2008. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat.Genet.*, 40, (10) 1216-1223 available from: PM:18794853

Raychaudhuri, S., Sandor, C., Stahl, E.A., Freudenberg, J., Lee, H.S., Jia, X., Alfredsson, L., Padyukov, L., Klareskog, L., Worthington, J., Siminovitch, K.A., Bae, S.C., Plenge, R.M., Gregersen, P.K., & de Bakker, P.I. 2012. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat.Genet.*, 44, (3) 291-296 available from: PM:22286218

Reich, D.E. & Lander, E.S. 2001. On the allelic spectrum of human disease. *Trends Genet.*, 17, (9) 502-510 available from: PM:11525833

Reinius, L.E., Acevedo, N., Joerink, M., Pershagen, G., Dahlen, S.E., Greco, D., Soderhall, C., Scheynius, A., & Kere, J. 2012. Differential DNA methylation in

purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS.One.*, 7, (7) e41361 available from: PM:22848472

Remmers, E.F., Plenge, R.M., Lee, A.T., Graham, R.R., Hom, G., Behrens, T.W., de Bakker, P.I., Le, J.M., Lee, H.S., Batliwalla, F., Li, W., Masters, S.L., Booty, M.G., Carulli, J.P., Padyukov, L., Alfredsson, L., Klareskog, L., Chen, W.V., Amos, C.I., Criswell, L.A., Seldin, M.F., Kastner, D.L., & Gregersen, P.K. 2007. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N.Engl.J.Med.*, 357, (10) 977-986 available from: PM:17804842

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., Fennell, T., Kirby, A., Latiano, A., Goyette, P., Green, T., Halfvarson, J., Haritunians, T., Korn, J.M., Kuruvilla, F., Lagace, C., Neale, B., Lo, K.S., Schumm, P., Torkvist, L., Dubinsky, M.C., Brant, S.R., Silverberg, M.S., Duerr, R.H., Altshuler, D., Gabriel, S., Lettre, G., Franke, A., D'Amato, M., McGovern, D.P., Cho, J.H., Rioux, J.D., Xavier, R.J., & Daly, M.J. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat.Genet.*, 43, (11) 1066-1073 available from: PM:21983784

Ropes M.W, Bennett, G.A., Cobb, S., Jacox, R., & Jessar, R.A. 1958. 1958 Revision of diagnostic criteria for rheumatoid arthritis. *Bull.Rheum.Dis.*, 9, (4) 175-176 available from: PM:13596783

Rosen, C.F., Mussani, F., Chandran, V., Eder, L., Thavaneswaran, A., & Gladman, D.D. 2012. Patients with psoriatic arthritis have worse quality of life than those with psoriasis alone. *Rheumatology.(Oxford)*, 51, (3) 571-576 available from: PM:22157469

Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., & Krumm, A. 2008. CTCF physically links cohesin to chromatin. *Proc.Natl.Acad.Sci.U.S.A*, 105, (24) 8309-8314 available from: PM:18550811

Ruperto, N., Lovell, D.J., Quartier, P., Paz, E., Rubio-Perez, N., Silva, C.A., bud-Mendoza, C., Burgos-Vargas, R., Gerloni, V., Melo-Gomes, J.A., Saad-Magalhaes, C.,

Chavez-Corrales, J., Huemer, C., Kivitz, A., Blanco, F.J., Foeldvari, I., Hofer, M., Horneff, G., Huppertz, H.I., Job-Deslandre, C., Loy, A., Minden, K., Punaro, M., Nunez, A.F., Sigal, L.H., Block, A.J., Nys, M., Martini, A., & Giannini, E.H. 2010. Long-term safety and efficacy of abatacept in children with juvenile idiopathic arthritis. *Arthritis Rheum.*, 62, (6) 1792-1802 available from: PM:20191582

Ruperto, N., Murray, K.J., Gerloni, V., Wulffraat, N., de Oliveira, S.K., Falcini, F., Dolezalova, P., Alessio, M., Burgos-Vargas, R., Corona, F., Vesely, R., Foster, H., Davidson, J., Zulian, F., Asplin, L., Baildam, E., Consuegra, J.G., Ozdogan, H., Saurenmann, R., Joos, R., Pistorio, A., Woo, P., & Martini, A. 2004. A randomized trial of parenteral methotrexate comparing an intermediate dose with a higher dose in children with juvenile idiopathic arthritis who failed to respond to standard doses of methotrexate. *Arthritis Rheum.*, 50, (7) 2191-2201 available from: PM:15248217

Sarembaud, J., Pinto, R., Rutledge, D.N., & Feinberg, M. 2007. Application of the ANOVA-PCA method to stability studies of reference materials. *Anal.Chim.Acta*, 603, (2) 147-154 available from: PM:17963834

Scher, J.U., Bretz, W.A., & Abramson, S.B. 2014. Periodontal disease and subgingival microbiota as contributors for rheumatoid arthritis pathogenesis: modifiable risk factors? *Curr.Opin.Rheumatol.* available from: PM:24807405

Seibl, R., Birchler, T., Loeliger, S., Hossle, J.P., Gay, R.E., Saurenmann, T., Michel, B.A., Seger, R.A., Gay, S., & Lauener, R.P. 2003. Expression and regulation of Toll-like receptor 2 in rheumatoid arthritis synovium. *Am.J.Pathol.*, 162, (4) 1221-1227 available from: PM:12651614

Seldin, M.F., Amos, C.I., Ward, R., & Gregersen, P.K. 1999. The genetics revolution and the assault on rheumatoid arthritis. *Arthritis Rheum.*, 42, (6) 1071-1079 available from: PM:10366098

Seldin, M.F., Shigeta, R., Laiho, K., Li, H., Saila, H., Savolainen, A., Leirisalo-Repo, M., Aho, K., Tuomilehto-Wolf, E., Kaarela, K., Kauppi, M., Alexander, H.C., Begovich, A.B., & Tuomilehto, J. 2005. Finnish case-control and family studies support PTPN22

R620W polymorphism as a risk factor in rheumatoid arthritis, but suggest only minimal or no effect in juvenile idiopathic arthritis. *Genes Immun.*, 6, (8) 720-722 available from: PM:16107870

Sham, P.C. & Purcell, S.M. 2014. Statistical power and significance testing in large-scale genetic studies. *Nat.Rev.Genet.*, 15, (5) 335-346 available from: PM:24739678

Shi, J., Knevel, R., Suwannalai, P., van der Linden, M.P., Janssen, G.M., van Veelen, P.A., Levarht, N.E., van der Helm-van Mil AH, Cerami, A., Huizinga, T.W., Toes, R.E., & Trouw, L.A. 2011. Autoantibodies recognizing carbamylated proteins are present in sera of patients with rheumatoid arthritis and predict joint damage. *Proc.Natl.Acad.Sci.U.S.A*, 108, (42) 17372-17377 available from: PM:21987802

Shi, J., van de Stadt, L.A., Levarht, E.W., Huizinga, T.W., Hamann, D., van, S.D., Toes, R.E., & Trouw, L.A. 2014. Anti-carbamylated protein (anti-CarP) antibodies precede the onset of rheumatoid arthritis. *Ann.Rheum.Dis.*, 73, (4) 780-783 available from: PM:24336334

Shi, J., van de Stadt, L.A., Levarht, E.W., Huizinga, T.W., Toes, R.E., Trouw, L.A., & van, S.D. 2013. Anti-carbamylated protein antibodies are present in arthralgia patients and predict the development of rheumatoid arthritis. *Arthritis Rheum.*, 65, (4) 911-915 available from: PM:23279976

Silman, A.J., MacGregor, A.J., Thomson, W., Holligan, S., Carthy, D., Farhan, A., & Ollier, W.E. 1993. Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Br.J.Rheumatol.*, 32, (10) 903-907 available from: PM:8402000

Smolen, J.S., Braun, J., Dougados, M., Emery, P., FitzGerald, O., Helliwell, P., Kavanaugh, A., Kvien, T.K., Landewe, R., Luger, T., Mease, P., Olivieri, I., Reveille, J., Ritchlin, C., Rudwaleit, M., Schoels, M., Sieper, J., Wit, M., Baraliakos, X., Betteridge, N., Burgos-Vargas, R., Collantes-Estevez, E., Deodhar, A., Elewaut, D., Gossec, L., Jongkees, M., Maccarone, M., Redlich, K., van den Bosch, F., Wei, J.C., Winthrop, K., & van der Heijde, D. 2014. Treating spondyloarthritis, including ankylosing spondylitis and psoriatic arthritis, to target: recommendations of an international task force. *Ann.Rheum.Dis.*, 73, (1) 6-16 available from: PM:23749611

Sroczynska, P., Lancrin, C., Kouskoff, V., & Lacaud, G. 2009. The differential activities of Runx1 promoters define milestones during embryonic hematopoiesis. *Blood*, 114, (26) 5279-5289 available from: PM:19858498

Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A., Zhernakova, A., Hinks, A., Guiducci, C., Chen, R., Alfredsson, L., Amos, C.I., Ardlie, K.G., Barton, A., Bowes, J., Brouwer, E., Burtt, N.P., Catanese, J.J., Coblyn, J., Coenen, M.J., Costenbader, K.H., Criswell, L.A., Crusius, J.B., Cui, J., de Bakker, P.I., De Jager, P.L., Ding, B., Emery, P., Flynn, E., Harrison, P., Hocking, L.J., Huizinga, T.W., Kastner, D.L., Ke, X., Lee, A.T., Liu, X., Martin, P., Morgan, A.W., Padyukov, L., Posthumus, M.D., Radstake, T.R., Reid, D.M., Seielstad, M., Seldin, M.F., Shadick, N.A., Steer, S., Tak, P.P., Thomson, W., van der Helm-van Mil AH, van der Horst-Bruinsma IE, van der Schoot, C.E., van Riel, P.L., Weinblatt, M.E., Wilson, A.G., Wolbink, G.J., Wordsworth, B.P., Wijmenga, C., Karlson, E.W., Toes, R.E., de, V.N., Begovich, A.B., Worthington, J., Siminovitch, K.A., Gregersen, P.K., Klareskog, L., & Plenge, R.M. 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat.Genet.*, 42, (6) 508-514 available from: PM:20453842

Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A., Kathiresan, S., Wijmenga, C., Gregersen, P.K., Alfredsson, L., Siminovitch, K.A., Worthington, J., de Bakker, P.I., Raychaudhuri, S., & Plenge, R.M. 2012. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat.Genet.*, 44, (5) 483-489 available from: PM:22446960

Stastny, P. 1976. Mixed lymphocyte cultures in rheumatoid arthritis. *J.Clin.Invest*, 57, (5) 1148-1157 available from: PM:1262462

Strand, V., Kimberly, R., & Isaacs, J.D. 2007. Biologic therapies in rheumatology: lessons learned, future directions. *Nat.Rev.Drug Discov.*, 6, (1) 75-92 available from: PM:17195034

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavare, S., Deloukas, P., &

Dermitzakis, E.T. 2007. Population genomics of human gene expression. *Nat.Genet.*, 39, (10) 1217-1224 available from: PM:17873874

Symmons, D., Turner, G., Webb, R., Asten, P., Barrett, E., Lunt, M., Scott, D., & Silman, A. 2002. The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century. *Rheumatology.(Oxford)*, 41, (7) 793-800 available from: PM:12096230

Symmons, D.P., Bankhead, C.R., Harrison, B.J., Brennan, P., Barrett, E.M., Scott, D.G., & Silman, A.J. 1997. 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.*, 40, (11) 1955-1961 available from: PM:9365083

Symmons, D.P., Jones, M., Osborne, J., Sills, J., Southwood, T.R., & Woo, P. 1996. Pediatric rheumatology in the United Kingdom: data from the British Pediatric Rheumatology Group National Diagnostic Register. *J.Rheumatol.*, 23, (11) 1975-1980 available from: PM:8923378

Szekanecz, Z. & Koch, A.E. 2008. Vascular involvement in rheumatic diseases: 'vascular rheumatology'. *Arthritis Res.Ther.*, 10, (5) 224 available from: PM:18947376

Taniuchi, I., Osato, M., Egawa, T., Sunshine, M.J., Bae, S.C., Komori, T., Ito, Y., & Littman, D.R. 2002. Differential requirements for Runx proteins in CD4 repression and epigenetic silencing during T lymphocyte development. *Cell*, 111, (5) 621-633 available from: PM:12464175

Taylor, P.C. & Feldmann, M. 2009. Anti-TNF biologic agents: still the therapy of choice for rheumatoid arthritis. *Nat.Rev.Rheumatol.*, 5, (10) 578-582 available from: PM:19798034

Taylor, W., Gladman, D., Helliwell, P., Marchesoni, A., Mease, P., & Mielants, H. 2006. Classification criteria for psoriatic arthritis: development of new criteria from a

large international study. *Arthritis Rheum.*, 54, (8) 2665-2673 available from: PM:16871531

The International HapMap Consortium 2003. The International HapMap Project. *Nature*, 426, (6968) 789-796 available from: PM:14685227

Thompson, S.D., Marion, M.C., Sudman, M., Ryan, M., Tsoras, M., Howard, T.D., Barnes, M.G., Ramos, P.S., Thomson, W., Hinks, A., Haas, J.P., Prahalad, S., Bohnsack, J.F., Wise, C.A., Punaro, M., Rose, C.D., Pajewski, N.M., Spigarelli, M., Keddache, M., Wagner, M., Langefeld, C.D., & Glass, D.N. 2012. Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. *Arthritis Rheum.*, 64, (8) 2781-2791 available from: PM:22354554

Thomson, W., Barrett, J.H., Donn, R., Pepper, L., Kennedy, L.J., Ollier, W.E., Silman, A.J., Woo, P., & Southwood, T. 2002. Juvenile idiopathic arthritis classified by the ILAR criteria: HLA associations in UK patients. *Rheumatology.(Oxford)*, 41, (10) 1183-1189 available from: PM:12364641

Thomson, W., Barton, A., Ke, X., Eyre, S., Hinks, A., Bowes, J., Donn, R., Symmons, D., Hider, S., Bruce, I.N., Wilson, A.G., Marinou, I., Morgan, A., Emery, P., Carter, A., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P., Strachan, D., & Worthington, J. 2007. Rheumatoid arthritis association at 6q23. *Nat.Genet.*, 39, (12) 1431-1433 available from: PM:17982455

Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F., Lowe, C.E., Szeszko, J.S., Hafler, J.P., Zeitels, L., Yang, J.H., Vella, A., Nutland, S., Stevens, H.E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L.J., Healy, B., Burren, O.S., Lam, A.A., Ovington, N.R., Allen, J., Adlem, E., Leung, H.T., Wallace, C., Howson, J.M., Guja, C., Ionescu-Tirgoviste, C., Simmonds, M.J., Heward, J.M., Gough, S.C., Dunger, D.B., Wicker, L.S., & Clayton, D.G. 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat.Genet.*, 39, (7) 857-864 available from: PM:17554260

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., & Raychaudhuri, S. 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat.Genet.*, 45, (2) 124-130 available from: PM:23263488

Tsoi, L.C., Spain, S.L., Knight, J., Ellinghaus, E., Stuart, P.E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J.E., Kang, H.M., Allen, M.H., McManus, R., Novelli, G., Samuelsson, L., Schalkwijk, J., Stahle, M., Burden, A.D., Smith, C.H., Cork, M.J., Estivill, X., Bowcock, A.M., Krueger, G.G., Weger, W., Worthington, J., Tazi-Ahnini, R., Nestle, F.O., Hayday, A., Hoffmann, P., Winkelmann, J., Wijmenga, C., Langford, C., Edkins, S., Andrews, R., Blackburn, H., Strange, A., Band, G., Pearson, R.D., Vukcevic, D., Spencer, C.C., Deloukas, P., Mrowietz, U., Schreiber, S., Weidinger, S., Koks, S., Kingo, K., Esko, T., Metspalu, A., Lim, H.W., Voorhees, J.J., Weichenthal, M., Wichmann, H.E., Chandran, V., Rosen, C.F., Rahman, P., Gladman, D.D., Griffiths, C.E., Reis, A., Kere, J., Nair, R.P., Franke, A., Barker, J.N., Abecasis, G.R., Elder, J.T., & Trembath, R.C. 2012. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat.Genet.*, 44, (12) 1341-1348 available from: PM:23143594

Valencia, X., Stephens, G., Goldbach-Mansky, R., Wilson, M., Shevach, E.M., & Lipsky, P.E. 2006. TNF downmodulates the function of human CD4+CD25hi T-regulatory cells. *Blood*, 108, (1) 253-261 available from: PM:16537805

Van Boxel, J.A. & Paget, S.A. 1975. Predominantly T-cell infiltrate in rheumatoid synovial membranes. *N.Engl.J.Med.*, 293, (11) 517-520 available from: PM:168488

van der Linden, M.P., van der, W.D., Ioan-Facsinay, A., Levarht, E.W., Stoeken-Rijsbergen, G., Huizinga, T.W., Toes, R.E., & van der Helm-van Mil AH 2009. Value of anti-modified citrullinated vimentin and third-generation anti-cyclic citrullinated peptide compared with second-generation anti-cyclic citrullinated peptide and rheumatoid factor in predicting disease outcome in undifferentiated arthritis and rheumatoid arthritis. *Arthritis Rheum.*, 60, (8) 2232-2241 available from: PM:19644872

van der, W.D., Houwing-Duistermaat, J.J., Toes, R.E., Huizinga, T.W., Thomson, W., Worthington, J., van der Helm-van Mil AH, & de Vries, R.R. 2009. Quantitative

heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheum.*, 60, (4) 916-923 available from: PM:19333951

van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C., Bethel, G., Holmes, G.K., Feighery, C., Jewell, D., Kelleher, D., Kumar, P., Travis, S., Walters, J.R., Sanders, D.S., Howdle, P., Swift, J., Playford, R.J., McLaren, W.M., Mearin, M.L., Mulder, C.J., McManus, R., McGinnis, R., Cardon, L.R., Deloukas, P., & Wijmenga, C. 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat.Genet.*, 39, (7) 827-829 available from: PM:17558408

Vang, T., Congia, M., Macis, M.D., Musumeci, L., Orru, V., Zavattari, P., Nika, K., Tautz, L., Tasken, K., Cucca, F., Mustelin, T., & Bottini, N. 2005. Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant. *Nat.Genet.*, 37, (12) 1317-1319 available from: PM:16273109

Vella, A., Cooper, J.D., Lowe, C.E., Walker, N., Nutland, S., Widmer, B., Jones, R., Ring, S.M., McArdle, W., Pembrey, M.E., Strachan, D.P., Dunger, D.B., Twells, R.C., Clayton, D.G., & Todd, J.A. 2005. Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *Am.J.Hum.Genet.*, 76, (5) 773-779 available from: PM:15776395

Viatte, S., Plant, D., Bowes, J., Lunt, M., Eyre, S., Barton, A., & Worthington, J. 2012. Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients. *Ann.Rheum.Dis.*, 71, (12) 1984-1990 available from: PM:22661644

Viken, M.K., Amundsen, S.S., Kvien, T.K., Boberg, K.M., Gilboe, I.M., Lilleby, V., Sollid, L.M., Forre, O.T., Thorsby, E., Smerdel, A., & Lie, B.A. 2005. Association analysis of the 1858C>T polymorphism in the PTPN22 gene in juvenile idiopathic arthritis and other autoimmune diseases. *Genes Immun.*, 6, (3) 271-273 available from: PM:15759012

Wallace, C.A., Giannini, E.H., Spalding, S.J., Hashkes, P.J., O'Neil, K.M., Zeft, A.S., Szer, I.S., Ringold, S., Brunner, H.I., Schanberg, L.E., Sundel, R.P., Milojevic, D., Punaro, M.G., Chira, P., Gottlieb, B.S., Higgins, G.C., Ilowite, N.T., Kimura, Y., Hamilton, S., Johnson, A., Huang, B., & Lovell, D.J. 2012. Trial of early aggressive therapy in polyarticular juvenile idiopathic arthritis. *Arthritis Rheum.*, 64, (6) 2012-2021 available from: PM:22183975

Wang, C., Ahlford, A., Laxman, N., Nordmark, G., Eloranta, M.L., Gunnarsson, I., Svenungsson, E., Padyukov, L., Sturfelt, G., Jonsen, A., Bengtsson, A.A., Truedsson, L., Rantapaa-Dahlqvist, S., Sjowall, C., Sandling, J.K., Ronnblom, L., & Syvanen, A.C. 2013a. Contribution of IKBKE and IFIH1 gene variants to SLE susceptibility. *Genes Immun.*, 14, (4) 217-222 available from: PM:23535865

Wang, J., Wang, X., Holz, J.D., Rutkowski, T., Wang, Y., Zhu, Z., & Dong, Y. 2013b. Runx1 is critical for PTH-induced onset of mesenchymal progenitor cell chondrogenic differentiation. *PLoS.One.*, 8, (9) e74255 available from: PM:24058535

Wang, J., Zhang, J., Li, K., Zhao, W., & Cui, Q. 2012. SpliceDisease database: linking RNA splicing and disease. *Nucleic Acids Res.*, 40, (Database issue) D1055-D1059 available from: PM:22139928

Wang, Z., Gerstein, M., & Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat.Rev.Genet.*, 10, (1) 57-63 available from: PM:19015660

Wehrens, E.J., Prakken, B.J., & van, W.F. 2013. T cells out of control--impaired immune regulation in the inflamed joint. *Nat.Rev.Rheumatol.*, 9, (1) 34-42 available from: PM:23390638

Weinblatt, M.E., Kavanaugh, A., Genovese, M.C., Jones, D.A., Musser, T.K., Grossbard, E.B., & Magilavy, D.B. 2013. Effects of fostamatinib (R788), an oral spleen tyrosine kinase inhibitor, on health-related quality of life in patients with active rheumatoid arthritis: analyses of patient-reported outcomes from a randomized, double-blind, placebo-controlled trial. *J.Rheumatol.*, 40, (4) 369-378 available from: PM:23378467

Weiss, K.M. & Clark, A.G. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.*, 18, (1) 19-24 available from: PM:11750696

Wellcome Trust Case Control Consortium 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447, (7145) 661-678 available from: PM:17554300

Wenink, M.H., Santegoets, K.C., Platt, A.M., van den Berg, W.B., van Riel, P.L., Garside, P., Radstake, T.R., & McInnes, I.B. 2012. Abatacept modulates proinflammatory macrophage responses upon cytokine-activated T cell and Toll-like receptor ligand stimulation. *Ann.Rheum.Dis.*, 71, (1) 80-83 available from: PM:21908454

Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., Zhernakova, A., Zhernakova, D.V., Veldink, J.H., Van den Berg, L.H., Karjalainen, J., Withoff, S., Uitterlinden, A.G., Hofman, A., Rivadeneira, F., 't Hoen, P.A., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A.B., Hernandez, D.G., Nalls, M.A., Homuth, G., Nauck, M., Radke, D., Volker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S.A., Enquobahrie, D.A., Lumley, T., Montgomery, G.W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R.C., Visscher, P.M., Knight, J.C., Psaty, B.M., Ripatti, S., Teumer, A., Frayling, T.M., Metspalu, A., van Meurs, J.B., & Franke, L. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat.Genet.*, 45, (10) 1238-1243 available from: PM:24013639

Wettenhall, J.M. & Smyth, G.K. 2004. limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics.*, 20, (18) 3705-3706 available from: PM:15297296

Wilson, C.L., Hine, D.W., Pradipta, A., Pearson, J.P., van, E.W., Robinson, J.H., & Knight, A.M. 2012. Presentation of the candidate rheumatoid arthritis autoantigen aggrecan by antigen-specific B cells induces enhanced CD4(+) T helper type 1 subset differentiation. *Immunology*, 135, (4) 344-354 available from: PM:22182481

Winchester, R., Minevich, G., Steshenko, V., Kirby, B., Kane, D., Greenberg, D.A., & Fitzgerald, O. 2012. HLA associations reveal genetic heterogeneity in psoriatic arthritis and in the psoriasis phenotype. *Arthritis Rheum.*, 64, (4) 1134-1144 available from: PM:22006066

Wong, W.F., Kohu, K., Nakamura, A., Ebina, M., Kikuchi, T., Tazawa, R., Tanaka, K., Kon, S., Funaki, T., Sugahara-Tobinai, A., Looi, C.Y., Endo, S., Funayama, R., Kurokawa, M., Habu, S., Ishii, N., Fukumoto, M., Nakata, K., Takai, T., & Satake, M. 2012. Runx1 deficiency in CD4+ T cells causes fatal autoimmune inflammatory lung disease due to spontaneous hyperactivation of cells. *J.Immunol.*, 188, (11) 5408-5420 available from: PM:22551552

Wright, C., Bergstrom, D., Dai, H., Marton, M., Morris, M., Tokiwa, G., Wang, Y., & Fare, T. 2008. Characterization of globin RNA interference in gene expression profiling of whole-blood samples. *Clin.Chem.*, 54, (2) 396-405 available from: PM:18089658

Xu, D., Jiang, H.R., Kewin, P., Li, Y., Mu, R., Fraser, A.R., Pitman, N., Kurowska-Stolarska, M., McKenzie, A.N., McInnes, I.B., & Liew, F.Y. 2008. IL-33 exacerbates antigen-induced arthritis by activating mast cells. *Proc.Natl.Acad.Sci.U.S.A*, 105, (31) 10913-10918 available from: PM:18667700

Yang, T.P., Beazley, C., Montgomery, S.B., Dimas, A.S., Gutierrez-Arcelus, M., Stranger, B.E., Deloukas, P., & Dermitzakis, E.T. 2010. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics.*, 26, (19) 2474-2476 available from: PM:20702402

Yano, F., Hojo, H., Ohba, S., Fukai, A., Hosaka, Y., Ikeda, T., Saito, T., Hirata, M., Chikuda, H., Takato, T., Kawaguchi, H., & Chung, U.I. 2013. A novel disease-modifying osteoarthritis drug candidate targeting Runx1. *Ann.Rheum.Dis.*, 72, (5) 748-753 available from: PM:23041841

Yeo, L., Lom, H., Juarez, M., Snow, M., Buckley, C.D., Filer, A., Raza, K., & Scheel-Toellner, D. 2014. Expression of FcRL4 defines a pro-inflammatory, RANKL-

producing B cell subset in rheumatoid arthritis. *Ann.Rheum.Dis.* available from: PM:24431391

Young, A., Koduri, G., Batley, M., Kulinskaya, E., Gough, A., Norton, S., & Dixey, J. 2007. Mortality in rheumatoid arthritis. Increased in the early course of disease, in ischaemic heart disease and in pulmonary fibrosis. *Rheumatology.(Oxford)*, 46, (2) 350-357 available from: PM:16908509

Zhang, X.J., Huang, W., Yang, S., Sun, L.D., Zhang, F.Y., Zhu, Q.X., Zhang, F.R., Zhang, C., Du, W.H., Pu, X.M., Li, H., Xiao, F.L., Wang, Z.X., Cui, Y., Hao, F., Zheng, J., Yang, X.Q., Cheng, H., He, C.D., Liu, X.M., Xu, L.M., Zheng, H.F., Zhang, S.M., Zhang, J.Z., Wang, H.Y., Cheng, Y.L., Ji, B.H., Fang, Q.Y., Li, Y.Z., Zhou, F.S., Han, J.W., Quan, C., Chen, B., Liu, J.L., Lin, D., Fan, L., Zhang, A.P., Liu, S.X., Yang, C.J., Wang, P.G., Zhou, W.M., Lin, G.S., Wu, W.D., Fan, X., Gao, M., Yang, B.Q., Lu, W.S., Zhang, Z., Zhu, K.J., Shen, S.K., Li, M., Zhang, X.Y., Cao, T.T., Ren, W., Zhang, X., He, J., Tang, X.F., Lu, S., Yang, J.Q., Zhang, L., Wang, D.N., Yuan, F., Yin, X.Y., Huang, H.J., Wang, H.F., Lin, X.Y., & Liu, J.J. 2009. Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. *Nat.Genet.*, 41, (2) 205-210 available from: PM:19169255

Zheng, G., Freidlin, B., Li, Z., & Gastwirth, J.L. 2005. Genomic control for association studies under various genetic models. *Biometrics*, 61, (1) 186-192 available from: PM:15737092

Zhernakova, A., Alizadeh, B.Z., Bevova, M., van Leeuwen, M.A., Coenen, M.J., Franke, B., Franke, L., Posthumus, M.D., van Heel, D.A., van der, S.G., Radstake, T.R., Barrera, P., Roep, B.O., Koeleman, B.P., & Wijmenga, C. 2007. Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. *Am.J.Hum.Genet.*, 81, (6) 1284-1288 available from: PM:17999365

Zhernakova, A., van Diemen, C.C., & Wijmenga, C. 2009. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat.Rev.Genet.*, 10, (1) 43-55 available from: PM:19092835