

Functional Genomics of Ankylosing Spondylitis

William Midwinter

Merton College

University of Oxford

A thesis submitted in partial

fulfilment of the requirements for the degree of

MSc by Research in Clinical Medicine

Trinity Term, 2020

Abstract

Ankylosing spondylitis (AS) is a chronic inflammatory arthritis primarily affecting the joints of the spine and pelvis. Its prevalence is highly correlated with the *HLA-B*27* allele but not all of the genetic risk for this complex disease is explained by this association. At least 113 other susceptibility loci have been identified, with the *NPEPPS* locus implicated by multiple high impact studies.

This thesis aimed to fine-map the AS disease-associated signal that overlaps the *NPEPPS* region on chromosome 17q21. Using statistical methods and epigenomic annotation it attempts to identify and prioritise the causal variant(s) and/or gene at this region. It then aims to validate the mechanistic role these prioritised variants play in AS.

Illumina CoreExome genotyping data from 8,244 AS cases and 14,542 healthy controls of European ancestry was used to fine-map the *NPEPPS* region (total cohort=22,786). Frequentist and Bayesian analysis were performed to identify sets of casual SNPs. The *NPEPPS* region was then annotated using epigenomic data from publicly available datasets and in-house disease-specific data to identify prioritised causal variants that overlapped with functional elements in the genomic landscape of AS. RT-qPCR was performed to compare the gene expression levels of *NPEPPS* and *TBKBP1* between AS cases and controls in PBMCs. Luciferase assays were performed to assess the function of the prioritised putative regulatory elements and the impact of specific genotypes.

Frequentist analysis in PLINK identified *rs4794057* (risk allele=T, OR=1.12, $p=2.4x10^{-11}$) as the lead SNP at the *NPEPPS* region. 96 SNPs were identified at genome-wide significance ($p<5x10^{-8}$), across 307,006bp. Bayesian analysis improved on this approach to identify a 99% credible set containing 62 SNPs across 211,521bp overlapping three genes (*NPEPPS*, *KPNB1*, *TBKBP1*). Annotation of the *NPEPPS* region with public epigenomic datasets and in-house AS-specific data identified three regulatory regions with a putative role in disease:

a CD4-specific putative enhancer region upstream of *NPEPPS*, a CD8-specific putative enhancer region between *NPEPPS* and *KPNB1*, and a putative promoter for *TBKBP1*. Following RT-qPCR experiments, lower expression of *NPEPPS* and *TBKBP1* was observed in AS cases compared to controls with a significant expression fold change of <0.64 and 0.02, respectively. In luciferase reporter assays the *NPEPPS-KPNB1* putative enhancer was shown to significantly increase relative luciferase activity (>2.5-fold).

This study has successfully improved on previous fine-mapping results of the *NPEPPS* region. Functional experiments identified lower expression of *NPEPPS* and *TBKBP1* in AS cases compared to healthy controls and an enhancer region was identified in the *NPEPPS-KPNB1* intergenic region tagged by the AS-associated SNP *rs4375701*. Further work is required to identify the regulatory region responsible for the differential expression observed and explain its mechanistic role in disease. This fine-mapping strategy should be implemented to improve the current understanding of known AS-associated loci which could potentially lead to the discovery of new disease pathways and new therapies for this debilitating disease.

Acknowledgements

Thank you to my supervisors; Julian Knight, Matt Brown and Paul Bowness for this wonderful opportunity and continued guidance and support throughout this project. A special mention to Carla Cohen who has unexpectedly provided me with the support, supervision, and constructive feedback to help drive this project to completion.

A huge thanks to both the Brown Group in Brisbane and Knight Group in Oxford for making me feel welcome in new environments and for all the support and knowledge you have shared with me throughout the two years.

Finally, thank you to my family and friends, old and new, who have supported me throughout this process and Jen who has always been there to constantly encourage me to achieve.

Declarations

I declare that all work presented in this thesis is my own unless otherwise stated.

The generation and quality control steps for the Illumina CoreExome genome-wide association study were performed by members of the Brown group, Queensland University of Technology. A subset of data covering a 1 Mb region overlapping the *NPEPPS* locus was provided for subsequent analysis in this study. Advice on frequentist and Bayesian analysis was provided by Dr Zhixiu Li and Prof Matt Brown.

AS patient and healthy volunteer samples used in gene expression assays were accessed from the Brown group biobank. AS patient and healthy control samples were previously collected, processed, and PMBCs stored by members of the Brown group. RNA from AS patients had previously been extracted. Dr Darrell Bessette and Associate Prof Tony Kenna provide guidance for RT-qPCR sample selection, assay design, and data analysis.

AS-specific annotation data was generated by members of the Knight group as part of the GENExpress ID study cohort. AS patient sample genomic DNA for luciferase assays was obtained from the GENExpress ID study cohort. Samples had previously been collected and processed as part of a collaborative effort by members of the Knight group. Advice for assay design and data analysis was provided by Dr Carla Cohen.

Contents

Abstract ..... **i**

Acknowledgements ..... **iii**

Declarations ..... **iv**

Contents ..... **v**

List of figures ..... **x**

List of tables ..... **xii**

Abbreviations ..... **xiv**

1 Introduction **1**

1.1 Ankylosing spondylitis 1

1.1.1 Clinical diagnosis of AS 2

1.1.2 Current treatments for AS 4

1.1.3 Aetiology and pathogenesis of AS 5

1.1.4 Genetics of AS 6

1.1.5 The MHC and HLA-B*27 7

1.1.6 Non-MHC genetics of AS 10

1.1.7 Immunological links to AS 12

1.1.8 Microbial Factors and their association with AS 14

1.2 Genome-wide association studies (GWAS) 15

1.2.1 GWAS in AS 16

1.3 Gene regulation landscape 19

1.3.1 Gene expression and regulation 19

1.3.2 Chromatin accessibility 21

1.3.3 Causal variants in gene regulation 22

1.4 Functional genomics 24

1.4.1 Approaches to fine-mapping 25

1.4.2 Statistical approaches to fine-map disease association 25

1.4.3 Functional annotation of the fine-mapping region 27

1.4.4 Functional experiments for prioritised variant investigation 29

1.5 Aims and objectives 31
1.5.1 Project aims 31

2 Materials and methods 32

2.1 Introduction 32

2.2 Statistical fine-mapping 33
2.2.1 Genotyping data 33
2.2.2 Frequentist analysis 34
2.2.3 Bayesian analysis 35

2.3 Functional Annotation 36
2.3.1 Regional annotation with available datasets 36
2.3.2 Publicly available datasets 36
2.3.3 Ankylosing spondylitis specific datasets 37

2.4 RT-qPCR gene expression assay 38
2.4.1 Samples 38
2.4.2 RNA extraction 40
2.4.3 cDNA synthesis 41
2.4.4 RT-qPCR 41
2.4.5 Data cleaning 44
2.4.6 Gene expression analysis 45

2.5 Luciferase dual-reporter assay 46
2.5.1 Samples 46
2.5.2 Generation of inserts 46
2.5.3 Sub-cloning into TOPO vector 47
2.5.4 Ligation into pGL4.23 vector 49
2.5.5 Transfection into HEK-293T cells 51
2.5.6 Luciferase assay 52
2.5.7 Data analysis of luciferase assay data 52

3 Computational approaches for fine-mapping the *NPEPPS* region 54

3.1 Introduction 54
3.1.1 Association of the *NPEPPS* region in previous GWAS studies 54
3.1.2 CoreExome GWAS cohort and dataset 57
3.1.3 Fine-mapping strategy 57
3.1.4 Statistical approaches for fine-mapping 58

3.1.5 Frequentist approaches 59
3.1.6 Bayesian approaches 59
3.1.7 Conditional analysis 60
3.1.8 Linkage disequilibrium 61
3.1.9 Aims 62
3.2 Results 63
3.2.1 Frequentist analysis of *NPEPPS* region prioritises 96 SNPs at genome-wide significance 63
3.2.2 Bayesian Analysis of *NPEPPS* region identifies 62 SNPs in the 99% credible set 65
3.2.3 Conditional analysis does not identify any independently associated signals in the *NPEPPS* region 69
3.2.4 Previously reported independent *TBKBP1*/*TBX21* signal not replicated 71
3.2.5 Linkage disequilibrium of the *NPEPPS* region 73
3.3 Discussion and conclusion 77
4 Annotation of *NPEPPS* region to identify variants for functional investigation 80
4.1 Introduction 80
4.1.1 Experimental methods to define DNA regulatory regions 81
4.1.2 Bioinformatic tools available for SNP prioritisation 83
4.1.3 Visualisation of publicly available data 85
4.1.4 In-house AS datasets used for annotation 86
4.1.5 Aim 87
4.2 Results 88
4.2.1 Annotation of *NPEPPS* region using publicly available SNP prioritisation tools 88
4.2.2 Annotation of *NPEPPS* region from in-house data 96
4.2.3 Regions selected for functional follow-up 101
4.3 Discussion and conclusion 106
5 Gene expression of *NPEPPS* and *TBKBP1* 111
5.1 Introduction 111
5.1.1 Assays for gene expression 112
5.1.2 RT-qPCR using SYBR green technology 114

5.1.3 Primer Design for RT-qPCR 114
5.1.4 Data analysis using $2^{-\Delta\Delta Ct}$ method for RT-qPCR 116
5.1.5 Aim 116
5.2 Results 117
5.2.1 AS case and healthy volunteer cohort 117
5.2.2 RT-qPCR experimental optimisation and primer efficiencies 117
5.2.3 Sample QC during RNA extraction and cDNA synthesis 118
5.2.4 Sample QC following RT-qPCR 118
5.2.5 Selection of appropriate reference genes for normalisation 120
5.2.6 Gene expression of *NPEPPS* and *TBKBP1* in AS cases and healthy controls 124
5.2.7 Genotype specific analysis of genome-wide significant SNPs 129
5.3 Discussion and conclusion 134
6 Functional follow up by luciferase reporter assays 139
6.1 Introduction 139
6.1.1 Luciferase reporter assay 139
6.1.2 The Dual-Glo luciferase assay system 140
6.1.3 Reporter constructs for investigating putative promoters and enhancers 141
6.1.4 Prioritised regions for investigation 144
6.1.5 Aim 145
6.2 Results 146
6.2.1 Haplotype analysis of the *NPEPPS* region 146
6.2.2 Generation of pGL4-based constructs containing putative regulatory regions 148
6.2.3 Luciferase assay experimental design and data QC 149
6.2.4 Selection of appropriate control vector: effect of insert size on luciferase activity 151
6.2.5 Region 2 insert DNA displays enhancer activity 153
6.2.6 The genotype at *rs4375701* does not significantly affect enhancer activity 155
6.2.7 The orientation of the insert significantly affects enhancer activity 156
6.3 Discussion and conclusion 158
7 General discussion 161
7.1 Successes and limitations of the fine-mapping approach 161

7.2 Cell-type and disease specific annotation is key to successful prioritisation 162
7.3 Functional experiments and the remaining challenge of identifying disease mechanism 163
7.4 Future work 165
7.4.1 Fine-mapping of all AS associated regions using CoreExome data 165
7.4.2 Cell-type and disease specific data are essential for annotation 167
7.4.3 Replication of functional experiments in disease-specific cell-types 168
7.4.4 Alternative and additional functional experiments to investigate putative regulatory variants 169
7.5 Conclusion 172
8 Appendices 173
8.1 Supplementary figures 173
8.1.1 Chapter 3 supplementary figures 173
8.1.2 Chapter 5 supplementary figures 175
8.2 Supplementary tables 177
8.2.1 Chapter 2 supplementary tables 177
8.2.2 Chapter 3 supplementary tables 178
8.2.3 Chapter 4 supplementary tables 182
8.2.4 Chapter 5 supplementary tables 186
8.2.5 Supplementary table of online tool websites 188
9 Bibliography 189

List of figures

Figure 2.1: Flow chart of steps performed in *NPEPPS* region fine-mapping study 32
Figure 2.2: Schematic of *NPEPPS* and *TBKBP1* RT-qPCR primer locations 44
Figure 2.3: pGL4.23 [luc2/minP] vector map showing *Bam*HI and *Sal*I restriction enzyme digest sites in relation to luciferase gene location 49
Figure 3.1: Association plots of *NPEPPS* region from previous studies 56
Figure 3.2: CoreExome frequentist association plot for the *NPEPPS* region 64
Figure 3.3: CoreExome frequentist association plot for the *NPEPPS* region depicting the lead SNPs from previous AS GWAS studies 65
Figure 3.4: CoreExome Bayesian analysis association plot for the *NPEPPS* region 66
Figure 3.5: Conditional analysis on the top four frequentist GWAS association SNPs at the *NPEPPS* region 70
Figure 3.6: Conditional analysis on the top four Bayesian association SNPs at the *NPEPPS* region 71
Figure 3.7: CoreExome frequentist association plot and conditional analysis plot for the *NPEPPS* region in relation to the *TBX21* lead SNP *rs11657479* 72
Figure 3.8: Frequentist association plot showing LD structure to the lead SNP .. 73
Figure 3.9: LD structure and haplotype estimation of 96 genome-wide significant SNPs as calculated in Haploview 75
Figure 4.1: Annotation of SNPs in *NPEPPS* region using publicly available tools 95
Figure 4.2: Chromatin landscape of the *NPEPPS* region and regions selected for further functional validation 100
Figure 4.3: Annotation of *NPEPPS* region around prioritised Region 1 103
Figure 4.4: Annotation of *NPEPPS* region around prioritised Region 2 104
Figure 4.5: Annotation of *NPEPPS* region around prioritised Region 3 105
Figure 5.1: Batch effects in AS cases due to cDNA synthesis preparations differences 120

Figure 5.2: Scatterplot showing the mean C_t values for three reference genes across all sample 122
Figure 5.3: Comparison of correlation observed in reference gene data 123
Figure 5.4: *NPEPPS* gene expression is lower in AS cases (n = 69) compared to healthy controls (n = 65) 126
Figure 5.5: *TBKBP1* gene expression is significantly lower in AS cases (n = 69) compared to healthy controls (n = 65) 128
Figure 5.6: eQTL analysis of gene expression of *NPEPPS* show no significant difference between genotypes for *rs11869940* at all four *NPEPPS* primer sets 131
Figure 5.7: eQTL analysis of gene expression of *TBKBP1* shows no significant difference between genotypes for *rs67919208* 133
Figure 6.1: Vector maps of Promega PGL4-series reporter constructs used for luciferase assays 143
Figure 6.2: Relative luciferase activity of pGL4.23 [luc2/minP] control vectors containing different size inserts 153
Figure 6.3: Relative luciferase activity of Region 2 inserts compared to size-matched control 154
Figure 6.4: Comparison of luciferase activity of genotype specific Region 2 inserts 156
Figure 6.5: Comparison of luciferase activity of orientation-specific Region 2 inserts 157
Supplementary Figure 3.1: CoreExome Bayesian analysis association plot for the *NPEPPS* region highlighting the prioritised region from van de Bunt 2015 173
Supplementary Figure 3.2: Correlation between Frequentist -log10(P-value) and log10(Bayes Factor) for 62 SNPs in 99% credible set. 174
Supplementary Figure 5.1: RT-qPCR mean Ct values for each primer set across two cell-line dilution series 175
Supplementary Figure 5.2: Log2(fold change) for *NPEPPS* and *TBKBP1* from healthy controls to AS cases showing a decrease in expression in AS cases across all primers 176

List of tables

Table 1.1: Summary of the main AS GWAS studies performed in Caucasian cohorts with further detail on *NPEPPS* locus 18
Table 2.1: Summary cohort data of AS patients and healthy controls used in RT-qPCR gene expression assay. 40
Table 2.2: Summary of primers used in RT-qPCR reaction 43
Table 2.3: Primer sequences used to amplify regions of interest for luciferase reporter assay experiment 48
Table 3.1: Summary statistics of top 10 ranked SNPs following frequentist analysis and Bayesian analysis 68
Table 3.2: Table to show LD blocks for genome-wide significant SNPs as calculated by PLINK and Haploview. 76
Table 4.1: Summary of Ensembl VEP output showing prioritised SNPs that overlap regulatory features 90
Table 4.2: SNPs classified as Category 1 or 2 variants by RegulomeDB 93
Table 4.3: Results of CADD SNP prioritisation showing SNPs that are classified as top 10% of deleterious substitutions in human genome 94
Table 4.4: H3K27ac master peak overlaps with computational prioritised SNPs 97
Table 4.5: ATAC master peak overlaps with computational prioritised SNPs 99
Table 5.1: Summary statistics for C_t values of reference genes across all samples. 122
Table 6.1: Genotype information for samples used as template to amplify regions of interest 147
Table 6.1: pGL4.23-based vectors used in luciferase dual-reporter assay 149
Table 6.2: Control vectors used in luciferase dual-reporter assay 151
Supplementary Table 2.1: *NPEPPS* primer capture table for *NPEPPS* transcripts 177
Supplementary Table 3.1: Frequentist Analysis genome-wide significant SNPs (n = 96). 178
Supplementary Table 3.2: Frequentist Analysis suggestive significant SNPs (n = 193). 179

Supplementary Table 3.3: Bayesian Analysis 99% credible set (n = 62). 180
Supplementary Table 3.4: Bayesian Analysis 95% credible set (n = 43). 181
Supplementary Table 4.1: HaploReg annotation of all prioritised SNPs 183
Supplementary Table 4.2: RegulomeDB classification and ranking of all prioritised SNPs. 184
Supplementary Table 4.3: CADD classification and ranking of all prioritised SNPs. 185
Supplementary Table 5.1: Primer efficiencies calculated for each primer for the two cell-lines (C1R-B27 and U937) 186
Supplementary Table 5.2: Summary eQTL analysis data of gene expression of *NPEPPS* for *rs11869940* at all four *NPEPPS* primer sets..... 187
Supplementary Table 5.3: Summary eQTL analysis data of gene expression of *TBKBP1* for *rs67919208*..... 187
Supplementary Table 8.2.5: Table of online tool websites used in project..... 188

Abbreviations

AS	Ankylosing spondylitis
ASAS	Assessment of Spondyloarthritis International Society
ASDAS	Ankylosing spondylitis disease activity score
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
BASDAI	Bath ankylosing spondylitis disease activity index
BASMI	Bath ankylosing spondylitis metrology index
BF	Bayes Factor
ChIP-seq	Chromatin immunoprecipitation sequencing
CRP	C-reactive protein
C_t	Cycle threshold
CTCF	CCCTC-binding factor
DMARD	Disease-modifying anti-rheumatic drug
DNase-seq	DNase I hypersensitive site sequencing
ENCODE	Encyclopedia of DNA elements consortium
eQTL	expression quantitative trait loci
ER	Endoplasmic reticulum
ERAP1/2	Endoplasmic reticulum protein 1/2
ESR	Erythrocyte sedimentation rate
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements by sequencing
GTEx	Genotype-Tissue Expression project
GWAS	Genome-wide association study
HRC	Haplotype Reference Consortium
IBD	Inflammatory bowel disease
IGAS	International genetics of ankylosing spondylitis consortium
LD	Linkage disequilibrium
MHC	Major histocompatibility complex
mNYAS	Modified New York criteria for ankylosing spondylitis
mRNA	messenger RNA
NSAID	Non-steroid anti-inflammatory drug
PBMC	Peripheral blood mononuclear cells
PO	Posterior odds
PP	Posterior probability
PPA	Posterior probability of association
PsA	Psoriatic arthritis
PSA	Puromycin sensitive aminopeptidase
QC	Quality control
ReA	Reactive arthritis
RNA-seq	RNA sequencing

RT-qPCR	Reverse transcription quantitative polymerase chain reaction
SpA	Spondyloarthropathy
SNP	Single nucleotide polymorphism
TASC	The Australo-Anglo-American Spondyloarthritis Consortium
TF	Transcription factor
TNF	Tumor necrosis factor
TRI	Translational Research Institute
TSS	Transcriptional start site
USpA	Undifferentiated spondyloarthritis
UTR	Untranslated region
VEP	Variant effect predictor
WashU	Washington University in St. Louis
WCHG	Wellcome Centre for Human Genetics
WTCCC2	Wellcome Trust Case Control Consortium 2

1 Introduction

This thesis aims to fine-map the ankylosing spondylitis disease-associated genetic association signal identified by genome-wide association studies (GWAS) in a European cohort that overlaps the *NPEPPS* gene on chromosome 17q21. Through the use of computational approaches, it will attempt to identify the variants responsible for the disease association signal, and then focus in on SNPs and genes prioritised at this region to functionally link the identified variants to mechanism and disease.

1.1 Ankylosing spondylitis

Ankylosing spondylitis (AS) is a chronic inflammatory arthritis primarily affecting the joints of the spine and pelvis. AS is the prototypic phenotype of the spondyloarthropathies (SpA) that include reactive arthritis (ReA), psoriatic arthritis (PsA), arthritis of inflammatory bowel disease (IBD), juvenile-onset spondyloarthritis and undifferentiated spondyloarthritis (USpA) (Brown et al., 2016, Li and Brown, 2017). This complex immune-mediated disease manifests as chronic inflammation causing pain and stiffness in the back, buttocks and neck which can lead to the formation of new bone and ankylosis (fusion) of the affected joints if left untreated. Generally, AS patients suffer from severe back pain, reduced spinal mobility, and impaired physical function which has an impact on their quality of life. Associated comorbidities that are known with AS affect sites including the eyes (uveitis), skin (psoriasis), gastrointestinal tract (inflammatory bowel disease), lungs, and heart. Symptoms develop in the late

teens and early twenties with an average age of onset at 24 years old. In the European population, men are likely to be affected 2-3 times more often than women (Lee et al., 2007).

1.1.1 Clinical diagnosis of AS

AS is diagnosed by the combination of clinical and radiographic features set out in the modified New York (mNYAS) classification criteria (van de Linden et al., 1984). These clinical criteria include; greater than 3 months reported back pain that improves with exercise but not relieved by rest, limited movement of the lumbar spine in the sagittal and frontal planes, and limited chest expansion measures for age and sex. The radiological criteria are assessed by signs of sacroiliitis visible on x-ray with differing weight given to the grade of sacroiliitis and whether it is bilateral or unilateral. These criteria are suitable for established disease that exhibits radiographic damage of the sacroiliitis but are not applicable in early disease. Here the Assessment of Spondyloarthritis International Society (ASAS) criteria should be used for the assessment of axial spondyloarthritis (Sieper et al., 2009). This criterion is applicable to patients who are below 45 years and have experienced greater than three months back pain. A positive diagnosis of axial spondyloarthritis is identified through either sacroiliitis on MRI imaging with one or more additional features or for *HLA-B*27* positive individuals with 2 or more additional features. The additional features for a positive diagnosis include, inflammatory back pain, arthritis, enthesitis, uveitis, psoriasis, Crohn's disease or ulcerative colitis, elevated C-reactive protein (CRP), family history of SpA. This may be confirmed as AS in later life with the development of

radiographic damage of the sacroiliitis. A test for the patient's *HLA-B*27* status is also a strong indicator in the diagnosis pathway as the combination of being *HLA-B*27* positive with any clinical or radiological features is highly suggestive of AS and a leading diagnostic identifier for axial spondyloarthritis.

Inflammation associated with active disease is measured using two clinical markers, CRP and erythrocyte sedimentation rate (ESR). These two biomarkers are not specific to disease associated inflammation so are not reliable measures of disease activity but are used to track active inflammation and control of disease.

Disease activity is assessed using the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), a six-question self-administered questionnaire regarding the major symptoms of AS (Garrett et al., 1994). This measure is used to assess the level of inflammation being experienced and is a key feature in the criteria for treatment with biologic drugs. The Ankylosing Spondylitis Disease Activity Score (ASDAS) uses BASDAI as a key element and build on it to create an index by which to assess disease activity. The Bath Ankylosing Spondylitis Metrology Index (BASMI), which consists of 5 reproducible and sensitive measures of spinal mobility, is used to assess and track patient's mobility and response to treatment.

These tests, assessment criteria, and scoring systems are used in combination to diagnose and track disease. As such, a confirmed diagnosis of AS my take many years as radiological signs may not be clear in early disease. It is estimated that the average delay in diagnosis from the start of symptoms is 8 to 10 years (Sykes

et al., 2015). No single test on its own is able to give a definitive diagnosis of AS and further discoveries may lead to greater understanding of this disease and the improvement of diagnosis.

1.1.2 Current treatments for AS

Current clinical practice involves treating the symptoms of AS and reducing inflammation, however patients still experience disease progression and continued ankylosis (Kenna et al., 2016).

There are two lines of treatment that are used initially for the treatment of suspected AS. The first line of treatment for patients with AS is physiotherapy and exercise to maintain and improve posture and spinal movement. The aim of targeted exercise is to prevent the stiffening of joints and reduce pain to help the AS patient to maintain function. Non-steroid anti-inflammatory drugs (NSAIDs) are used to manage pain and reduce swelling at the joints. Stronger analgesics may also be prescribed to help deal with pain, but these drugs have no effect in halting disease progression.

If symptoms are not controlled by exercise and painkillers, then biological treatment may be recommended. The main biological treatment in current use are anti-Tumour Necrosis Factor (anti-TNF) drugs which aim to prevent the inflammatory signalling of TNF (Tumor Necrosis Factor). Anti-TNF therapy is advised for people with high disease activity despite treatment with NSAIDs. Anti-TNF drugs include etanercept, adalimumab, golimumab, certolizumab and

infliximab. Alternative pathways that are the target of biologic drugs are the IL-23 and IL-17 inflammatory pathways, with IL-17 inhibitors in widespread clinical use for the treatment of AS and psoriasis. Biological drugs targeting the IL-17 pathway include secukinumab and ixekizumab. Disease-modifying anti-rheumatic drugs (DMARDs) may be used to treat synovitis at sites other than the spine.

If AS is left untreated or cannot be controlled by drugs, the disease may lead to spinal deformities that require surgical treatments. Surgical interventions are only required in rare cases with extreme spinal deformations or at alternative sites such as the hips to improve pain and movement following severe damage to joint sites.

To date, there have been no pharmaceutical treatments that can stop ankylosis or result in drug-free remission (Kenna et al., 2016). This still leaves a window of opportunity for the discovery and characterisation of new drug targets and treatment pathways to improve the therapeutic options for patients with AS. A clear starting point to investigate alternative targets are the genetic regions that have previously been associated with disease.

1.1.3 Aetiology and pathogenesis of AS

AS is an immune-mediated disease that evidence suggests develops due to a complex interplay between genetic and environmental factors. Understanding how a disease develops and the factors involved in its progression is crucial in attempting to identify the cause of the disease and potential routes and targets to

treat and cure the disease. In ankylosing spondylitis genetic factors, combined with an environmental insult, trigger inflammation. Over time these flares of inflammation can eventually lead to new bone growth and fusion of the sacrio-illiac joints and vertebrae. The interaction between genetic factors and the inflammatory response are crucial to understanding how AS develops, and of particular interest is the interaction between HLA-B*27 and the T-cell response.

1.1.4 Genetics of AS

In 1961, the first evidence that genetic risk factors were involved in AS was established with the report that the disease has a tendency to reoccur within families (de Blecourt et al., 1961). The reported risk ratio for an AS patient's sibling or first-degree relatives is >52 compared to the general population (Geirsson et al., 2010, Thjodleifsson et al., 2007, Brown et al., 2000) and the heritability of AS established by twin studies is >90% (Brown et al., 1997). Further to this, the disease concordance rate in monozygotic twins is 63% compared to 12.5% for dizygotic twins (Brown et al., 1997). This evidence indicates that the risk of developing disease is strongly dependent on genetic factors.

This confirmation that genetic risk factors play a key role in the development of disease was demonstrated in the 1970s with the association of the major histocompatibility complex (MHC) class I allele *HLA-B*27* with AS (Schlosstein et al., 1973, Brewerton et al., 1973, Caffrey and James, 1973). *HLA-B*27* is the most important genetic factors associated with AS and research into this alleles

role in disease has dominated research in the field since its discovery. The prevalence of AS is closely correlated with the frequency of *HLA-B*27* in a population and varies in different countries, with a prevalence of 0.55% reported in the European population (Braun et al., 1998, Stolwijk et al., 2016). Subsequent GWAS studies have identified more than 100 genetic associations with AS (see section 1.2.1 GWAS in AS below). This section will discuss the key genetic risk factors linked to disease and the evidence supporting these associations.

1.1.5 The MHC and HLA-B*27

AS has been associated with the inheritance of *HLA-B*27* since the early 1970's (Schlosstein et al., 1973, Brewerton et al., 1973) and represents one of the strongest known genetic associations in common disease. The major genetic variant associated with AS is the MHC allele *HLA-B*27*, with 80-95% of AS patients carrying this HLA allele (International Genetics of Ankylosing Spondylitis Consortium et al., 2013). *HLA-B* has the strongest association with AS driven by the *HLA-B*27* allele. There are more than 200 *HLA-B*27* subtypes described (Robinson et al., 2020) with the key ancestral subtype *HLA-B*27:05*. Other *HLA-B*27* alleles that have been associated with disease risk include; **27:02*, **27:03*, **27:04*, **27:07*, **27:08*, **27:10*. The key variant that defines these alleles is located in the protein binding grove at amino acid position 97 and this location is hypothesised to be critical in the mechanism of disease (Cortes et al., 2015). Other *HLA-B* alleles have been associated with AS including; *B*13*, *B*40*, *B*47*, and *B*51* as risk alleles and *B*57* and *B*07* in a protective context (Cortes et al., 2015). The association of the MHC class I alleles implicate CD8+ T cells as a cell

type in disease. There are also associations of MHC class II genes following conditional analysis, *HLA-DRB1* and *HLA-DPB1*, supporting the investigation of CD4+ T cells in AS (Cortes et al., 2015).

Although HLA-B*27 has been linked to a pathogenic role in AS for almost half a century its exact role still remains unclear. Multiple hypotheses have been proposed and the likelihood is that more than one may be involved in disease. Four main hypotheses have been proposed to date to explain the mechanism behind how HLA-B*27 induces AS:

(i) Arthritogenetic peptide theory – HLA-B*27 presents peptides that drive a cell-mediated immune response causing AS (Benjamin and Parham, 1992). In this theory, HLA-B*27 has a central role in presenting peptides to CD8+ cytotoxic T cells. However, the exact self or environmental peptides that potentially bind to HLA-B*27 and are presented to CD8+ T cells still remain to be elucidated (Benjamin and Parham, 1992).

(ii) HLA-B*27 homodimer model – an abnormal interaction occurs between KIR and homodimers of HLA-B*27, this drives IL-23 production and subsequent inflammation leading to disease (Allen et al., 1999).

(iii) Endoplasmic reticulum (ER) stress model – slow folding and frequent misfolding of HLA-B*27 lead to accumulation of the class I molecule in the ER. This induces an ER-stress response and the production of IL-23 leading to inflammation and disease (Mear et al., 1999). The ER-

stress activates an unfolded protein response which leads to the upregulation of inflammatory pathways.

(iv) Mucosal Immunodeficiency theory – HLA-B*27 fails to present bacterial peptides leading to increased mucosal invasion of bacteria. This leads to excess IL-23 pathway activation in individuals with upregulated IL-23 pathways (Kenna and Brown, 2013).

Although 80-95% of AS patients are *HLA-B*27* positive, 1-5% of the general population are also carriers of *HLA-B*27* and not all of these will develop AS. Alongside this there are 5-20% of AS patients who are *HLA-B*27* negative yet still develop the disease. This indicates that there are numerous other factors that influence the development of AS including genetics factors that are non-*HLA-B*27* variants. The MHC determines ~20% of the risk of developing AS, with a further ~7% currently explained by other known non-MHC loci (Ellinghaus et al., 2016) highlighting the polygenic nature of this disease.

To date, at least 114 genetic loci have been associated with AS (Brown, 2018, Li and Brown, 2017, Ellinghaus et al., 2016). The vast majority of these are non-MHC genetic factors and this has led to the discovery of novel molecular pathways that may be implicated in disease progression.

1.1.6 Non-MHC genetics of AS

1.1.6.1 Aminopeptidase genes associated with disease

Genes of the M1-aminopeptidase family were first associated with AS in 2007 with the high level of significance of two SNPs in the *ERAP1* gene (then known as ARTS1) (Wellcome Trust Case Control Consortium and The Australo-Anglo-American Spondylitis Consortium, 2007). This was one of the first non-MHC genes associated with AS. Three genes in this family have been shown to be associated with AS; *ERAP1* and *ERAP2* (Endoplasmic Reticulum Aminopeptidases 1 and 2) on chromosome 5q15 and *NPEPPS* (Puromycin Sensitive Aminopeptidase) on chromosome 17q21. These genes all encode proteins with a role in peptide trimming to appropriate lengths prior to presentation by HLA class I molecules (Stoltze et al., 2000).

ERAP1 is active in the endoplasmic reticulum (ER). The most strongly AS-associated variant is *rs30187* (International Genetics of Ankylosing Spondylitis Consortium et al., 2013) and this finding has been widely replicated in different ethnic groups including; Han Chinese (Davidson et al., 2009), Korean (Bang et al., 2011), and Iranian populations (Mahmoudi et al., 2012). The replication of this disease association across multiple ethnicities is supportive of the concept that AS is driven by combinations of common variants rather than rare variants. Disease-associated *ERAP1* variants increase AS risk only in *HLA-B*27* and *HLA-B*40* individuals (Evans et al., 2011). Such variants are commonly located in the coding region for the active site of the enzyme, supporting the idea of disrupted

peptide presentation as a key molecular mechanism in disease (Nguyen et al., 2011).

Variants in *ERAP2* have also been associated with AS in multiple studies (Wellcome Trust Case Control Consortium and The Australo-Anglo-American Spondylitis Consortium, 2007, Tsui et al., 2010). The nonsense mutation of *rs2248374* lead to lower expression of ERAP2 with absence of ERAP2 proposed to be protective in AS (Andres et al., 2010). Unlike *ERAP1*, there has been no specific epistasis observed with HLA-B*27. Both *ERAP1* and *ERAP2* have been extensively studied, and key genetic variants for these genes have been determined.

NPEPPS, located on chromosome 17q21 encodes the enzyme puromycin-sensitive aminopeptidase (PSA), a ubiquitously expressed cytoplasmic enzyme. PSA is reported to trims the N-terminal amino acids of its substrates and plays a role in MHC class I antigen processing and presentation (Stoltze et al., 2000). PSA is hypothesised to act in the same functional pathway as the other M1-aminopeptidases, ERAP1 and ERAP2, either by acting to reduce HLA class I presentation or by altering peptide handling and presentation leading to autoreactivity against an arthritogenic peptide (International Genetics of Ankylosing Spondylitis Consortium et al., 2013).

The association of *NPEPPS* with AS has not been investigated in detail and the key genetic variants at this locus are yet to be determined. This region has not currently been shown to be associated with AS in other populations at genome-

wide significance although there is suggestive evidence of disease-association in East Asians (International Genetics of Ankylosing Spondylitis Consortium et al., 2013). In Ellinghaus and colleagues (2016) analysis of five chronic inflammatory disease this region was only significantly associated with disease in individuals with AS. Although, SNPs in this region have been associated with multiple sclerosis at genome-wide significance (International Multiple Sclerosis Genetics Consortium et al., 2013). Due to its link in function to other disease associated aminopeptidases the associated region around *NPEPPS* is a clear candidate for fine-mapping and follow-up investigations and is the subject of this thesis.

1.1.7 Immunological links to AS

AS is an immune-mediated disease linked to other diseases including inflammatory bowel disease, anterior uveitis, and psoriasis. This suggests that this group of diseases may share similar genetics and therefore share some common immunological processes in the cause of disease.

1.1.7.1 The IL-17/IL-23 pathway in AS

The discovery of the association of *IL23R* with AS in 2007 (Wellcome Trust Case Control Consortium and The Australo-Anglo-American Spondylitis Consortium, 2007) was the first genetic link to potential causative pathways in disease other than through HLA-B*27. The *IL23R* gene encodes the IL-23 receptor, the receptor for the pro-inflammatory cytokine IL-23. This cytokine has an important role in the activation of pro-inflammatory cells including Th17 cells, gamma-delta T cells, and NK cells. When ligated to IL-23, IL23R signals a cascade through

JAK2 and STAT3 which leads to the production of inflammatory cytokines including IL-17 and IL-22. *IL23R* gene polymorphisms have also been associated with other autoimmune diseases including IBD and psoriasis.

Further genes linked to the IL-23/IL-27 pathway have subsequently been associated with AS; *IL6R*, *IL12B*, *IL27*, *CARD9*, *RUNX3*, *STAT3*, *TYK2*. These associations are evidence of the importance of this inflammatory pathway in the development and progression of this disease. Such pathways are the target of biological drug in the treatment of disease. Current anti-IL-17 drugs include secukinumab and ixekizumab along with the IL-12/IL-23 inhibitor ustekinumab.

1.1.7.2 The TNF pathway in disease

The TNF pathway has been linked to AS with genetic associations at *TNFRSF1A* (Evans et al., 2011, Wellcome Trust Case Control Consortium and The Australo-Anglo-American Spondylitis Consortium, 2007), *TRADD* (Pointon et al., 2010), and *TNFSF15* (Zinovieva et al., 2009). TNF is a cytokine that is an important mediator of inflammation and is proposed to play a role in disease during the initial inflammation of affected joints (Sieper et al., 2008). Overexpression of TNF has been reported in the sacroiliac joints of AS patients (Braun et al., 1995) implicating this cytokine at the site of disease. TNF has also been the main target of biologic treatment in AS (Braun et al., 2002). Current anti-TNF drugs include etanercept, adalimumab, and infliximab.

TBKBP1 overlaps the region of genetic association studied by this project. *TBKBP1* encodes TBK1 binding protein 1, an adaptor protein that binds TBK1

(TANK-binding kinase 1), a serine/threonine protein kinase in the TNF/NFKB pathway. *TBKBP1* variants are proposed to contribute to disease through an altered role on this inflammatory pathway, a pathway targeted in current AS treatments using anti-TNF drugs.

1.1.8 Microbial Factors and their association with AS

In addition to the above genetics, environmental factors may also make a sizeable contribution to the pathogenesis of AS. The microbiome, mechanical stress, drug and toxin exposure, and infection all may play a role.

AS has been known to coexist with intestinal inflammation with 5-10% of patients with AS also developing clinically diagnosed IBD, and a further 70% developing subclinical gut inflammation (Thomas and Brown, 2010, Mielants et al., 1991). There is strong genetic overlap between AS and IBD. These two diseases commonly occur together in families (Thjodleifsson et al., 2007). Microbial infection may act as a triggering factor of the host immune system and the location of the initial antigenic exposure leading to the development of AS driven by the IL-23 pathway (Brown, 2018). Multiple genes associated with AS are involved in mucosal immunity, including differentiation and activation of gut lymphoid cells (*TBX21*, *RUNX3*, *EOMES*, *ICOSLG*, *NKX2-3* and *ZMIZ1*) (Brown et al., 2016) and bacterial sensing or with alterations of the gut microbiome (*TLR4*, *AHR*, *NOD2*, *GPR35*, *GPR65*, *FUT2*) (Brown, 2018) further supporting a role of the gut in disease. In the 2010 study investigating an overlap between AS and Crohn’s disease strong association was identified for variants at 1q32 near

the *KIF21B* gene and *STAT3* as well as the confirmation of the *IL23R* gene (Danoy et al., 2010).

1.2 Genome-wide association studies (GWAS)

Genome-wide association studies examine the whole genome for regions where DNA sequence variation is associated with the phenotype of the trait being explored. It builds on the common disease common variant hypothesis, that common disorders are likely to be influenced by common genetic variation and that each variant has a small effect that combine together to cause disease. This hypothesis-free genetic association study is most frequently performed on genotyping array, where thousands of individuals can be genotyped at hundreds of thousands to millions of sites of variation with the aim of identifying the genetic differences between groups of individuals. This method has been widely used in the study of common complex human traits to identify genomic regions on chromosomes that identify the genetic markers of that trait.

The key findings of GWAS is the identification of a lead single nucleotide polymorphism (SNP) at each trait associated region. At any one region the number of significantly associated SNPs can range from one to several hundred and can span large genomic regions. Genes that are located in close proximity to the lead SNP are typically reported as being associated with the trait. However, these genes/lead SNPs are not necessarily causal. This is because the genotyped SNPs are tag SNPs for multiple other SNPs in the region due to linkage disequilibrium (LD). The genotyped SNPs therefore act as surrogates for

larger genomic regions containing untested SNPs. It is therefore plausible that the tag SNP is itself not associated with the trait but is associated to a secondary SNP which in turn is associated to a trait (Spain and Barrett, 2015). The results of GWAS alone are therefore not sufficient to assign causality to individual SNPs without further computational analysis and functional follow-up. Additionally, the majority of variants associated with disease are found within non-coding regions of the genome (Spain and Barrett, 2015), so the impact of these variants may not be obvious. It is important to tease apart this structure of association to attempt to identify the SNP(s) specifically associated with a trait at a genomic locus. This process is known as fine-mapping.

1.2.1 GWAS in AS

GWAS have been employed in AS in order to identify the multiple associated variants to the disease. Multiple published studies exist that investigate AS in European cohorts (Ellinghaus et al., 2016, Evans et al., 2011, International Genetics of Ankylosing Spondylitis Consortium et al., 2013, The Australo-Anglo-American Spondyloarthritis Consortium, 2010, Wellcome Trust Case Control Consortium and The Australo-Anglo-American Spondylitis Consortium, 2007). These studies are summarised in Table 1.1 and have identified more than 100 genetic associations (Brown, 2018, Li and Brown, 2017). The use of this hypothesis free technique has led to the discovery of unsuspected pathways and biological mechanisms, including aminopeptidases and the IL-23/IL-17 pathways, which prior to this form of investigation had not been implicated in disease. Also,

comparison across immune-mediated disease hits has contributed to the idea of a shared genetic background amongst diseases.

Most of these associations are yet to be explained biologically but understanding the mechanisms that have led to these associations is of great importance in understanding the causes of this complex disease. Further work is required to fine-map these associations, identify the causative genes and variants driving the association signal, and experimentally show their impact on function and how this contributes to ankylosing spondylitis.

Of the 114 loci identified by GWAS to be associated with AS in Europeans, SNPs in the 17q21 region have been associated with disease in multiple studies (see Table 1.1). This region captures the genes *NPEPPS*, *TBKBP1*, and *TBX21*, all of which have been previously linked as the key gene driving the association at this locus (Evans et al., 2011, International Genetics of Ankylosing Spondylitis Consortium et al., 2013).

This thesis will focus on the 17q21 region and fine-map this locus to attempt to identify the causative gene and variants driving the association signal at 17q21. For the purpose of this report this region will be referred to as the *NPEPPS* region.

Study	Year	Sample size (case/control)	Ethnicity	Number of variants	Number of genome-wide significant loci	Lead SNP at *NPEPPS* loci	Lead SNP p-value
WTCCC & TASC (Wellcome Trust Case Control Consortium and The Australo-Anglo-American Spondylitis Consortium, 2007)	2007	922/1,466	Caucasian	15,333	2	-	-
TASC (The Australo-Anglo-American Spondyloarthritis Consortium, 2010)	2010	2,053/5,140	Caucasian	286,662	6	-	-
Evans et al. (Evans et al., 2011)	2011	3,023/8,779	Caucasian	2,223,620	8	*rs8070463*	$5.3x10^{-8}$
IGAS (International Genetics of Ankylosing Spondylitis Consortium et al., 2013)	2013	10,619/15,145	Caucasian	128,935	24	*rs9901869*	$6.0x10^{-15}$
Ellinghaus et al. (Ellinghaus et al., 2016)	2016	52,262(8,726*)/34,213	Caucasian	130,052	113	*rs12943464*	$2.3x10^{-14}$

Table 1.1: Summary of the main AS GWAS studies performed in Caucasian cohorts with further detail on *NPEPPS* locus. Study is the consortium or lead author of the GWAS study. Sample size ascribes the number of cases and controls in the study. Number of variants lists the number of SNPs genotyped in the experimental procedure. Number of genome-wide significant loci is given ($P<5x10^{-8}$). Where a disease association was identified for the *NPEPPS* region the lead SNP and p-value are given. Adapted from (Costantino et al., 2018).

*AS specific cases in cross inflammatory disease study.

1.3 Gene regulation landscape

Although GWAS results are generally presented as a list of genome-wide significant loci associated with disease and defined by a lead SNP, each of these single variants are generally indicative of a large group of SNPs associated with disease. The majority of GWAS SNPs are found within introns and intergenic regions of the genome (Maurano et al., 2012) with 65% to 72% of disease-associated SNPs identified in AS likely to lie in non-coding regions (Li et al., 2017). These non-coding variants may overlap with functional sequence in these regions which are likely to impact on disease through an effect on the regulation of gene expression. In AS, disease-associated SNPs have been shown to have significant enrichment in immune cell types and gut tissue when restricted to specific markers of regulatory function (Li et al., 2017).

1.3.1 Gene expression and regulation

Gene expression is the process by which DNA is transcribed into messenger RNA (mRNA) and subsequently translated into a functioning protein. This process begins with transcription as RNA polymerase II produces the mRNA molecule from a gene's DNA template. Multiple complex steps are involved in transcription from initiation through elongation to termination. These steps are regulated by transcription factors (TFs), enzymes, and proteins which interact with specific sites of the DNA sequence to carry out the transcriptional process Each of these steps and interactions are points at which the process can be regulated and affect the amount of mRNA produced during transcription.

Gene regulation is the process by which cells control the production of specific gene products to meet the changing requirements of the cell. Most gene regulation is believed to occur at transcriptional initiation (Maston et al., 2006) and is controlled by the binding of multiple factors at specific regulatory elements upstream of the genes transcriptional start site (TSS). These regulatory elements may be proximal or distal to the gene on which they act. Cis-acting regulatory elements generally comprise of a promoter region proximal to the TSS that defines where transcription is initiated and more distal elements such as enhancers, silencers, insulators, or locus control regions that amplify or suppress the transcriptional initiation (Maston et al., 2006). The promoter is a region at the start of the gene that acts as the site of binding for transcriptional machinery and defines the position of the TSS and direction of transcription. The promoter is located proximal to the gene it transcribes. Enhancers are regulatory elements that positively influence transcription at a distal promoter by recruiting TFs and chromatin regulators (Visel et al., 2009). These regulatory regions have a crucial role in controlling the level and timing of gene expression in a cell type specific manner (Zhou et al., 2011a). Silencers are regulatory elements that impact negatively to repress the effects of transcription on the target gene (Maston et al., 2006). Insulators are regulatory elements that act like boundaries and block enhancer activity from affecting neighbouring genes (Maston et al., 2006). Enhancers, silencers, and insulators are less defined by genomic distance to the gene they influence and may act over long distances. For these elements to influence the magnitude of transcription they require direct or indirect communication with the promoter region (Spielmann and Mundlos, 2016). For

direct communication of distal regulatory elements physical contact looping of the DNA is required to bridge together the promoter and regulatory element as a chromatin structure. This is mediated by CCCTC-binding factor (CTCF), a transcription factor, classed as an insulator as it can interfere with the interaction between enhancer and promoter, that binds together strands of DNA to form loops (Ong and Corces, 2014).

These areas of regulatory control contain the DNA sequence that acts as the binding sites of the proteins, enzymes, and transcription factors that make up the molecular machinery for transcription. Gene expression may be upregulated or down regulated at the transcriptional level by any process that has an effect on these regulatory elements. One factor that must be considered is whether the transcriptional molecular machinery can access the DNA at the regulatory element to bind. This is controlled by the structure of the DNA in the nucleus.

1.3.2 Chromatin accessibility

DNA is stored in a complex with proteins, primarily histone, to package the long DNA molecules in a compact form for storage in the nucleus of cells. The DNA forms a complex with a histone octamer as it coils around the protein structure to form a nucleosome. In this compact, dense structure the DNA sequence is not accessible for interactions with proteins and other molecules meaning processes such as transcription for gene expression are restricted. For the transcriptional machinery to bind to DNA to initiate transcription and gene regulation, the DNA-protein complexes of chromatin must open. This dynamic process, the opening

and restriction of the chromatin structure, is associated with modifications of the histone proteins in the complex. The degree to which molecules are able to physically interact with DNA at these areas of open chromatin is known as chromatin accessibility (Klemm et al., 2019). Areas of chromatin accessibility can be used to identify the location of regulatory elements across the genome and how these differ between cell-type and context. Histone modifications are highly informative and in combination with open chromatin can be used to identify the different forms of regulatory element present. Tri-methylation of the 4th lysine of Histone H3 (H3K4me3) is associated with active promoters near a TSS (Heintzman et al., 2009). Whereas, mono-methylation of the 4th lysine of Histone H3 (H3K4me1) is associated with active enhancers (Heintzman et al., 2009). Acetylation at the 27th lysine of histone H3 (H3K27ac) is a histone mark associated with both activate promoters and enhancers (Creyghton et al., 2010). It is also informative to identify sites of CTCF binding in these areas of open chromatin as these can identify the location of promoters and enhancers (Spielmann and Mundlos, 2016, Tak and Farnham, 2015, Ong and Corces, 2014). As gene expression is cell-type and context specific it is tangible that chromatin accessibility is similarly specific and understanding this chromatin landscape across cell-types is valuable in understanding changes in gene expression.

1.3.3 Causal variants in gene regulation

Variants that are located in regulatory elements may have a range of effects that impact on gene regulation but understanding the impact of these non-coding

variants is complex. SNPs that change the DNA sequence within regulatory elements can disrupt the binding sites of existing TFs and transcriptional machinery which can change the affinity of these DNA-protein interactions or can introduce new binding sites leading to the recruitment of new proteins to the region. These changes in DNA-protein, and subsequent protein-protein, interactions can change the process of gene regulation and result in different levels of gene expression depending on genotype. Variants in regulatory regions have the potential to cause up- and down-regulation of the transcription of a gene by altering levels of function at a specific element but also have the ability to disrupt the function of regulatory element altogether resulting in the gain or loss of specific regulatory elements (Spielmann and Mundlos, 2016). An example of causal variants in regulatory regions is demonstrated in the study by Vecellio and colleagues (2016) who investigated the effect of the AS-associated SNP in a putative regulatory region upstream of *RUNX3*. They were able to show that the presence of the risk allele at *rs4648889* was able to decrease TF binding and reduce *RUNX3* expression. A further example from AS was the identification of an enhancer in the *IL23R-IL12RB2* intergenic region where reduced luciferase activity was associated with the presence of the risk allele at *rs11209032*.

Through fine-mapping of disease-associated regions identified through GWAS it is possible to improve on the identification of a lead SNPs for a region and identify the putative causal variants with a role in disease. To truly define a functional effect of a variant experimental approaches will be required. Understanding the expression levels of the gene in the prioritised region will be informative when attempting to interpret non-coding variants prioritised through

fine-mapping. Differential expression between AS cases and healthy controls may further implicate a specific gene in disease. Alternatively specific genomic regions or SNPs could be identified as expression quantitative trait loci (eQTL) (Rockman and Kruglyak, 2006) through the identification of variants that correlate with allele-specific gene expression differences (Tak and Farnham, 2015). These approaches will assist in the fine-mapping strategy of this region.

1.4 Functional genomics

Building on the findings of previous GWAS studies, this project aims to fine-map the disease-associated region surrounding *NPEPPS* with the use of CoreExome genotyping data (full GWAS study manuscript in preparation by the Brown Group). The Illumina CoreExome BeadChip is designed for large-scale genotyping studies offering genome-wide SNP coverage across diverse world populations. This array genotypes over 510,000 marks offering greater genome-wide coverage than the Illumina ImmunoChip array. The CoreExome genotyping panel will provide greater genotyped locations at the *NPEPPS* region providing improved coverage and a higher likelihood of genotyping the causal variant. In this thesis, statistical fine-mapping processes will be used to attempt to prioritise variants, then go on to investigate potential causative roles through functional experiments.

1.4.1 Approaches to fine-mapping

Fine-mapping is used to determine the genetic variant (or variants) responsible for disease given the association of a genomic region with that trait assuming at least one causal variant exists (Schaid et al., 2018). Fine-mapping is a statistical approach to assign probability of causality to candidate SNPs and identify target genes and variants driving GWAS signals. The goal is to identify a minimal set of SNPs that include the causal variant(s). When combined with a better understanding of the functional landscape around an association signal a fine-mapping approach will allow for the narrowing down of the signal and identification of a shortlist of potentially causal variants and target genes.

By using a combination of statistical approaches, this study aims to prioritise a minimal set of SNPs for the *NPEPPS* region. This study will then use functional annotation to prioritise variants to investigate using functional experiments.

1.4.2 Statistical approaches to fine-map disease association

Spain and Barrett (Spain and Barrett, 2015) categorise statistical methods for fine-mapping into two groups: frequentist approaches that prioritise variants based on p-values or LD to identify a lead SNP, and Bayesian approaches that calculate the posterior probability of causality for each SNP.

The frequentist approach is to consider all SNPs with a p-value less than a specific threshold as candidate variants for causality. For genome-wide studies the established p-value threshold for genome-wide significance is $p < 5x10^{-8}$. In

addition to this hierarchical prioritisation, it is also important to consider the LD structure of the region to the lead SNP and to each other in a pairwise manner. This is because LD plays a key role in GWAS, and genotyped SNPs act to tag alternative variants that may have a causative role. By combining the p-value threshold SNP and variants in strong LD the prioritisation method should capture the causal SNP in the set.

The Bayesian approach measures the evidence of association at each SNP by calculating the Bayes Factor (BF). This is used to calculate the posterior probability (PP) of each variant being causal in the region. The PP is the ratio of evidence for each variant versus all others at the region of association and can be used to rank the SNPs in order of their evidence of being causal.

Bayesian methods are used to determine credible sets of SNPs, a minimal set of SNPs with a set probability that it contains all causal SNPs for the region. A 99% credible set is generally calculated, that is the set of SNPs whose cumulative combined PP accounts for 99% of the ratio of evidence for the test region. An example of Bayesian analysis in AS was performed by van de Bunt and colleagues (van de Bunt et al., 2015) who used an approximate Bayesian approach to investigate the twenty six non-MHC loci, including the *NPEPPS* region, identified in a European cohort by the IGAS study (International Genetics of Ankylosing Spondylitis Consortium et al., 2013). This technique generated 95% credible sets with less than 10 SNPs for seven of the twenty-six loci. This demonstrates the ability of the Bayesian method to prioritise variants.

Stepwise conditional analysis is an additional process that can be employed in a fine-mapping strategy. By applying a stepwise approach by conditioning on the variant with the lowest p-value (or largest Bayes Factor) and subsequently controlling for the variant to the model it is possible to identify secondary signals at the associated region. In a region with a set of SNPs with genome-wide significant associations, if following conditional analysis on the lead SNP a variant remains at genome-wide significance it is likely to be independently associated with disease. Conditional analysis was performed in the IGAS study (International Genetics of Ankylosing Spondylitis Consortium et al., 2013) and identify 12 secondary signals at known AS-associated loci in Europeans. Following statistical fine-mapping and the identification of a prioritised variant set it is highly useful to identify if these SNPs overlap with functional elements of the genome.

1.4.3 Functional annotation of the fine-mapping region

Assigning biological function to DNA sequence can inform about the likely role of SNPs that have been prioritised by fine-mapping steps and can further aid in the selection of variants for functional follow up.

Annotation of SNPs can be categorised into two broad categories; those that occur in protein-coding sequence versus those that occur in non-protein-coding sequence. Protein-coding variants occur in exons and a change in genomic sequence may result in either a change in amino acid sequence, non-synonymous mutation, or no change in amino acid sequence, synonymous

mutation. Non-synonymous changes may result in modification to the structure of the protein produced effecting its ability to perform its role. It may lead to alter splice sites or introduce premature stop codons resulting in different proteins isoforms being produced. Or this form of variation may result in no protein being produced at all. Non-coding variants occur in introns and intergenic regions between genes. These variants are likely to have an effect on the regulation of gene expression, but their impact is harder to predict. A variant may be located at a known or putative functional element, e.g. an enhancer, and the allelic variant may act to alter the ability of the region to bind tissue specific and general TFs and/or accessory proteins. This may impact on the ability of the functional element to perform it role and in turn influence gene expression.

Tools that query large databases, such as Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) can be used to categorise variants and predict the effect of protein-coding variants. Non-coding variants are more complex to annotate and predict the impact of the variation. Genetic variation in non-coding regions are often involved in genetic regulation and through annotation may align with functional elements such as promoters, enhancers, silencers, transcription start sites, and transcription factor binding sites. These forms of non-coding regions can be annotated using tools such as RegulomeDB (Boyle et al., 2012), Combined Annotation-Dependent Depletion (CADD) (Rentzsch et al., 2019) and HaploREG (Ward and Kellis, 2012) which query data from large public datasets of functional elements like the Encyclopedia of DNA Elements (ENCODE) Consortium (The ENCODE Project Consortium, 2012) and the Roadmap Epigenomics Project (Bernstein et al., 2010). With the addition of disease-specific

regulatory information, overlapping regions between prioritised variants and regulatory elements can be identified and investigated further through functional experiments.

1.4.4 Functional experiments for prioritised variant investigation

Following statistical fine-mapping and functional annotation, follow up experiments are required to understand the functional consequence of the prioritised variant(s). The functional approach will depend on the variant identified and the evidence for any putative function that has led to its selection. Variants that are located in the coding region of genes require experimental approaches that assess the impact of the variant on the protein produced. The majority of disease-associated SNPs are located in the non-coding regions and are likely to impose regulatory effects on the expression of genes. Alternative techniques are required to investigate the impact of these non-coding variants and how they influence the regulation of genes.

A wide array of investigative approaches have been used to assess the effect of disease-associated variants. In AS, Hanson and colleagues (Hanson et al., 2018) assessed the influence of coding variants in *ERAP1* and *ERAP2* using RNA sequencing (RNA-seq) to assess their impact on isoform-specific expression and mass spectrometry to assess their impact on protein levels. Vecellio and colleagues (Vecellio et al., 2018, Vecellio et al., 2016) have explored *RUNX3* non-coding regulatory polymorphisms associated with AS. They used reverse

transcription quantitative polymerase chain reaction (RT-qPCR) to compare *RUNX3* mRNA levels in AS risk and protective genotypes, followed by luciferase reporter assays to investigate the functional effect of the SNPs, and electrophoretic mobility gel shift assays and chromatin immunoprecipitation to assess their effect on transcription factor binding. In 2017, Lau et al. (Lau et al., 2017) investigated the risk allele, *rs11657579*, located in the 3` untranslated region (UTR) of *TBX21* and whether it alters gene and protein expression in AS patients. They used RT-qPCR to quantify the expression of total *TBX21* mRNA transcripts, followed by flow cytometry to assess the frequency of TBX21-expressing cells. This led on to modelling in mice to examine the role of TBX21 on disease development and progression.

Although significant progress has been made, the exact cause(s) of AS remains unclear. The use of GWAS has identified key disease associated region and novel pathways previously not explored in a disease context. Further work is required to investigate additional associated regions and attempt to establish the biological mechanisms behind their associations to AS. The work in this thesis will contribute valuable evidence to understanding the causes of disease and identifying novel drug targets to expand and improve the available treatment pathways for this debilitating condition.

1.5 Aims and objectives

1.5.1 Project aims

Aim 1: To fine-map the AS disease association signal from CoreExome GWAS data and identify and prioritise the causative genetic variants.

1.1 To analyse the CoreExome GWAS data for the *NPEPPS* locus using frequentist analysis in PLINK, perform conditional analysis, and prioritise the key SNPs driving this signal.

1.2 To analyse the CoreExome GWAS data for the *NPEPPS* locus using Bayesian analysis in SNPTEST, perform conditional analysis, and prioritise the key SNPs driving this signal.

Aim 2: To use epigenomic data generated by the Knight group and from publicly available databases to further annotate the *NPEPPS* locus to identify functional regions of interest that overlap with the key SNPs driving this signal.

Aim 3: To characterise the causative genetic variants through experimental approaches to validate their functional role in disease.

3.1 To use RT-qPCR to assess whether the expression levels of the genes in the *NPEPPS* locus differ between AS patients and healthy controls.

3.2 To investigate the regulatory role of specific regions of the *NPEPPS* locus, identified through computational analysis, using luciferase reporter assays.

2 Materials and methods

2.1 Introduction

To achieve the aims of this study, the process to fine-map the *NPEPPS* region was conducted in four distinct steps (see Figure 2.1). In step 1 statistical approaches were used to perform fine-mapping using the CoreExome genotyping data with the aim of identifying putative causal SNPs. Step 2 builds on the results of step 1, by functionally annotating the region with public and in-house data to prioritise variants from the identified SNPs. These two steps were performed computationally with experimental approaches performed in steps 3 and 4 with the aim of functionally validating a role in disease for the prioritised variants. In step 3, RT-qPCR was used to identify differential gene expression of proximal genes in the prioritised region. This was followed by step 4, which made use of luciferase dual-reporter assays to see if the prioritised regions exhibit enhancer activity. Further details of the methods used in each of these steps are provided in this chapter.

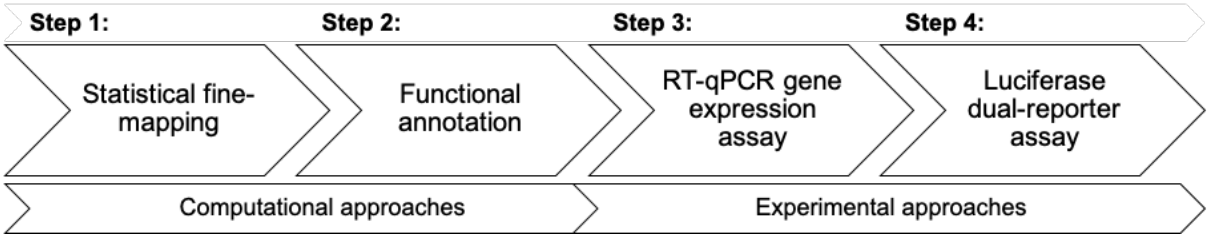


Figure 2.1: Flow chart of steps performed in *NPEPPS* region fine-mapping study. Step 1 and Step 2 are the computational steps performed to identify and prioritise putative causative genetic variants. Step 3 and Step 4 are the experimental assays performed to investigate the functional role of the variants identified in Steps 1 and 2.

2.2 Statistical fine-mapping

2.2.1 Genotyping data

Genotyping data had previously been generated using the Illumina CoreExome SNP genotyping microarray (Illumina) at the Translational Research Institute (TRI) by Brown and colleagues (currently unpublished). Prior to data access for this project, data was processed and normalised for each sample, and genotypes called using Illumina Genome Studio software. SNPs were excluded if the quality score <0.15, and samples were excluded if genotyping success <98%. Standard quality control steps were performed to identify and exclude samples with cryptic relatedness, outlying heterozygosity rate (>0.37 or <0.32) or excessive missingness (>5%). SNPs were identified and excluded with Hardy-Weinberg equilibrium P-value $<10^{-7}$ or with a minor allele frequency (MAF) <1%. The dataset was then imputed with the Haplotype Reference Consortium (HRC) reference data (McCarthy et al., 2016) using the Sanger imputation server and loci with quality score <0.6 were excluded from further testing.

A subset of this data was generated by Dr Zhixiu Li to be used for this research project. This subset only included samples of European descent and was specific to the 1 Mb region around the *NPEPPS* gene: Chr17:45,000,000-46,000,000. This window fully captured the disease associated signal at this region and its boundaries were manually defined. This region-specific dataset contained genotyping data from 8,244 AS patients and 14,542 healthy controls (total cohort = 22,786) covered by 246 genotyped SNPs and 4,834 imputed SNPs (total SNP coverage = 5,080). This dataset was analysed to fine-map the region around *NPEPPS*.

2.2.2 Frequentist analysis

Genome-wide association analysis was performed on the *NPEPPS* region data using the command line version of PLINK 1.9 (Chang et al., 2015). A logistic regression model was used with command `--logistic` in a case-control test to compare the allele frequencies of samples from patients with AS to samples from healthy volunteers. A genotype threshold of 0.05 was set using the command `--geno 0.05`, with a Hardy-Weinberg exact threshold of $1x10^{-6}$ using the command `--hwe 1e-6`, and 2 covariates of PC1 and PC2 loaded. A total of 8067 variants were analysed with a total genotyping rate of 0.999967. No variants were removed due to missing genotype data or due to Hardy-Weinberg exact test, other than those already removed as part of the initial quality control (QC) process. The calculated p-value was used to indicate the significance of difference in frequency of the SNP tested. Following frequentist analysis, SNPs were ranked by -log(p-value) and grouped by those that reached genome-wide significance (p-value < $5x10^{-8}$) and those that reached suggestive significance (p-value < $5x10^{-5}$).

Conditional analysis was performed using PLINK 1.9. Each SNP in the genome wide significant set was individually added to the logistic regression analysis as a covariate and controlled for using command `--condition rsID`. The residual signal was re-assessed to identify any SNPs that remained significant following conditional analysis.

Data visualisation was generated in R (Chang et al., 2015, Team, 2014) using packages including; dplyr, readr, ggplot2, and personally developed scripts.

2.2.3 Bayesian analysis

Bayesian analysis was performed on the *NPEPPS* region data in SNPTEST 2.5.4 (Marchini et al., 2007). Input files were re-formatted from the PLINK input files using R. The Bayesian case-control test method was used with the additive model and inbuilt priors to generate a Bayes Factor for each SNP using command `-bayesian 1 -method score`. The 2 PCA covariates were included in the model as for frequentist analysis. The posterior probability was calculated for each SNP as described in WTCCC 2012 (Wellcome Trust Case Control Consortium et al., 2012) (see below equation) and ranked by their contribution into a 99% and 95% credible set.

$$Posterior\ probability\ (PP) = BF / \sum (all\ BF\ for\ region)$$

Conditional analysis was performed in SNPTEST 2.5.4 controlling for each of the genome-wide significant SNPs identified by frequentist analysis using the addition of command `-condition_on rsID` to the analysis. The residual signal was re-assessed to identify any SNPs that remained significant following conditional analysis.

2.3 Functional Annotation

2.3.1 Regional annotation with available datasets

Following frequentist and Bayesian analysis, the prioritised set of 96 SNPs were analysed using publicly available databases and tools as well as in-house AS-specific data. This approach identified overlaps and proximity between SNPs and known or putative functional elements. Datasets were aligned and visualised using the Washington University in St. Louis (WashU) epigenome legacy browser (Zhou et al., 2011b). Bedtools (Quinlan and Hall, 2010) intersects function was used where coordinate data was available to identify specific overlaps between datasets.

2.3.2 Publicly available datasets

Several online tools were used to annotate the prioritised SNPs with functional information, and where possible rank the SNPs based on in-built scoring systems. Ensembl VEP (McLaren et al., 2016) was initially used to categorise the SNPs as coding or non-coding, identify any overlaps with known gene transcripts or intergenic regions, and give an initial consequence if a coding variant was identified. As the majority of SNPs were classed as non-coding, epigenomic tools were subsequently used to identify overlaps with functional elements and rank the SNPs based on predicted impact. The tools used were: HaploREG from the Broad Institute and MIT (Ward and Kellis, 2012) and RegulomeDB from Stanford University (Boyle et al., 2012). These tools annotate SNPs with known and predicted regulatory elements in non-coding region from public datasets providing information on regions of LD from the 1000 Genomes project, open chromatin binding sites of transcription factors from

Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects, and the effect of SNPs on expression from eQTL studies.

Once the prioritised SNPs had been annotated, the functional genomic landscape of the *NPEPPS* region was visualised on the WashU epigenome browser to identify other proximal functional elements in the region and confirm overlaps between SNPs and regulatory regions. The following datasets were used: NIH Epigenomic Roadmap cell-type specific chromHMM.

2.3.3 Ankylosing spondylitis specific datasets

Epigenomics datasets were available on 17 AS patients and 44 healthy controls from previous work performed by the Knight group (Wellcome Centre for Human Genetics (WCHG)). AS patient samples were obtained from the Genetics of gene expression in immune-related disease (GENExpress ID) study cohort. Patients gave written informed consent at the time of collection based on the GENExpress ID study criteria (National Research Ethics Service Committee (REC) South Central Reference Number: 12/SC/0063). Healthy volunteers were recruited from the Oxford Biobank (REC South Central Reference Number: 06/Q1605/55). All patients and healthy volunteers gave written informed consent at the time of collection. All epigenomic data were generated as part of on-going studies performed by the Knight group.

Individual cell-type specific sample data from AS patients and healthy controls was available for assay for transposase-accessible chromatin using sequencing (ATAC-seq) experiments identifying open chromatin, chromatin immunoprecipitation

sequencing (ChIP-seq for H3K27ac and H3K4me3 identifying specific histone marks, and *NPEPPS* promoter specific and *TBKBP1* promoter specific Capture-C sequencing for CD4, CD8, and CD14 cell populations for each individual sample. Tracks for these experiments were visualised using the WashU epigenome browser.

Master peak files, a coordinate file of well-defined peaks, were available for cell-type specific (CD4, CD8, CD14) open chromatin from ATAC-seq and H3K27ac marks from ChIP-seq. This master peak file was generated where non overlapping 500bp peaks were present in at least 30% of the samples included regardless of disease state. Overlaps between these files and prioritised SNPs informed on regions of interest for functional follow-up.

2.4 RT-qPCR gene expression assay

2.4.1 Samples

140 patients with a diagnosis of AS according to the modified New York criteria and 74 healthy controls were included in the gene expression experiment from the biobank of the Brown group (TRI, Brisbane). See Table 2.1 for summary cohort data. Human ethical approval was granted by the Princess Alexandra Hospital and University of Queensland Ethics Committees (ethics no. Metro South HREC/05/QPAH/221 and UQ 2006000102), and all subjects provided written informed consent.

Peripheral venous blood was previously collected by members of the Brown group from AS patients at the Princess Alexandra Hospital Ankylosing Spondylitis Clinic, and PBMCs isolated using density gradient centrifugation over Ficoll-Paque Plus (GE Healthcare) and stored in liquid nitrogen prior to downstream processing.

	AS Cases	Healthy Controls
Count	140	74
Male (%)	121 (86)	25 (34)
*HLA-B*27* positive (%)	124 (89)	-
Anti-TNF treatment (%)	20 (14)	-

Table 2.1: Summary cohort data of AS patients and healthy controls used in RT-qPCR gene expression assay.

2.4.2 RNA extraction

RNA was extracted from peripheral blood mononuclear cells (PBMCs) of AS patients and healthy volunteer samples. All PBMC samples were stored in liquid nitrogen prior to extraction. RNA for AS patient samples was previously extracted in the Brown group and stored at -80°C. PBMC samples for healthy controls were removed from liquid nitrogen onto dry ice. The cell pellet was quickly thawed and transferred to 10 mL falcon tube. PBMCs were resuspended in 8 mL RPMI+FBS and centrifuged at 400 x *g* for 5 min. The supernatant was discarded without disturbing the cell pellet and then resuspended in 10 mL PBS. 2 million cells were resuspended in RLT Plus buffer (Qiagen) for RNA extraction. Cells were homogenised by vortexing the sample on max speed for 1 min and stored at -80°C until extraction. The remaining cells were stored at -80°C in relevant buffers for future RNA/DNA extraction. Samples were extracted on the QIAsymphony robot (Qiagen) following the manufacturers protocol for RNA extraction (Programme: RNA_CT_400) using an input volume of 400 µL and elution volume of 50 µL. Following extraction, RNA was quantified, and material quality assessed by NanoDrop (ThermoFisher Scientific).

2.4.3 cDNA synthesis

cDNA synthesis was performed using the Bioline Tetro cDNA Synthesis kit (BIO-65043, Bioline) and oligo (dT) primers. For each sample a 20 µL reaction was performed following the manufacturers protocol, with an input mass of 1 µg RNA. If RNA material was limited, reactions were prepared with 500 ng RNA. The reaction mix was prepared on ice and run on the recommended thermal cycler programme; 45°C for 30 mins, 85°C for 5 mins, 4°C until removal from machine. Following cDNA synthesis, samples were stored at -20°C.

2.4.4 RT-qPCR

Following sample preparation, a total of 138 AS cases and 67 healthy controls were available as input material for the RT-qPCR assay. Primers were designed manually from transcript alignments and checked using primer-BLAST (Ye et al., 2012) to target all protein coding transcripts (n = 6) of *NPEPPS* (4 primer pairs) and *TBKBP1* as well as the reference gene *HPRT1*. In-house primers for reference genes *GAPDH* and *RPL32* were also used. Specific attention was given to the design of the primers for *NPEPPS* as this gene has 25 reported transcripts of varying type, length, exon combinations, and supporting evidence. Of the 25 predicted transcripts: 6 are protein coding, 6 undergo nonsense mediated decay, 4 are processed transcripts, and 9 undergo alternative splicing and contain intronic sequence. To capture the expression signal from all *NPEPPS* transcripts a combination of multiple primers was required as no consensus region is shared across all transcripts. By targeting exons that are conserved across multiple transcripts, a panel of 4 primers were designed to capture all six of the predicted protein coding *NPEPPS* transcripts as well as the

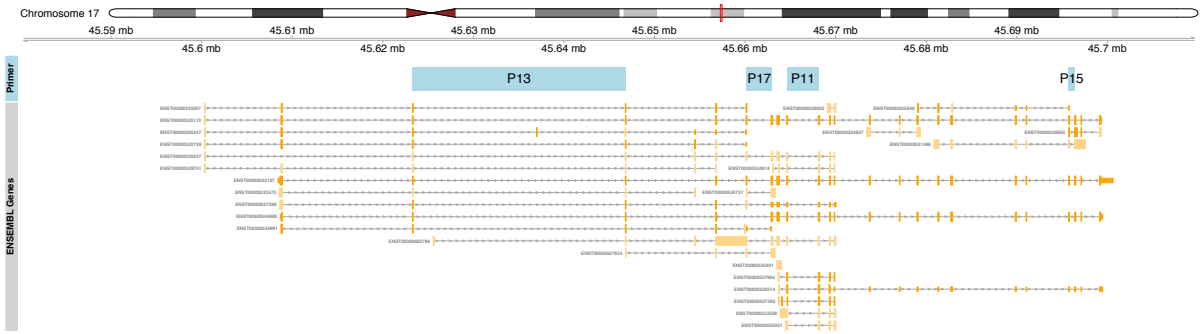
most abundant transcript observed in in-house RNA-seq data. This combination of primers is predicted to capture signal from 20 of the 25 reported transcripts (see Supplementary Table 2.1 for transcripts captured by specific primers). *TBKBP1* and the three reference genes had fewer reported transcripts and therefore a single primer pair for each gene was used to capture all transcripts. Details of primers are given in Table 2.2 and the binding locations for *NPEPPS* and *TBKBP1* shown in Figure 2.2(A) and (B) respectively. All primer pairs were designed to generate a product that spanned an intron to minimise genomic DNA contamination. Primers were tested and the RT-qPCR optimised to ensure all primer pairs had an efficiency of 90-110%.

RT-qPCR reactions were set up in triplicate for each sample for each of the 8 primer pairs using Applied Biosystem SYBR Green PCR Mastermix (Ref# 1811520, Applied Biosystem). SYBR Green reagent from a single manufacturer lot number (1811520) was used across all 16 reaction plates. Reactions were set up in a total volume of 10.0 µL, consisting of 5.0 µL of 2x SYBR Green PCR Mastermix, 0.5 µL primer mix (5mM forward primer and 5mM reverse primer), 3.5 µL deionised water, and 1.0 µL cDNA. Reactions were run in 384-well plates and AS samples and healthy control samples were mixed across all plates with additional control samples run on each plate. Reaction plates were all run on the same QuantStudio 5 instrument. The RT-qPCR reaction conditions were as follows: an initial step of 50°C for 2 mins and then 95°C for 10 mins to denature the cDNA, followed by 40 cycles of 95°C for 30 secs, annealing at 60°C for 30 secs, and elongation at 72°C, and then a melt-curve programme decreasing from 95°C to 45°C.

Target Gene	Primer Name	Primer Sequence	BP	Tm	GC%	Products length	Exon location
NPEPPS	NPEPPS-P11_F	TTGCTCTGGTTGTGGGACATG	21	60.8	52.4	134	10
NPEPPS	NPEPPS-P11_R	GAAGCAGTGGTCTACACACAG	21	58.6	52.4		11
NPEPPS	NPEPPS-P13_F	GTGAGGCAGGCGACTAATCAG	21	60.8	57.1	152	3
NPEPPS	NPEPPS-P13_R	GTACTAGGGAAAGACAAGGTGAC	23	58.2	47.8		4
NPEPPS	NPEPPS-P15_F	ACAAGCAGATATGCAAGAAGAGA	23	57.9	39.1	149	21
NPEPPS	NPEPPS-P15_R	CAGCTACTCCACCAATTACCGA	22	59.8	50.0		22
NPEPPS	NPEPPS-P17_F	TGTTCCTAAAGACAGAGTAGCT	22	56.2	40.9	134	6
NPEPPS	NPEPPS-P17_R	CAACAAATGCCACCAGATATGT	22	57.3	40.9		7
TBKBP1	TBKBP1-P1 _F	GCCTCAAAGTCTACGAGATCA	21	57.2	47.6	193	1
TBKBP1	TBKBP1-P1 _R	TGGATCATCTCTCCAAGCTGT	21	58.5	47.6		3
GAPDH	GAPDH-P1_F	TGCACCACCAACTGCTTA	18	54.5	50.0	177	7
GAPDH	GAPDH-P1_R	GGATGCAGGGATGATGTTC	19	53.3	52.6		8
HPRT1	HPRT1-P1_F	GACACTGGCAAAACAATG	18	49.5	44.4	125	6
HPRT1	HPRT1-P1_R	ACAAAGTCTGGCTTATATCC	20	49.4	40.0		7
RPL32	RPL32-P1_F	GATTCAAGGGCCAGATCTTGA	21	57.7	47.6	174	2
RPL32	RPL32-P1_R	TGGAGGAAACATTGTGAGCGA	21	59.9	47.6		3

Table 2.2: Summary of primers used in RT-qPCR reaction. Target gene for primer pair identified. Primer name identifies the individual primer. BP ascribes the number of base pairs of the primer. Tm ascribes the predicted melting temperature of the primer (℃). GC% ascribes the percentage of guanine and cytosine base pairs in the primer sequence. Product length ascribes the base pair length of the product generated by the primer pair. Exon location ascribes the exon in which the primer is located.

(A)



(B)

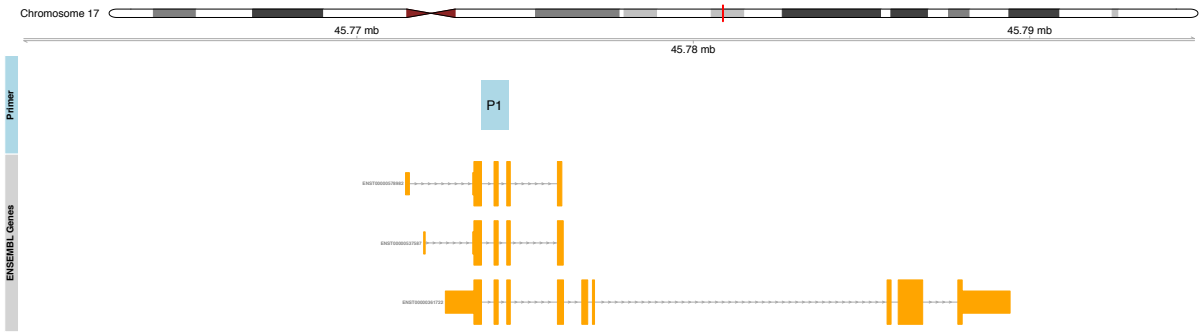


Figure 2.2: Schematic of *NPEPPS* and *TBKBP1* RT-qPCR primer locations. (A) Location of *NPEPPS* primer pairs (P13, P17, P11, P15) in relation to know transcripts of the genes. (B) Location of *TBKBP1* primer pair (P1) in relation to know transcripts of the gene.

2.4.5 Data cleaning

Initial quality control steps were performed in QuantStudio Design and Analyse software (Applied Biosystems) to calculate C_t for each reaction and remove any outliers from forward analysis. Samples that failed to amplify, displayed unusual amplification plots, and/or displayed unusual melt curves were removed from the dataset. Following data export from QuantStudio, data was imported into R and cleaned to remove anomalous results. The acceptable variance for each primer set was calculated as +/- 2 X standard deviation from the corresponding mean observed in the cell-line sample for that primer run across all reaction plates. This acceptable range was applied to the raw C_t values as a percentage of overall signal observed. As a result, 13 AS patient samples and one healthy control sample were removed.

2.4.6 Gene expression analysis

Gene expression analysis was performed using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001) on the software R unless otherwise stated. Data was normalised to the geometric mean of two reference genes (*RPL32* and *HPRT1*) for each reaction through equation 1. Once ΔC_t was calculated for all reactions, the mean ΔC_t for all AS cases was compared to the mean ΔC_t for all healthy controls for each primer set by a one-way ANOVA.

$$\Delta C_t = mean(test\ C_t) - mean((RPL32\ C_t) + (HPRT1\ C_t)) \qquad (1)$$

Using the mean ΔC_t for AS cases and healthy controls, the $\Delta\Delta C_t$ was calculated for each primer pair by normalising to the calibrator sample (healthy control mean) using equation 2.

$$\Delta\Delta C_t = mean(AS\ case\ \Delta C_t) - mean(Healthy\ control\ \Delta C_t) \quad (2)$$

$\Delta\Delta C_t$ was then used to calculate the fold expression change for AS cases compared to healthy controls using equation 3.

$$Fold\ expression\ change = 2^{-(\Delta\Delta Ct)} \quad (3)$$

2.5 Luciferase dual-reporter assay

A luciferase reporter assay was used to test whether regions at the *NPEPPS* region identified through *in silico* prioritisation exhibited enhancer activity.

2.5.1 Samples

Genomic DNA from AS patient samples was obtained from the GENExpress ID study cohort. Patient blood samples had been previously collected at the Nuffield Orthopaedic Centre, Oxford University Hospitals NHS Trust. DNA had been extracted previously from CD14 cells and stored at -20°C. The two patient samples were selected based on their genotype at selected SNPs in the *NPEPPS* region. AS patient AS050 was homozygous for the risk allele at these SNPs, whereas AS051 was homozygous for the protective allele at these SNPs.

2.5.2 Generation of inserts

PCR was used to amplify regions of interest from the AS patient genomic DNA. Primer sequences are given in Table 2.3. A tag-sequence containing a restriction enzyme was added to the end of each primer to facilitate downstream cloning into the reporter vector in an orientation-specific manner.

Invitrogen Platinum SuperFi polymerase (Invitrogen) was used in a 25 µL amplification reaction. 50 ng of genomic DNA was amplified in a reaction mix containing 12.5 µL 2x Platinum SuperFi Mastermix, 0.625 µL 10 mM forward primer, 0.625 µL 10 mM reverse primer and deionised water to the total reaction volume.

The amplification reaction programme was as follows: an initial denaturing step of 30 secs at 98°C, then 35 cycles of 10 secs at 98°C, annealing for 10 secs at 68°C, and elongation of 1 min 45 secs at 72°C followed by a final elongation step of 5 mins at 72°C and hold at 10°C. PCR products were analysed by gel electrophoresis to confirm the presence of a single clean band. PCR products were purified by gel extraction using MinElute Gel Extraction Kit (Qiagen).

2.5.3 Sub-cloning into TOPO vector

Insert DNA was ligated into the pCR-Blunt II-TOPO vector (Invitrogen) following the manufacturers protocol. 4 µL PCR product was combined with 1 µL salt solution and 1 µL pCR-Blunt II-TOPO vector in a 6 µL reaction. The reaction mix was incubated at room temperature for 30 mins and then stored on ice until subcloning into competent cells was performed. DH5 Competent cells (Invitrogen #18265-017) were thawed on ice and 5 µL of TOPO-insert ligation reaction was added to 50 µL competent cells. This mixture was left to incubated on ice for 30 mins and then heat shocked at 42°C water bath for 20 secs. The reaction was then placed on ice for 2 mins before 950 µL pre-warmed LB broth was added and the reaction incubated at 37°C for 1 hr at 225 rpm. 50 µL and 200 µL of each culture was spread onto LB agar plates containing kanamycin (50 µg/mL) and incubated overnight at 37°C for single colonies to form. Overnight 3 mL cultures in LB + kanamycin were grown from single colonies and plasmid DNA extracted by plasmid miniprep (Qiagen QIAPrep miniprep). Restriction enzyme digests were performed with *NcoI*-HF (NEB #R3193) to ascertain whether inserts were present. Three constructs for each region of interest were validated by Sanger sequencing with primers designed at ~500 bp intervals along the region of interest.

Target	Primer name	Primer Sequence	Restriction enzyme tag	BP	Tm	GC%	Product Length
Region 1	NPEPPS-enhancer_F2_BamHI	GGATCCAGGGAAGCTGTGCGAATG	*Bam*HI (GGATCC)	24	71.8	58.0	3545
Region 1	NPEPPS-enhancer_F2_SalI	GTCGACAGGGAAGCTGTGCGAATG	*Sal*I (GTCGAC)	24	71.2	58.0	
Region 1	NPEPPS-enhancer_R2_BamHI	GGATCCTTGGGATTGCAGCCCTG	*Bam*HI	23	71.7	61.0	
Region 1	NPEPPS-enhancer_R2_SalI	GTCGACTTGGGATTGCAGCCCTG	*Sal*I	23	71.1	61.0	
Region 2	NPEPPS-KPNB1_F2_BamHI	GGATCCCATTGGTGCCTGCTTTA	*Bam*HI	24	68.2	50.0	3003
Region 2	NPEPPS-KPNB1_F2_SalI	GTCGACCATTGGTGCCTGCTTTA	*Sal*I	24	67.4	50.0	
Region 2	NPEPPS-KPNB1_R1_BamHI	GGATCCATATTTTCAGGGCCAAGT	*Bam*HI	24	65.3	46.0	
Region 2	NPEPPS-KPNB1_R1_SalI	GTCGACATATTTTCAGGGCCAAGT	*Sal*I	24	64.5	46.0	
Region 3	TBKBP1-Promoter_F2_NheI	GCTAGCAACGAAGAGCAGGTCTAATG	*Nhe*I (GCTAGC)	26	65.6	50.0	2676
Region 3	TBKBP1-Promoter_F2_XhoI	CTCGAGCAACGAAGAGCAGGTCTAATG	*Xho*I (CTCGAG)	27	69.1	52.0	
Region 3	TBKBP1-Promoter_R2_NheI	GCTAGCCGTTCCTTGATGTCTCC	*Nhe*I	23	65.5	57.0	
Region 3	TBKBP1-Promoter_R2_XhoI	CTCGAGCCGTTCCTTGATGTCTCC	*Xho*I	24	69.4	56.0	

Table 2.3: Primer sequences used to amplify regions of interest for luciferase reporter assay experiment. Target identifies the name of putative regulatory element target region for the primer. Primer name ascribes the unique identifier of the primer. Restriction enzyme tag identifies the restriction enzyme sequence designed into the primer sequence. BP ascribes the number of base pairs of the primer. Tm ascribes the predicted melting temperature of the primer (℃). GC% ascribes the percentage of guanine and cytosine base pairs in the primer sequence. Product length ascribes the base pair length of the product generated by the primer pair.

2.5.4 Ligation into pGL4.23 vector

In the final cloning step, *NPEPPS* regions of interest were excised from pCR-Blunt II-TOPO and ligated into pGL4.23 [luc2/minP] (Promega). This vector (hereafter referred to as pGL4.23) contains a firefly luciferase gene under the control of a minimal promoter. Test regions were cloned into the *Bam*HI/*Sal*I restriction site (Figure 2.3 shows a map of the vector).

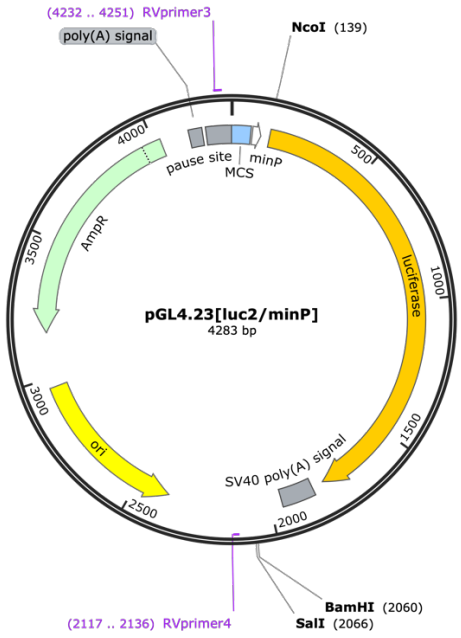


Figure 2.3: pGL4.23 [luc2/minP] vector map showing *Bam*HI and *Sal*I restriction enzyme digest sites in relation to luciferase gene location. Vector visualised using Snapgene software.

To ligate the DNA of interest into the pGL4.23 vector backbone, pCR-Blunt II-TOPO vector containing the correct insert and empty pGL4.23 were linearised by sequential restriction enzyme digest with *Bam*HI-HF and *Sal*I-HF (NEB #R3136 and #R3138 respectively). A 50 µL reaction was prepared for each insert and the pGL4 vector following the manufacturers protocol. 1 µg of DNA was digested in a reaction mix containing 5 µL 10x CutSmart Buffer, 1 µL *Sal*I-HF and deionised

water to 50µL. The reaction mix was incubated at 37°C for 1 hour followed by 20mins at 65°C to head inactivate the enzyme. As *Bam*HI is optimal in the same 10x CutSmart buffer, 1 µL of *Bam*HI-HF was added to the reaction mix and incubated at 37°C overnight to perform the second restriction enzyme digestion.

Following the sequential restriction enzyme digestion, the digested DNA was run on a 0.6% agarose gel to separate out the different sized products. The relevant linearised insert and pGL4 products were extracted from the agarose gel using the Qiagen QIAprep gel extraction kit and product was quantified using the Nanodrop.

The DNA insert was ligated into the pGL4 backbone using NEB T4 DNA ligase. Linearised Insert DNA and pGL4 vector DNA were combined at a 3:1 molar ratio calculated from the DNA size using the NEBioCalculator (v1.10.1; https://nebiocalculator.neb.com/). The ligation reaction was performed in a 20µL total volume containing 2 µL 10x T4 DNA ligase buffer, 1µL T4 DNA ligase and deionised water to total volume. The reaction mixture was incubated at room temperature for 2 hrs and then heat inactivated at 65°C for 10 mins. The pGL4.23 constructs were then transformed into DH5 Competent cells as above and 50 µL and 200 µL of each culture was spread onto LB agar plates containing ampicillin (10 µg/mL) and incubated overnight at 37°C for single colonies to form. Overnight 3 mL cultures in LB + ampicillin were grown from single colonies and plasmid DNA extracted by plasmid miniprep (Qiagen QIAPrep miniprep). Restriction enzyme digests were performed with *NcoI*-HF (NEB #R3193) to ascertain whether inserts were present. If insert was present, a 40 mL overnight

culture in LB + ampicillin was grown and a plasmid midiprep (Invitrogen PureLink HiPure Plasmid DNA Purification Kit) performed to extract high quality plasmid DNA.

Size-matched control pGL4.23 constructs had previously been made in the Knight lab using λDNA inserted at the *Bam*HI site in pGL4.23. Three size-matched controls were used in the luciferase assay: pGL4.23+1.6kb λDNA, pGL4.23+2.4kb λDNA, and pGL4.23+ 4.0kb λDNA.

2.5.5 Transfection into HEK-293T cells

HEK-293T cells were maintained in growth media containing DMEM, 10% FBS, pen-strep, L-Glu and grown to 70-90% confluency in a T75 culture plate. Cultured cells were harvested, diluted to $4x10^5$ cells/mL in growth media, and then 100 µL seeded into each well of a Corning 96-well flat, clear bottomed white tissue culture treated microplate (#3610) ($4x10^4$ cells/well). The seeded plate was placed in 37°C incubator for 24 hours to reach 70-90% confluency. pGL4.23 constructs were diluted to 100 ng/µL in sterile water for subsequent transfection. pRL-TK control plasmid was diluted to 10 ng/µL in sterile water. 200 ng plasmid DNA and 20 ng pRL-TK were co-transfected into the HEK-293T cells in triplicate using lipofectamine 3000 reagent (Invitrogen #100022234) in a total volume of 75 µL as per the manufacturer's instructions. The cells were returned to the incubator at 37°C for 48 hours.

2.5.6 Luciferase assay

Transfected HEK-293T cells were equilibrated to room temperature for use in the Dual-Glo luciferase reporter assay (Promega #E2920). 75 µL Dual-Glo Luciferase reagent was added to each reaction well and the sample gently mixed by pipetting. The plate was incubated for 12 mins 30 secs before being read on the Clariostar microplate reader (BMG Labtech) to detect firefly luminescence. Subsequently, 75µL Dual-Glo Stop & Glo reagent was added and the reaction plate incubated at room temperature for 12 mins 30 secs before being read to detect Renilla luminescence.

2.5.7 Data analysis of luciferase assay data

Firefly luminescence readings were normalised to the corresponding Renilla luminescence reading for each well using equation 4.

$$Relative\ Luciferase\ Units\ (RLU) = \frac{Firefly\ Luciferase}{Renilla\ Luciferase} \quad (4)$$

The average ratio was then calculated across the triplicate reactions for each pGL4.23 construct. For each pGL4.23 construct and controls, the mean relative luciferase units were calculated across all replicate reactions. All results were then normalised to the pGL4.23 vector results (equation 5) to compare the expression across all pGL4.23 constructs.

$$RLU\ normalised\ to\ pGL4.23 = \frac{mean\ RLU\ pGL4.23\ construct}{mean\ RLU\ pGL4.23\ control} \quad (5)$$

Following the initial analysis steps, it was identified that an appropriate size-matched control should be used to calculate the fold-change. All fold-change results are reported in relation to the size-matched control pGL4.23+2.4kb λDNA (equation 6). A paired t-test was performed to assess whether the fold change observed was significantly different.

$$Fold\ change = \frac{mean\ RLU\ pGL4.23\ construct}{mean\ RLU\ pGL4.23\ (2.4kb\ \lambda DNA)\ control} \quad (6)$$

3 Computational approaches for fine-mapping the *NPEPPS* region

3.1 Introduction

3.1.1 Association of the *NPEPPS* region in previous GWAS studies

The association of AS with a region at chromosome 17 overlapping *NPEPPS-TBKBP1* was first identified by GWAS studies in 2011 by The Australo-Anglo-American Spondyloarthritis Consortium (TASC) and the Wellcome Trust Case Control Consortium 2 (WTCCC2) (Evans et al., 2011). In this combined study there was strong evidence of association of SNPs in the chromosome 17q21 region in the WTCCC2 sample set, the TASC sample set, and the replication sample set with a combined p-value of 5.3×10^{-8} for the lead SNP *rs8070463* (Odds ratio (OR) = 1.10, risk allele = C, WTCCC2 P = 0.0025, TASC P = 1.6×10^{-5}, replication P = 0.015). *TBKBP1* was identified as the putative gene for this association signal (Evans et al., 2011). This was the first discovery of an association signal at this region and the first link to a potential role in disease pathogenesis.

This finding was subsequently confirmed in 2013 by the International Genetics of Ankylosing Spondylitis Consortium (IGAS) and the association of lead SNP

rs9901869 at genome-wide significance in a European cohort ($P = 2.3x10^{-12}$, OR = 1.15, risk allele = G), suggestive significance in the East Asian cohort (P = 0.002, OR = 1.18), and genome-wide significant association in the combined analysis ($P = 6.0x10^{-15}$, OR = 1.14) (International Genetics of Ankylosing Spondylitis Consortium et al., 2013). In this study there was uncertainty over the gene contributing to the association signal and the association was attributed to *NPEPPS-TBKBP1-TBX21*. This study was the first to draw the link between the role of *NPEPPS* as an aminopeptidase with a role in peptide processing prior to HLA class I presentation and the *ERAP1* and *ERAP2* associations. The IGAS study also attributed SNPs for *NPEPPS* and *TBX21* to be independently associated with AS, with *rs9901869* tagging the *NPEPPS* signal and *rs11657479* tagging the *TBX21* signal following conditional analysis.

Ellinghaus and colleagues (Ellinghaus et al., 2016) confirmed this disease associate in their cross disease analysis of five chronic inflammatory diseases (AS, Crohn's disease, psoriasis, primary cholangitis and ulcerative colitis). This region on chromosome 17q21 was specifically associated with AS with the lead SNP *rs12943464* at genome-wide significance ($p = 2.3x10^{-14}$, OR = 1.15, risk allele = A) and the signal attributed to *NPEPPS* (Ellinghaus et al., 2016).

The region-specific association plots from these three studies are shown in Figure 3.1. In the 2011 study, genotyping was performed on the Illumina Human660W-Quad array, a custom chip designed by WTCCC2 consisting of the Human550 array plus a set of ~6,000 common copy number variants (Evans et al., 2011). This array gave genome-wide coverage. The 2013 and 2016 studies

were both performed on the Illumina Immunochip genotyping array, a targeted high-density array with focused coverage of SNPs within 186 autoimmune disease-associated loci (Ellinghaus et al., 2016, International Genetics of Ankylosing Spondylitis Consortium et al., 2013). This array does not offer genome-wide coverage, instead offering dense SNP coverage of previously described autoimmune associated loci (Cortes and Brown, 2011). These GWAS offered different levels of genotyping coverage for the *NPEPPS* region (as seen in Figure 3.1), with the recent Immunochip based approaches targeting few SNPs across the region and insufficient data for fine-mapping.

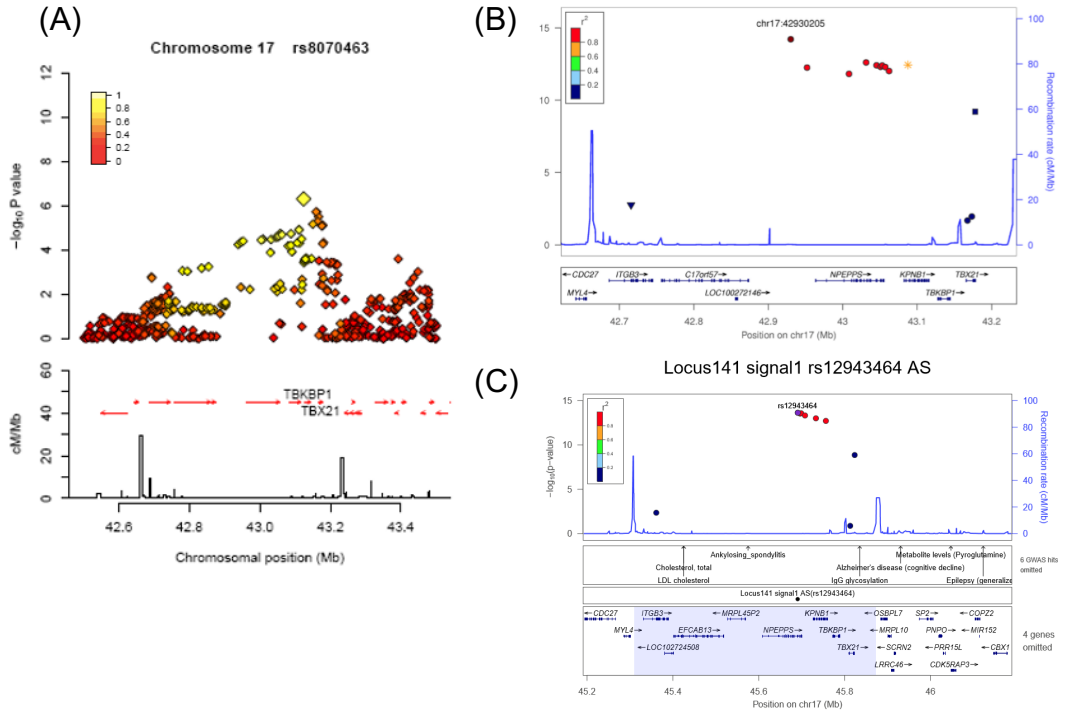


Figure 3.1: Association plots of *NPEPPS* region from previous studies. (A) TASC & WTCCC2 2011 study (lead SNP = *rs8070463*, p = 5.3×10^{-8}), (B) IGAS 2013 study (lead SNP = *rs9901869*, p = 6.0×10^{-15}), (C) Ellinghaus 2016 study (lead SNP = *rs8070463*, p = 2.3×10^{-14}).

3.1.2 CoreExome GWAS cohort and dataset

As described in Chapter 2, the genotyping dataset used in this project was generated on the Illumina CoreExome array and imputed with the HRC reference data using Sanger imputation server by the Brown group (manuscript in preparation). For the purpose of this project, the cohort was restricted to the European individuals to generate a subset of the large dataset to work with. This consisted of 8,244 AS patient samples and 14,542 healthy controls.

The full CoreExome dataset was truncated to the region of chromosome 17 that spanned 1 Mb and overlapped the disease association signal at this region. This region was from chromosome 17 position 45,000,000 to 46,000,000, based on the GRCh37 (hg19) coordinate systems, and captures a section of the genome that overlaps 17 protein coding genes. This 1 Mb window contained data from 5080 SNPs (246 genotyped and 4834 imputed).

3.1.3 Fine-mapping strategy

Fine-mapping is defined in Schaid et al. (2018) as “to refine the genomic localization of causal variants by the use of statistical, bioinformatic or functional methods”. It is the process taken to attempt to identify the causative genetic variant (or variants) for a disease. It builds on the results from GWAS, the identification of SNPs significantly associated with disease, to further explore the identified regions in greater detail with the aim of localising the association to a specific gene, functional element, or SNP that is driving the signal and is therefore implicated in disease (Schaid et al., 2018). Spain and Barrett (Spain

and Barrett, 2015) highlight three essential components that are required for fine-mapping: all common SNPs need to be genotyped or imputed with high confidence, stringent quality control are adhered to, and large sample size is required to provide sufficient power to differentiate between SNPs in high LD. The CoreExome data used in this study meets these requirements, providing good coverage in the 1 Mb window under investigation. With datapoints for over 5000 SNPs this genotyping array, following imputation, has roughly one SNP for every 200 bp of the genomic window. All appropriate quality control steps where followed and performed by Dr Zhixiu Li of the Brown group and a clean subset of the wider dataset was provided for this fine-mapping project. Finally, the European cohort that make up this dataset is sufficiently large and comparable with the cohort used in the IGAS study (see Table 1.1, Introduction). This chapter will use the CoreExome dataset and perform statistical fine-mapping approaches to prioritise disease associated SNPs into putatively causal SNP sets.

3.1.4 Statistical approaches for fine-mapping

The fine-mapping approach used in this chapter follows two arms of analysis as suggested by Spain and Barrett (Spain and Barrett, 2015). The first approach was to rank SNPs based on their p-value and consider all SNPs within a certain threshold as being potentially causative. This approach is covered under the frequentist analysis approach below. The second approach was to assign each SNP with a posterior probability of causality and to then use this value to identify credible sets with a specified likelihood of containing the causative SNP. This approach is covered under the Bayesian analysis approach below. It is also important to examine the LD relationship among the SNPs surrounding the lead

SNP. Once prioritised SNPs were identified through frequentist and Bayesian approaches the LD structure within these sets was investigated.

3.1.5 Frequentist approaches

GWAS analysis was performed using PLINK employing a logistic regression model to perform the case-control test. This test compares the allele frequency of AS patients to healthy controls in this cohort and output a p-value for the significance of the difference. This value is used to rank the SNPs in the test window and threshold cut-offs will be applied to exclude SNPs. The SNPs were initially filtered to identify those that met a suggestive significance level p-value less than or equal to 5×10^{-5}. A more stringent well-established threshold was then applied to identify SNPs that are associated with AS at genome-wide significance. These are SNPs with a p-value less than or equal to 5×10^{-8}. This frequentist approach was used as the initial step of prioritisation.

3.1.6 Bayesian approaches

Bayesian analysis was used to identify a minimal set of SNPs that contains the causal SNP by determining credible sets. The CoreExome genotyping data was analysed in SNPTEST using the additive model and inbuilt priors to calculate a Bayes Factor (BF) at each SNP. The Bayes Factor, the evidence of association, was used to calculate the posterior probability of the test SNP being causal at the region (Stephens and Balding, 2009). The posterior probability, the weight of evidence that a SNP should be included as potentially causative, was then used to rank the SNPs and generate credible sets. Two credible sets were generated:

(i) a 99% credible set (the minimal set of SNPs with 99% probability of containing the causal SNP) and (ii) a 95% credible set.

Bayesian analysis has previously been performed on this *NPEPPS* region by van Bunt and colleagues (van de Bunt et al., 2015). Their analysis identified *rs4239162* as most significant variant in the region (log10(BF) = 9.95, OR = 1.15, posterior probability (PP) = 0.04) and identified a 99% credible set containing 141 SNPs (95% credible set = 113) located across a 247,366 bp region. This analysis was used as a benchmark for comparison of the performance of the Bayesian analysis in this study.

3.1.7 Conditional analysis

Following these two approaches conditional analysis was performed to identify whether the observed disease association signal is due to a single signal or if the main observation is masking a secondary association. This is an extension of GWAS logistic regression models where a SNP of interest (usually the lead SNP) is included as a covariate and controlled for in the statistical model. The technique is used to identify if other SNPs in the region are independently associated with disease or if all SNPs are connected at a given signal. When controlling on the lead SNP, if all the association signal is diminished the association is controlled by the SNP included in the model but if an association signal remains then there is evidence of a secondary association at this region. PLINK includes functionality to perform conditional analysis on the frequentist model whilst SNPTEST has similar commands to be used for the Bayesian analysis.

3.1.8 Linkage disequilibrium

LD, the non-random association of alleles at different genomic locations, can be used to assist in the fine-mapping process. The lead SNP at a disease associated loci may be tagging the causal SNP which are in strong LD to each other. As such, it is prudent to consider all SNPs that are in strong LD to the lead SNPs when prioritising causal SNP sets. In this study the SNPs prioritised into casual sets were investigated as to their pairwise correlation (R^2) to the lead SNP. The SNPs in this list were also examined for their pairwise correlation to each other to establish the haploblock structure observed in the region.

Although these statistical fine-mapping approaches may not directly identify causal variants, and do not provide direct evidence of biological processes linking a variant to disease they are able to narrow down the list of potential SNPs to those with the highest association to disease and therefore statistical likelihood of a role in disease. By forming prioritised sets of SNPs, using multiple sources of statistical evidence of association, future steps of fine-mapping can be focused on the most plausible variants and specific genomic region where efforts are likely to be the most productive.

3.1.9 Aims

Aim 1: To fine-map the AS disease association signal from CoreExome GWAS data and identify and prioritise the causative genetic variants.

1.1 To analyse the CoreExome GWAS data for the *NPEPPS* locus using frequentist analysis in PLINK, perform conditional analysis, and prioritise the key SNPs driving this signal.

1.2 To analyse the CoreExome GWAS data for the *NPEPPS* locus using Bayesian analysis in SNPTEST, perform conditional analysis, and prioritise the key SNPs driving this signal.

3.2 Results

3.2.1 Frequentist analysis of *NPEPPS* region prioritises 96 SNPs at genome-wide significance

Frequentist analysis was performed in PLINK 1.9 in a logistic regression model. Following GWAS analysis the lead SNP at the *NPEPPS* region was *rs4794057* (AS risk allele = T, OR = 1.12, p-value = 2.4×10^{-11}). This SNP is a synonymous variant in a single *TBKBP1* transcript (ENST00000361722.3) and an intronic variant in the remaining 2 *TBKBP1* transcripts.

At this locus, 96 SNPs reached genome-wide significance ($p<5 \times 10^{-8}$) spanning a region of 307,006 bp and overlapping 5 genes; *EFCAB13*, *MRPL45P2*, *NPEPPS*, *KPNB1*, and *TBKBP1* (see Figure 3.2). 193 SNPs reaching the suggestive significance threshold ($p<5 \times 10^{-5}$). See Supplementary Table 3.1 and 3.2 for full list of SNPs.

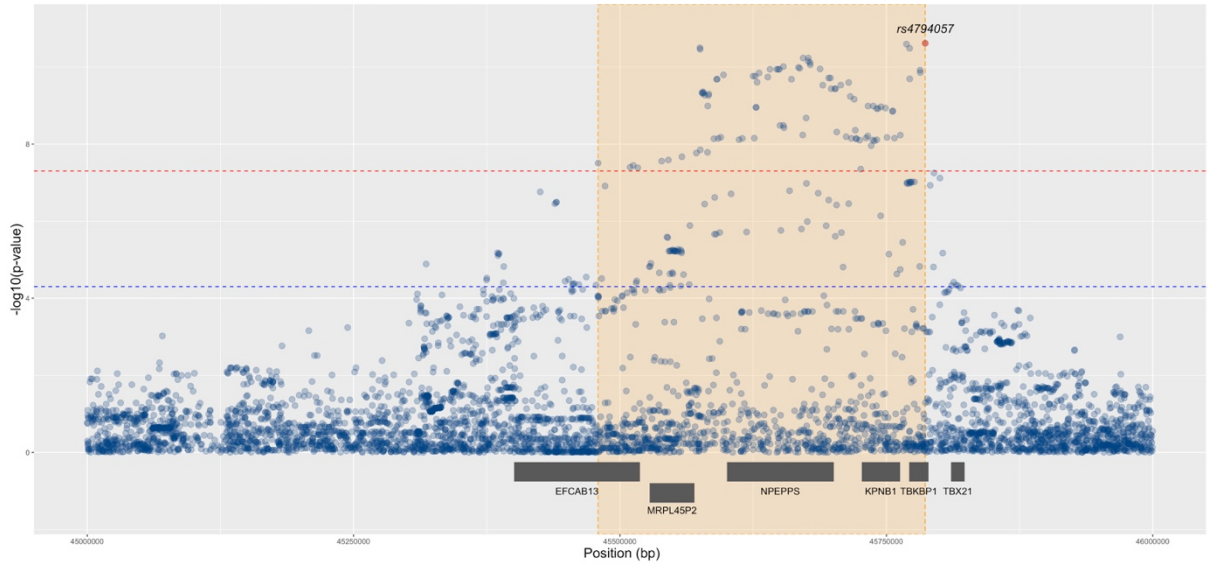


Figure 3.2: CoreExome frequentist association plot for the *NPEPPS* region. The lead SNP, *rs4794057*, has been highlighted in red with the 307,006 bp window capturing the genome-wide significant SNPs highlighted in orange. Red dotted line denotes the genome-wide significance threshold ($p<5x10^{-8}$). Blue dotted line denotes the suggestive significance threshold ($p<5x10^{-5}$).

The lead SNPs for the *NPEPPS* region identified in previous studies (see Table 1.1, Chapter 1) were all associated at genome-wide significance in this analysis (*rs8070463*, $p=2.56x10^{-11}$; *rs9901869*, $p=3.49x10^{-11}$; *rs12943464*, $p=2.98x10^{-10}$) as well as the lead SNP identified by an approximate Bayesian analysis of the IGAS 2013 dataset (*rs4239162*, $p=1.43x10^{-9}$) (van de Bunt et al., 2015). These SNPs are highlighted in Figure 3.3.

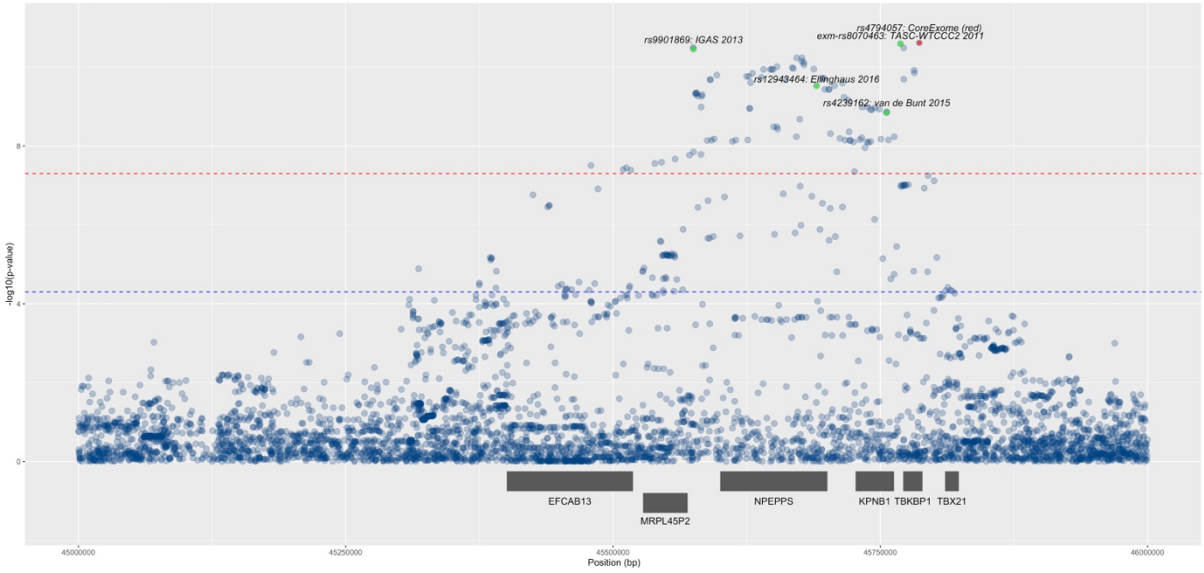


Figure 3.3: CoreExome frequentist association plot for the *NPEPPS* region depicting the lead SNPs from previous AS GWAS studies. The lead SNP, *rs4794057*, has been highlighted in red with the lead SNPs from previous GWAS studies highlighted in green. All previous lead SNPs were identified at genome-wide significance in the current study.

3.2.2 Bayesian Analysis of *NPEPPS* region identifies 62 SNPs in the 99% credible set

Bayesian analysis was performed in SNPTEST v2.5.4 (Marchini et al., 2007, Wellcome Trust Case Control Consortium and The Australo-Anglo-American Spondylitis Consortium, 2007) using the CoreExome genotyping data and applying an additive model with inbuild priors to calculate the log10(Bayes Factor) for each SNP. The lead SNP was identified as *rs4794057* (log10(BF) = 8.61, posterior probability (PP) = 0.09). This is the same lead SNP as observed following frequentist analysis in PLINK.

Post-analysis was performed in R to calculate the Bayes Factor, posterior odds (PO), posterior probability of association (PPA), and posterior probability (PP) for

each SNP as described in Stephens and Balding (Stephens and Balding, 2009) and Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium et al., 2012). The SNPs were then ranked by PP and the cumulative sum of PP calculated to establish the 99% and 95% credible sets of SNPs. The 99% credible set contained 62 SNPs and spanned a region of 211,521 bp. The 95% credible set contained 43 SNPs and spanned the same genomic region. This genomic region overlaps 3 genes; *NPEPPS*, *KPNB1*, and *TBKBP1* (see Figure 3.4). See Supplementary Table 3.3 and 3.4 for full list of SNPs in credible sets. All SNPs in the Bayesian analysis credible set are present in the frequentist analysis genome-wide significant SNPs list (n = 93).

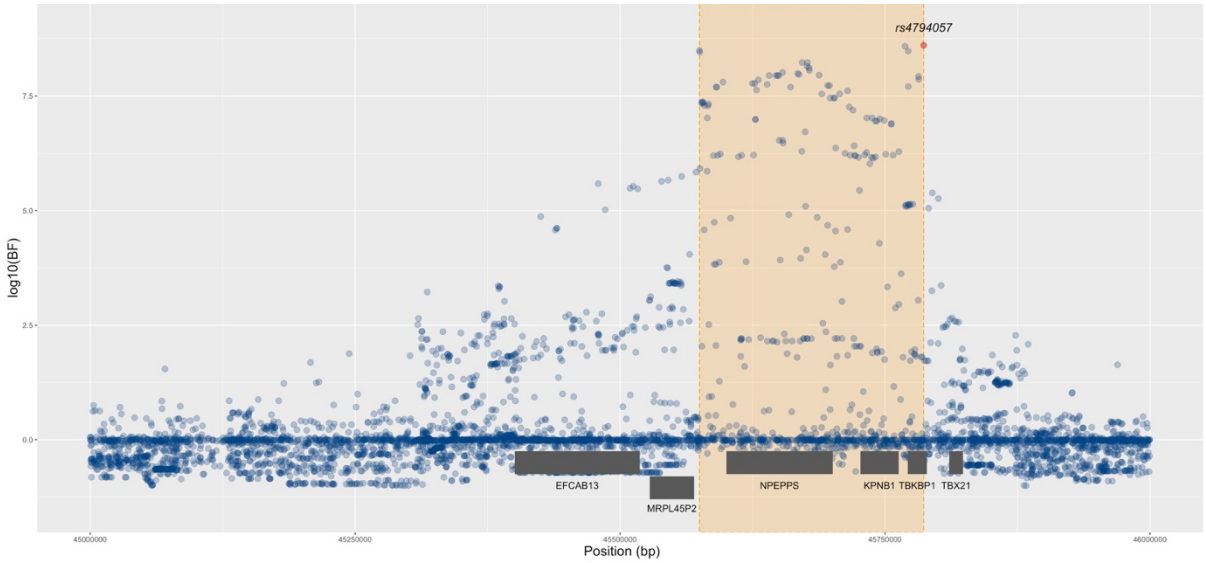


Figure 3.4: CoreExome Bayesian analysis association plot for the *NPEPPS* region. The lead SNP, *rs4794057*, has been highlighted in red with the 211,521 bp window capturing the 99% and 95% credible sets highlighted in orange.

The top 10 ranked SNPs following Bayesian analysis were consistent with those from frequentist analysis and account for 53% of the association signal observed in the tested window (cumulative (PP) = 0.534). These SNPs and their test statistics and ranks from frequentist and Bayesian analysis are summarised in Table 3.1.

								Frequentist Analysis			Bayesian Analysis			
Position on chr17	**SNP**	**Ref allele**	**Alt allele**	**TEST**	**OR**	**OR Lower**	**OR Upper**	**P**	**standard error**	**Rank**	**log10(Bayes Factor)**	**standard error**	**PP**	**Rank**
45786452	*rs4794057*	C	T	ADD	1.12	1.08	1.17	2.43E-11	0.02	1	8.61	0.02	0.09	1
45768836	*exm-rs8070463*	T	C	ADD	1.13	1.08	1.17	2.56E-11	0.02	2	8.59	0.02	0.09	2
45574931	*rs11651766*	G	T	ADD	1.13	1.09	1.17	3.17E-11	0.02	3	8.49	0.02	0.07	3
45771873	*rs67919208*	T	C	ADD	1.13	1.08	1.17	3.28E-11	0.02	4	8.48	0.02	0.07	4
45575206	*rs9901869*	G	A	ADD	1.13	1.09	1.17	3.49E-11	0.02	5	8.45	0.02	0.06	5
45671869	*rs12937560*	T	C	ADD	1.13	1.08	1.17	5.86E-11	0.02	6	8.23	0.02	0.04	6
45676760	*rs8075411*	T	C	ADD	1.13	1.08	1.17	5.86E-11	0.02	7	8.23	0.02	0.04	7
45676526	*rs8078880*	C	G	ADD	1.13	1.08	1.17	7.13E-11	0.02	8	8.15	0.02	0.03	8
45678408	*rs4793908*	A	G	ADD	1.13	1.08	1.17	7.73E-11	0.02	9	8.11	0.02	0.03	9
45678579	*rs4793913*	T	C	ADD	1.12	1.08	1.17	8.80E-11	0.02	10	8.06	0.02	0.03	10

Table 3.1: Summary statistics of top 10 ranked SNPs following frequentist analysis and Bayesian analysis. Ranking is agreeable across both methods.

3.2.3 Conditional analysis does not identify any independently associated signals in the *NPEPPS* region

Conditional analysis was performed in PLINK to identify any independent association signals in this region. Each SNP of the 96 genome-wide significant SNPs identified by frequentist analysis was independently controlled for in the logistical regression model. Conditional analysis showed that each of the genome-wide significant SNPs was able to control for the GWAS signal at this region with all previous association signal reduced to below the suggestive significance threshold (residual signal after independently controlling for each of the top 4 associated SNPs see Figure 3.5). No independent signal was observed for this region and the disease association signal can be classed as one signal for subsequent analysis.

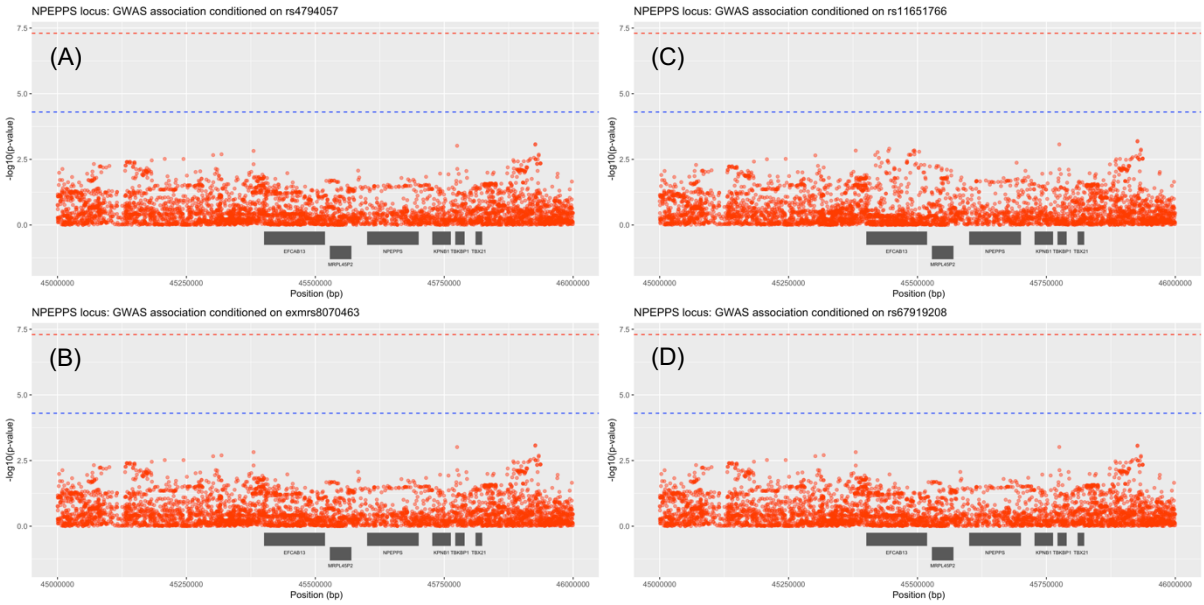


Figure 3.5: Conditional analysis on the top four frequentist GWAS association SNPs at the *NPEPPS* region. Plots (A) to (D) show the resulting association signal when controlling for the top four SNPs at this locus. (A) *rs4794057*, (B) *exm-rs8070463*, (C) *rs11651766*, (D) *rs67919208*.

Conditional analysis was subsequently performed in SNPTEST using the Bayesian analysis data to verify that only one association signal is present at the *NPEPPS* region. As can be seen in Figure 3.6, when any of the 4 top Bayesian analysis SNPs is independently controlled for a minimal residual signal remains. This confirms the result observed in the frequentist conditional analysis, that the association of the *NPEPPS* region can be classified as one signal.

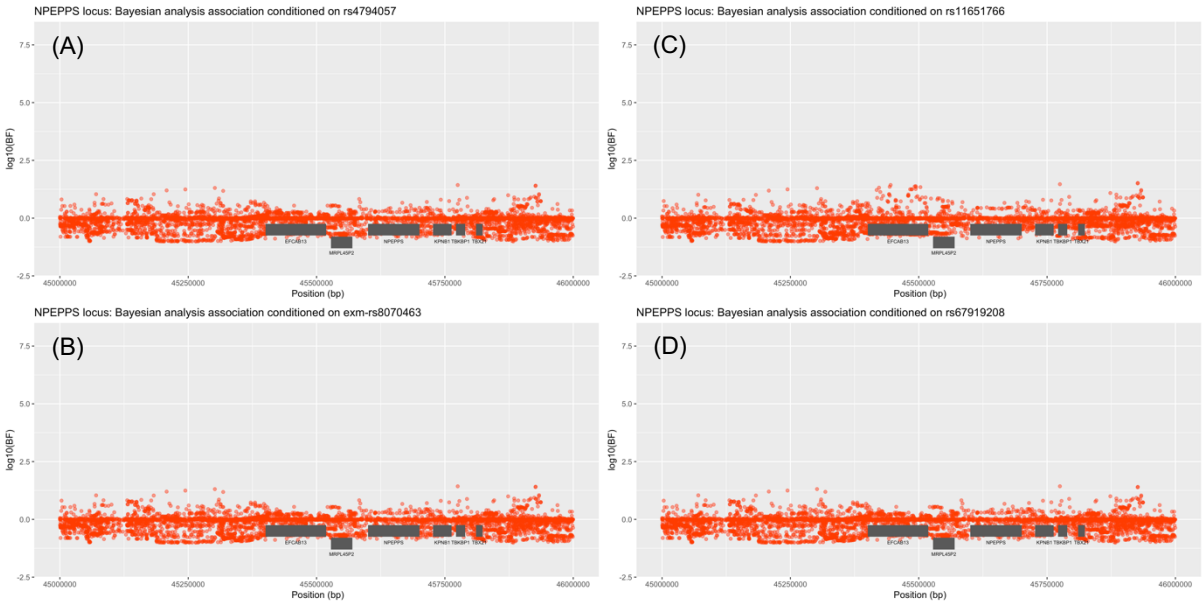
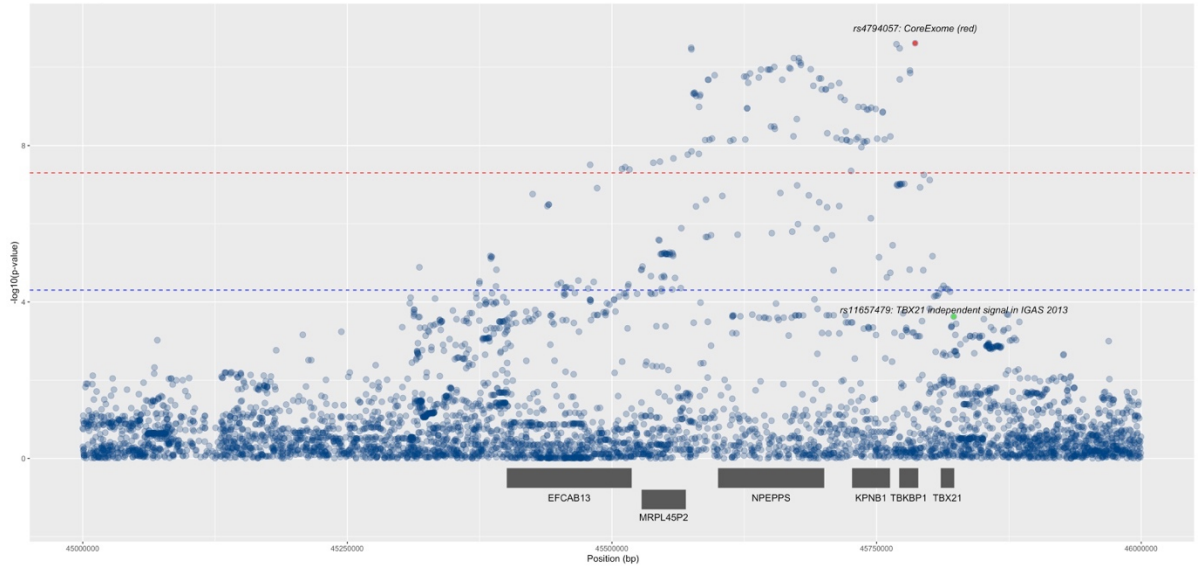


Figure 3.6: Conditional analysis on the top four Bayesian association SNPs at the *NPEPPS* region. Plots (A) to (D) show the resulting association signal when controlling for the top four SNPs at this locus. (A) *rs4794057*, (B) *exm-rs8070463*, (C) *rs11651766*, (D) *rs67919208*.

3.2.4 Previously reported independent *TBKBP1*/*TBX21* signal not replicated

It was previously reported that *NPEPPS* and *TBKBP1/TBX21* are associated with AS at this locus independently from each other (International Genetics of Ankylosing Spondylitis Consortium et al., 2013). The CoreExome data was reviewed to see if this independently observed association could be confirmed. In frequentist analysis the lead *TBX21* SNP, *rs11657479*, was in low LD to the lead SNP (R^2=0.21) and was not significantly associated with disease (p=2.36x10^{-4}). Figure 3.7(a) shows the association signal for the *NPEPPS* region with the *TBX21* associated SNP highlighted. Following conditional analysis, the lead residual SNP was *rs62076137* (position = chr17:45926752, p=8.4x10^{-4}), no independent signal was observed for *TBKBP1/TBX21* as previously described (see Figure 3.7(b)).

(A)



(B)

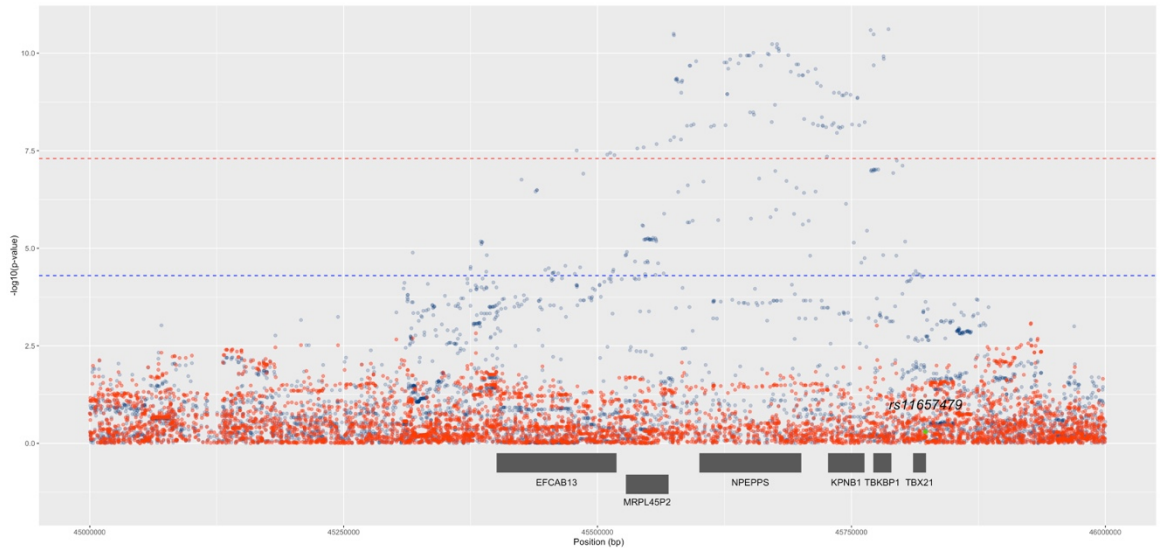


Figure 3.7: CoreExome frequentist association plot and conditional analysis plot for the *NPEPPS* region in relation to the *TBX21* lead SNP *rs11657479*. (A) CoreExome association plot highlighting the previously observed *TBX21* lead SNP *rs11657479* ($p=2.36x10^{-4}$). (B) The residual signal for the region after conditional analysis, controlling for the lead SNP *rs4794057*. No independent signal observed for the *TBX21* SNP *rs11657479* (p=0.50).

3.2.5 Linkage disequilibrium of the *NPEPPS* region

3.2.5.1 PLINK identifies strong LD structure

The LD structure to the lead SNP (rs4794057) was calculated in PLINK 1.9 and visualised as a zoomplot where R^2 values were assigned to colour (Figure 3.8A). All genome-wide significant SNPs and those in the 99% credible set have an R^2 value of greater or equal to 0.68, indicating a high level of pairwise LD between each genome-wide significant SNP and the lead SNP.

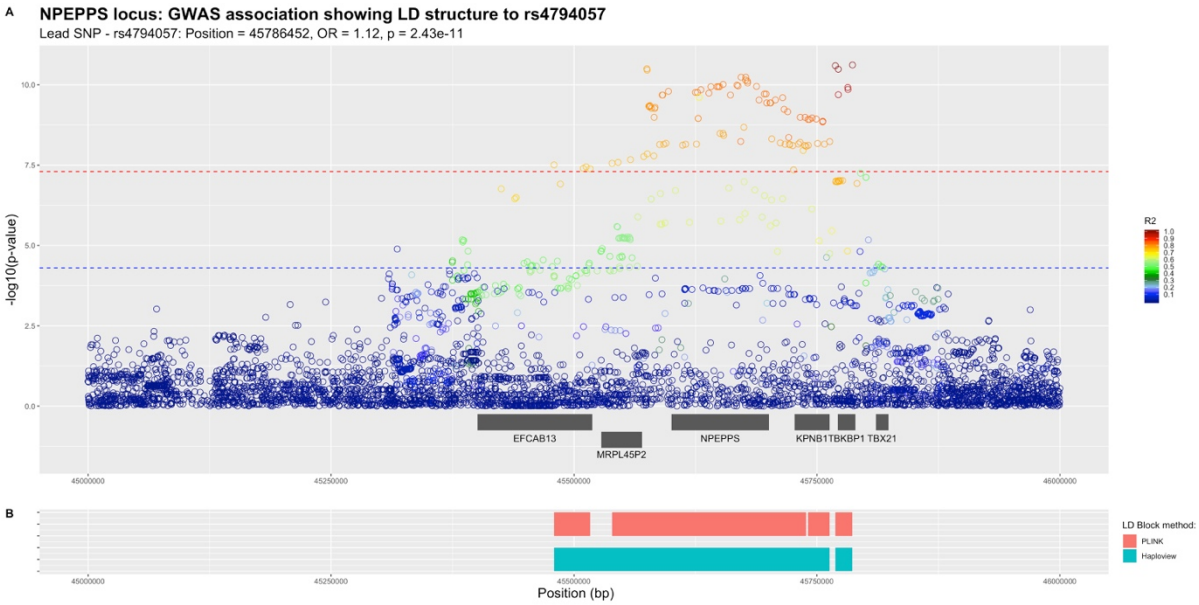


Figure 3.8: Frequentist association plot showing LD structure to the lead SNP. (A) LD structure to lead SNP, *rs4794057*, as calculated in PLINK. (B) Corresponding LD blocks for genome-wide significant SNPs aligned to region as calculated in PLINK and Haploview.

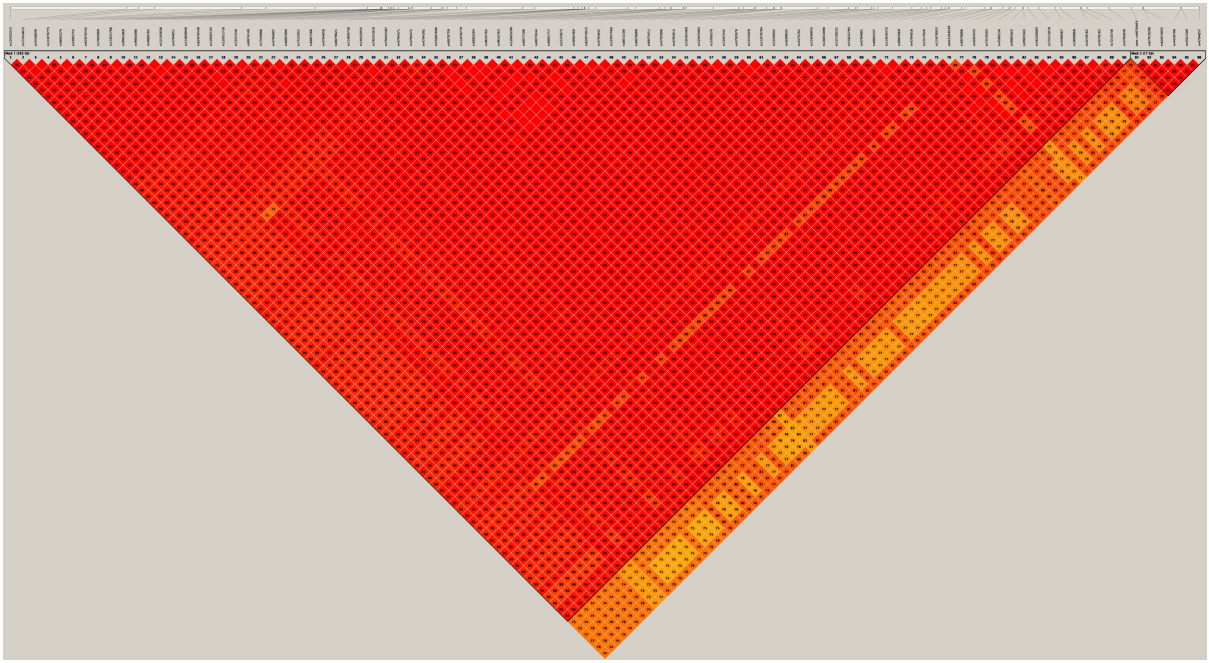
Linkage disequilibrium block estimation was also performed in PLINK using the `--blocks` function restricted to the 96 genome-wide significant SNPs. Four LD blocks were identified across the region; Block 1 contained 4 SNPs and spanned 37kb overlapping *EFCAB13*, Block 2 (the largest block) contained 76 SNPs and spanned nearly 200kb overlapping the 3`end of *MRPL45P2*, *NPEPPS*, and 5`

KPNB1, Block 3 contained 10 SNP overlapping 22kb of *KPNB1*, and Block 4 contained 6 SNPs and spanned 17.6kb localised to *TBKBP1*. This data is summarised in Table 3.2 and the relative position of the LD blocks indicated in Figure 3.8B.

3.2.5.2 Haploview confirms LD structure identified by PLINK

The LD structure of the region was also examined using Haploview (Barrett et al., 2005), where 2 LD blocks were identified in the 96 genome-wide significant SNPs. Block 1 contains 90 SNPs and capture a region from *rs12452315* to *rs4794048* covering 283kb. This single block captures LD blocks 1 to 3 identified in the PLINK analysis and overlaps 4 genes: *EFCAB13*, *MRPL45P2*, *NPEPPS*, and *KPNB1*. Block 2 contains the remaining 6 SNPs and covers a region 17kb in length from exm-rs8070463 to *rs4794057*. This block is the same as LD Block 4 from the PLINK analysis and overlaps *TBKBP1*. The LD blocks are shown in Figure 3.8B and the LD structure visualised in Haploview format in Figure 3.9A. Two main haplotypes are observed across both LD blocks due to the strong LD observed throughout the entire region (Figure 3.9B).

(A)



(B)



Figure 3.9: LD structure and haplotype estimation of 96 genome-wide significant SNPs as calculated in Haploview. (A) Haploview pairwise LD structure. (B) Haploview haplotype estimations. Two blocks of strong LD can be observed the LD structure map corresponding to the LD block estimation in Figure 3.8. Haplotype estimations provided with two haplotypes estimated at a higher frequency.

Chr	Start (BP)	End (BP)	Size (kb)	No. SNPs	PLINK block	Haploview Block	SNPs in block
17	45479446	45516667	37.222	4	1	1	*rs12452315, rs77454872, rs4338849, rs11079775*
17	45539117	45738861	199.745	76	2		*rs9895274, rs2085755, rs62076542, rs7405889, rs11651766, rs9901869, rs9903898, rs4060785, rs12103836, rs4794051, rs11869940, rs62076549, rs12602134, rs12051716, rs7221548, rs8074149, rs7350906, rs4794007, rs8065990, rs7222851, rs8067286, rs4794058, rs8079821, rs11869730, rs10432035, rs78551919, rs79443387, rs4794371, rs4794372, rs4794392, rs55687600, rs8070759, rs11079779, rs8082284, rs9635762, rs9635763, rs12603290, rs8077106, rs8072644, rs8081717, rs2125873, rs4092460, rs62074015, rs4793842, rs12937560, rs8072100, rs8078880, rs8075411, rs4793908, rs4793913, rs10163469, rs12943464, rs3760370, rs7207542, rs4793978, rs7223029, rs11079784, rs7220935, rs4289035, rs4375701, rs11656855, rs10853090, rs12150231, rs55633793, rs4794001, rs8065437, rs11650372, rs8065669, rs4794016, rs2317826, rs11870935, exm2264630, rs8078686, rs4264433, rs7219303, rs4580230*
17	45740805	45763073	22.269	10	3		*rs9905922, rs9905583, rs56226803, rs73315878, rs4399567, rs3809868, rs4239162, rs4239163, rs7210738, rs4794048*
17	45768836	45786452	17.617	6	4	2	*exm-rs8070463, rs67919208, rs62074055, rs4439799, rs4561509, rs4794057*

Table 3.2: Table to show LD blocks for genome-wide significant SNPs as calculated by PLINK and Haploview.

3.3 Discussion and conclusion

Statistical fine-mapping was performed on the *NPEPPS* region using the CoreExome dataset to identify and prioritise the causative SNPs for this region. A combination of frequentist approaches and Bayesian analysis were successfully performed and both methods identified sets of variants with a high likelihood of containing the causal SNP. Both analyses identified *rs4794057* as the lead SNP for this disease association signal. Frequentist analysis identified 96 SNPs at genome-wide significance within a strong LD block. Bayesian analysis identified a credible set of 62 SNPs with a 99% probability of containing the causal SNP. There was strong correlation between both approaches with all 62 SNPs identified by Bayesian analysis contained within the frequentist genome-wide significant set and the top 10 SNPs consistently identified across both methods (see Supplementary Figure 3.2 for correlation between results). Following conditional analysis in both approaches, no secondary disease associated signal was identified.

This analysis further confirms the presence of a disease association signal at this region and improves on previous GWAS analyses by offering increased coverage. Unfortunately, the current fine-mapping analysis was not able to attribute the signal to a single gene, with the main signal overlapping *NPEPPS*, *KPNB1*, and *TBKBP1*. The sets of SNPs prioritised by both techniques remains large and lie over a sizeable genomic region. When directly compared to the Bayesian analysis performed by van Bunt and colleagues (van de Bunt et al., 2015) using the IGAS European dataset, the analysis from the CoreExome data

improved on the fine-mapping outputs. The number of SNPs in the 99% credible set is reduced from 141 to 62 and the overlapping window of the credible set was slightly condensed from 247,336 bp to 211,521 bp, with the genomic window reduced at the 5` end and focused over 3 genes (see Supplementary Figure 3.1).

The LD observed in the human genome is an element that makes GWAS analysis possible but is also a barrier to the identification of causal SNPs in a region like this where the LD remains strong across a large region. With access to a dataset from an alternative ethnic group, that has an alternative LD structure in this region, it may be possible to prioritise the current SNP sets further.

Further approaches are required to assign causality to the variants in the prioritised SNP sets. The next logical step in this process is to annotate the region and identify where these SNPs lie in terms of the genes in the region and known functional elements which may be suggestive of the role of the variants in disease.

This statistical fine-mapping approach has only been applied to the *NPEPPS* region in this study. The approach used to fine-map this region should be applied using the CoreExome dataset to all other disease associated loci previously identified. For the *NPEPPS* region the 1 Mb window for analysis was defined manually which would not be feasible when considering multiple regions. An alternative approach to define the fine-mapping window when considering additional loci would be to select the region 500 kb downstream and upstream of the lead SNP with manual adjustment where required to ensure the full disease-

association is captured. This approach would have been suitable for the *NPEPPS* region in this study. With greater understanding of computational approaches, it would be possible to automate an analysis pipeline that included the steps performed in this manual procedure and apply this to fine-map all disease-associated loci.

4 Annotation of *NPEPPS* region to identify variants for functional investigation

4.1 Introduction

Annotation of the genomic landscape is crucial to aid in the prioritisation of variants and identification of their potential functional roles in disease. Functional annotation brings in independent sources of information about each variants biological role (Spain and Barrett, 2015). This process aligns the statistical association of a variant to genomic coordinates with biological context allowing further prioritisation and design of studies to determine the functional mechanisms by which the variants influence disease.

As discussed in Section 1.4.3, variants have traditionally been initially categorised as protein coding or non-protein coding. Coding variants have the potential to change the coding sequence of a gene and consequentially affect protein structure. This generally makes them easier to support for prioritisation as a clear consequence of the variant can be predicted (Spain and Barrett, 2015).

However, the majority of GWAS SNPs associated with disease are found in introns and non-coding regions of the genome (Maurano et al., 2012) and are

likely to have an impact on gene regulation. DNA regulatory elements are characterised by accessible chromatin, histone modifications, and TF binding motifs (Trynka and Raychaudhuri, 2013). Assigning function and causality to these variants is more complex than for coding variants as the intronic and non-coding genomes are less clearly defined or annotated, and can be highly cell-type, context, and disease specific (Andersson et al., 2014). Thus, annotation is not as fixed, and the source of the data must be considered.

4.1.1 Experimental methods to define DNA regulatory regions

DNA is organised into compact, highly organised chromatin structures by coiling around histone proteins in the nucleus of cells. For transcription to occur, chromatin must become less compact, enabling assembly of the transcriptional machinery. These changes are associated with different modifications of histone proteins. As a result, experimental processes to define regulatory features in the genomic landscape identify areas of open chromatin or changes in histone modifications. Accessible chromatin can be assayed using techniques including DNase I hypersensitive site sequencing (DNase-seq), formaldehyde-assisted isolation of regulatory elements by sequencing (FAIRE-seq) or ATAC-seq. DNase-seq identifies areas of the genome that are accessible to cleavage by DNase I (Thurman et al., 2012). FAIRE-seq uses formaldehyde to cross-link DNA to proteins in nucleosome-depleted regions of the genome before shearing the DNA by sonication, and extraction followed by sequencing of the remaining DNA which is enriched for regions of accessible chromatin (Giresi et al., 2007). ATAC-

seq uses Tn5, a hyperactive transposase enzyme, to tag and fragment accessible DNA which is then sequenced (Buenrostro et al., 2013). Histone modifications known to mark regulatory regions are detected using ChIP-seq (Aparicio et al., 2005). In this method, proteins are crosslinked to DNA by formaldehyde fixation, DNA is fragmented by sonication, and antibodies specific to the target histone mark are used to precipitate the DNA. This is followed by sequencing. The pattern of histone modifications that are enriched in various chromatin states has been well documented (see Chapter 1.3.2). This chapter will focus on histone modifications that are enriched at enhancers and promoters. Examples of the histone modifications at these regulatory regions include; acetylation at the 27th lysine of histone H3 (H3K27ac) a histone mark associated with activated promoters and enhancer (Creyghton et al., 2010), mono-methylation of the 4th lysine of Histone H3 (H3K4me1) a histone mark associated with enhancers (Heintzman et al., 2009), and tri-methylation of the 4th lysine of Histone H3 (H3K4me3) a histone mark associated with active promoters near TSS (Heintzman et al., 2009).

Three large epigenomic mapping consortia, ENCODE (The ENCODE Project Consortium, 2012), NIH Epigenomic RoadMap (Bernstein et al., 2010) and BLUEPRINT (Adams et al., 2012) have produced comprehensive databases on various epigenetic features assayed using DNase-seq, ATAC-seq and ChIP-seq in a multitude of cell lines and primary cells. These serve as a fantastic resource to begin the process of prioritising SNPs that might have functional roles in AS. In this chapter these databases have been used alongside in-house epigenomic data (described below).

4.1.2 Bioinformatic tools available for SNP prioritisation

The set of AS-associated SNPs at the *NPEPPS* region defined in Chapter 3 will be analysed using the following tools in order to predict and rank which are most likely to have functional regulatory effects.

The Ensembl Variant Effect Predictor (VEP) is a toolset for the *"analysis, annotation, and prioritization of genomic variants in coding and non-coding regions"* (McLaren et al., 2016). VEP predicts the gene and transcript a variant is located in using the GENCODE gene set (GRCh37 hg19) and identifies any protein sequence changes. For non-coding variants, VEP reports variants in non-coding RNA, genomic regulatory regions, or transcription factor binding motifs. This regulatory annotation uses data from ENCODE (The ENCODE Project Consortium, 2012), BLUEPRINT (Adams et al., 2012), and the NIH Epigenomics Roadmap (Bernstein et al., 2010).

HaploReg (Ward and Kellis, 2012) is a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants (Ward and Kellis, 2012) and will be employed to give further detail on the known regulatory context of the prioritised SNPs. This will act as an additional level of annotation on top of that provided by VEP.

RegulomeDB is a database that uses regulatory information to infer SNP function, and classifies variants using an in-built scoring system to assist in

variant comparison (Boyle et al., 2012). The RegulomeDB database combines ENCODE transcription factor ChIP-seq, histone ChIP-seq, FAIRE, and DNase I hypersensitivity data with additional NCBI Sequence Read Archive TF ChIP-seq data and a large collection of eQTL data from over 100 tissues and cell lines (Boyle et al., 2012). The output of this tool is an annotation score based on the confidence that a variant lies in a functional location and is likely to result in a functional consequence based on its annotation. The scoring system is from 1, variants likely to affect binding of a TF and linked to expression of a gene target, to 7, no annotation support.

CADD is a tool that uses integrative annotation from genomic features to score the deleteriousness of single nucleotide variants and short insertions and deletions (Rentzsch et al., 2019). This measure can be used to prioritise causal variants in genetic analysis. It is a single score derived from the integration of multiple genomic features which are obtained from VEP, conservation and selection scores, UCSC genome browser tracks, and epigenomic information from ENCODE and NIH RoadMap projects (Rentzsch et al., 2019). The CADD score generated is primarily used to identify variants that are most likely to be deleterious and potentially pathogenic and will be used in the annotation process as a score to rank the prioritised variants. The PHRED-scaled score (the CADD score ranked in order of magnitude) is used for the ranking process. In this scoring system, a PHRED score of >10 indicates a SNP with a raw score in the top 10% of all deleterious reference genome single nucleotide variants (SNVs).

Once ranked using the two annotation scoring systems, the ranked prioritised SNPs were transformed into visualisation tracks that were aligned in the WashU epigenome browser.

4.1.3 Visualisation of publicly available data

Data from the NIH Epigenomic Roadmap database was visualised in chromHMM format using the WashU Epigenome browser to establish a visual representation of the functional landscape. ChromHMM (Ernst and Kellis, 2012) is a computational method for the visualisation of genome-wide chromatin-state annotation. The model learns the combination of different chromatin marks that are present or absent at known functional elements and uses this information to predict and assign biological function to each genomic position based on the chromatin marks present. Data for CD4, CD8, and CD14 cells was uploaded as these cell-types have been previously implicated in ankylosing spondylitis (Wright et al., 2009, Szalay et al., 2012, Shen et al., 2009) and are the focus of ongoing cell-type specific research by the Knight group. Further cell-types, including NK cells and gamma delta T cells, have been implicated in disease but these have not been the focus of work in the Knight group and in-house data is not available for comparison. Overlaps can be identified by aligning the prioritised SNP set to these chromHMM cell-type specific tracks to located regions with predicted function.

4.1.4 In-house AS datasets used for annotation

Epigenomic datasets from AS patients and healthy controls were available from work performed by the Knight group (see Materials & Methods 2.3.3). These data were used to establish the epigenomic landscape of the *NPEPPS* region specifically in AS. Cell-type specific data, for CD4, CD8, and CD14 cells, were available as post-analysis peak coordinate files for ATAC-seq experiments, and ChIP-seq for H3K27ac and H3K4me3 histone marks. For the ATAC-seq experiments and H3K27ac ChIP-seq experiments, master peak files had been generated for CD4, CD8, and CD14 cell-types containing peaks present in greater than 30% of samples. These files were used to identify overlaps between ATAC-seq or ChIP-seq peaks and the prioritised SNP set. This data was also aligned as tracks in the WashU epigenome browser for visual representation of overlaps.

In this chapter, the prioritised SNP set defined by statistical methods in Chapter 3 was annotated using publicly available datasets and tools. This enabled further prioritisation by predicting the likely regulatory function of each SNP. This chapter uses (i) computational tools that predict overlap with coding or regulatory features (ii) overlap with published experimental data (iv) overlap with in-house AS-specific experimental data. The results were visualised on the WashU epigenome browser to facilitate biological interpretation and select SNPs for validation with experimental methods in Chapter 5 and 6.

4.1.5 Aim

Aim 2: To use epigenomic data generated by the Knight group and from publicly available databases to further annotate the *NPEPPS* locus to identify functional regions of interest that overlap with the key SNPs driving this signal.

4.2 Results

4.2.1 Annotation of *NPEPPS* region using publicly available SNP prioritisation tools

4.2.1.1 Ensembl Variant Effect Predictor and HaploReg outputs

The genome-wide significant SNPs prioritised for the *NPEPPS* region (n = 96) were run through Ensembl VEP (McLaren et al., 2016) using the standard annotation settings. A single SNP, *rs4794057*, was identified as overlapping a coding region of the *TBKBP1* transcript ENST00000361722.3. This SNP is a synonymous variant with no impact on the protein structure. It is located central to exon 8 (position 362 of 728 bp) in a gene structure containing 9 exons and is unlikely to have an effect on splicing. The remaining 95 SNPs were identified as either intronic or intergenic variants. Of these non-coding variants, 20 SNPs overlapped with 15 known regulatory features (Table 4.1). Five SNPs overlapped promoters, five SNPs overlapped with promoter flanking regions, four SNPs overlapped enhancers, three SNPs overlapped CTCF binding sites, one SNP overlapped an open chromatin region, and a single SNP overlapped an open chromatin region that also contained a CTCF binding site.

As the majority of prioritised SNPs (98.9%) were aligned to non-coding regions of the genome, tools designed to annotate and inform on regulatory function were used for further annotation and to rank SNPs to identify those with a functional role. The genome-wide significant SNPs were run through HaploReg (version 4.1) (Ward and Kellis, 2012) using standard inbuilt settings to provide further

annotation. HaploReg confirmed *rs4794057* as a synonymous mutation in *TBKBP1* and identified 55 of the 96 SNPs as being located in intron. Of the 96 SNPs, 78 were labelled as overlapping chromatin marks, eight SNPs were annotated as under selective constraint of which two overlapped the VEP annotated SNPs that overlapped with regulatory features (*rs67919208* and *rs3760370*), and one SNP (*exm-rs8070463*) was annotated as being associated with disease in previous GWAS studies (ankylosing spondylitis and multiple sclerosis). The summary data for all SNPs are provided in Supplementary Table 4.1.

SNP	Position in chr17	Category	Ensembl FeatureID	Regional context
rs2085755	45545378	CTCF binding site	ENSR00001129227	*MRPL45P2* intronic
rs4794051	45577914	CTCF binding site	ENSR00000283591	*MRPL45P2-NPEPPS* intergenic
rs11869940	45578435	CTCF binding site	ENSR00000283591	*MRPL45P2-NPEPPS* intergenic
rs8074149	45583375	promoter flanking region	ENSR00001129229	*MRPL45P2-NPEPPS* intergenic
rs8079821	45611704	promoter	ENSR00000095186	*NPEPPS* intronic
rs9635763	45649694	enhancer	ENSR00001129242	*NPEPPS* intronic
rs12603290	45650196	enhancer	ENSR00001129242	*NPEPPS* intronic
rs10163469	45687767	promoter flanking region	ENSR00001129249	*NPEPPS* intronic
rs3760370	45695832	open chromatin region	ENSR00001129252	*NPEPPS* intronic
rs3760370	45695832	CTCF binding site	ENSR00001129253	*NPEPPS* intronic
rs4793978	45698175	promoter flanking region	ENSR00000095189	*NPEPPS* intronic
rs7223029	45701731	open chromatin region	ENSR00001129255	*NPEPPS-KPNB1* intergenic
rs4289035	45707211	enhancer	ENSR00001129256	*NPEPPS-KPNB1* intergenic
rs4375701	45712218	enhancer	ENSR00001129257	*NPEPPS-KPNB1* intergenic
rs8065669	45724996	promoter	ENSR00000095190	*NPEPPS-KPNB1* intergenic
rs4794016	45725931	promoter	ENSR00000095190	*NPEPPS-KPNB1* intergenic
rs9905922	45740805	promoter flanking region	ENSR00000376579	*KPNB1* intronic
rs9905583	45741016	promoter flanking region	ENSR00000376579	*KPNB1* intronic
rs56226803	45742233	promoter flanking region	ENSR00000376579	*KPNB1* intronic
rs67919208	45771873	promoter	ENSR00000095197	*KPNB1-TBKBP1* intergenic
rs62074055	45771933	promoter	ENSR00000095197	*KPNB1-TBKBP1* intergenic

Table 4.1: Summary of Ensembl VEP output showing prioritised SNPs that overlap regulatory features. SNP name and position are provided along with VEP information on category of regulatory feature, Ensembl feature ID, and regional context of SNP location.

4.2.1.2 Ranking of prioritised SNPs by RegulomeDB

Following this initial categorisation of SNPs by VEP, the entire set of 96 SNPs was passed to RegulomeDB 2.0 (Boyle et al., 2012) and the output used to rank the SNPs based on the level of supporting evidence for each to be functional (Table 4.2 and Figure 4.1). Six SNPs were ranked in Category 1, meaning they are known eQTLs for genes, and therefore have been shown to be associated with expression and are most likely to be significant (Boyle et al., 2012). Two SNPs were labelled with the "b" sub-category as they had additional supporting annotation from TF binding, DNase footprint overlaps, and DNase peak data. These SNPs were *rs11079784* (score = 0.90, position = 45702280: *NPEPPS-KPNB1* intergenic region) and *rs4794048* (score = 0.46, position = 45763073: *KPNB1-TBKBP1* intergenic region). The remaining 4 SNPs (*rs8074149*, *rs8072644*, *rs11871606*, *rs8070463*) were categorised as 1f as these had additional annotation evidence from either TF binding or DNase peak data. These SNPs are considered likely to affect binding and are linked to expression of a gene target. A further 5 SNPs were classed as Category 2 variants and are likely to affect binding but without eQTL data. The Category 1 and 2 SNPs are summarised in Table 4.2 with all other SNPs summarised in Supplementary Table 4.2.

4.2.1.3 Ranking of prioritised SNPs by CADD

The SNP set was also passed through CADD to generate a score for the predicted deleteriousness of each SNP. The SNPs were then ranked based on the PHRED score generated. Seven SNPs had a PHRED score >10 indicating

that these SNPs are predicted to be in the 10% most deleterious substitutions in the human genome (Table 4.3 and Figure 4.1). Of these, *rs67919208*, had a PHRED score of 21.3 and is predicted to be in the top 1% of deleterious substitutions. This SNP was identified in VEP to overlap with a promoter of *TBKBP1*. The CADD output is summarised in Table 4.3, which also provides comparison for how the corresponding SNP ranked in statistical prioritisation. The full prioritised SNP ranking are provided in Supplementary Table 4.3.

Figure 4.1 summarises SNP prioritisation data from each method by visualisation in the WashU epigenome browser. Tracks are shown for prioritised SNPs using statistical approaches, RegulomeDB rank and CADD rank, along with protein coding genes and chromHMM tracks for CD4, CD8, and CD14 cell-types. This allows for visual comparison of the different ranking systems in the genomic context of where the SNP align.

SNP	Position on chr17	Raw score	Category	Rank	ChIP	DNase	Footprint	eQTL	eQTL Gene (Biosample)	Computational SNP rank	Comment
rs11079784	45702280	0.90	1b	1	TRUE	TRUE	TRUE	TRUE	*TBKBP1* & *LOC653479* (monocytes)	36	
rs4794048	45763073	0.46	1b	2	TRUE	TRUE	TRUE	TRUE	*TBKBP1* & *LOC653479* (monocytes)	66	
rs8070463	45768836	0.55	1f	3	TRUE	TRUE	FALSE	TRUE	*TBKBP1* & *LOC653479* (monocytes)	2	
rs11871606	45732774	0.55	1f	4	TRUE	FALSE	FALSE	TRUE	*TBKBP1* & *LOC653479* (monocytes)	49	
rs8072644	45653364	0.22	1f	5	FALSE	TRUE	FALSE	TRUE	*TBKBP1* & *LOC653479* (monocytes)	11	
rs8074149	45583375	0.04	1f	6	FALSE	TRUE	TRUE	TRUE	*TBKBP1* & *LOC653479* (monocytes)	43	VEP: Promoter flanking region
rs10163469	45687767	0.66	2b	7	TRUE	TRUE	TRUE	FALSE		14	VEP: Promoter flanking region
rs11870935	45732605	0.60	2b	8	TRUE	TRUE	TRUE	FALSE		97	
rs7219303	45737701	0.57	2b	9	TRUE	TRUE	TRUE	FALSE		49	
rs4793978	45698175	0.50	2b	10	TRUE	TRUE	TRUE	FALSE		36	VEP: Promoter flanking region
rs4794057	45786452	0.50	2b	11	TRUE	TRUE	TRUE	FALSE		1	Lead SNP

Table 4.2: SNPs classified as Category 1 or 2 variants by RegulomeDB. SNP name and position provided. Raw score as calculated by RegulomeDB. Category is the RegulomeDB category as described in text. Rank is the ranking order of SNPs based on RegulomeDB category and raw score. ChIP, DNase, Footprint, eQTL columns show further detail of the evidence for RegulomeDB category assignment according to the presence of supporting data for each method. eQTL gene ascribes the gene associated with the eQTL and the cell type where the eQTL is found. Computational SNP rank refers to the prioritisation from Chapter 3. Comments ascribes additional SNP annotation provided.

SNP	Position on chr17	Raw Score	PHRED	Rank	Consequence	Gene	Ensembl FeatureID	Computational SNP rank	Comment
rs67919208	45771873	2.19	21.30	1	Regulatory	NA	ENSR00000283592	4	VEP: Promoter
rs4794057	45786452	1.58	16.16	2	Synonymous	*TBKBP1*	ENST00000361722	1	Lead SNP
rs11870935	45732605	1.57	16.13	3	Intronic	*KPNB1*	ENST00000290158	67	
rs4561509	45781799	1.36	15.20	4	Regulatory	NA	ENSR00000095199	20	
rs4794058	45597098	1.05	13.60	5	Upstream	*NPEPPS*	ENST00000530173	22	
rs4794371	45627693	0.65	10.68	6	Intronic	*NPEPPS*	ENST00000322157	53	
rs62074055	45771933	0.57	10.04	7	Regulatory	NA	ENSR00000283592	28	VEP: Promoter

Table 4.3: Results of CADD SNP prioritisation showing SNPs that are classified as top 10% of deleterious substitutions in human genome. SNP name and position is provided. CADD raw score for deleteriousness of SNP. PHRED score shows scaled rank of raw score in context all genomic deleterious SNPs. Rank of SNPs based on CADD output. Consequence indicates the predicted effect of the SNP. Gene shows the nearest gene. Ensembl feature ID is the ID of the overlapping genetic feature. Computational SNP rank refers to the prioritisation from Chapter 3. Comments ascribes additional SNP annotation provided.

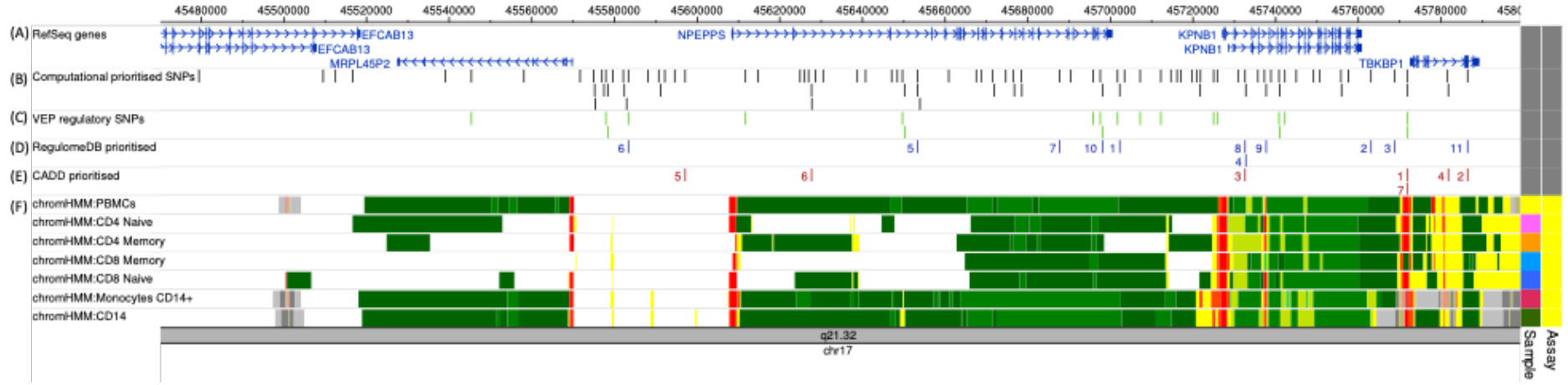


Figure 4.1: Annotation of SNPs in *NPEPPS* region using publicly available tools. Screenshot from WashU epigenome browser displaying (A) aligned to RefSeq genes, (B) computationally prioritised SNPs, (C) VEP regulatory SNPs, (D) category 1 and 2 ranked SNPs using RegulomeDB, (E) top 10% deleterious SNPs ranked using CADD PHRED scores, (F) chromHMM chromatin-states for CD4, CD8, and CD14 cell-types (green = strong transcription, dark green = weak transcription, yellow = enhancers, yellow green = generic enhancer, red = active TSS, orange red = flanking active TSS, grey = weak repressed PolyComb, dark grey = repressed PolyComb, white = quiescent/low).

4.2.2 Annotation of *NPEPPS* region from in-house data

4.2.2.1 Annotation from cell-type specific H3K27ac master peak

Bedtools intersect function was used to identify overlaps between the prioritised SNP positions and cell type specific H3K27ac master peak data generated from AS cases and healthy controls. A total of 16 master peaks for H3K27ac were identified across CD4, CD8, and CD14 cells in the fine-mapped window for the 96 prioritised SNPs (Chr17:45479446-45786452). These aligned to six genomic regions of which five (15 peaks) were shared across all three cell-types. Of the six genomic regions, five were overlapped by prioritised SNPs (Table 4.4 and Figure 4.2).

Notably, one CD4-specific peak overlapped two SNPs, *rs11869940* and *rs62076549*, in the intergenic region upstream of *NPEPPS*. This region was identified as a regulatory feature in VEP as it overlaps a CTCF binding site.

H3K27ac master peak position on chr17	Cell-Type	SNP	SNP position on chr17	Overlap in public data annotation
45567864-45571831 45567862-45571835 45567871-45571821	CD4 CD8 CD14	*rs7405889*	45571676	
45578178-45581570	CD4	*rs11869940* *rs62076549*	45578435 45579555	VEP: CTCF binding site
45724350-45729832 45722768-45729793 45722508-45729742	CD4 CD8 CD14	*rs8065669* *rs4794016*	45724996 45725931	VEP: Promoter VEP: Promoter
45734624-45740009 45735131-45739903 45735321-45739329	CD4 CD8 CD14	*rs8078686* *rs4264433* *rs7219303* *rs4580230*	45735706 45737275 45737701 45738861	RegulomeDB: 9th (Category = 2b)
45769609-45773873 45768075-45824390 45770064-45774361	CD4 CD8 CD14	*rs67919208* *rs62074055* **exm-rs8070463*	45771873 45771933 45768836	VEP: Promoter / CADD: 1st (PHRED = 21.30) VEP: Promoter / CADD: 7th (PHRED = 10.04)

Table 4.4: H3K27ac master peak overlaps with computational prioritised SNPs. Cell-type specific master peak position information is provided along with name and position of overlapping SNP. Overlap in public data annotation provides information for SNPs identified in VEP, RegulomeDB, and CADD. *exm-rs8070463 only overlaps the CD8 master peak for this region as it extends further downstream.

4.2.2.2 Annotation of *NPEPPS* region using cell-type specific ATAC master peak

The bedtools intersect function was also used to identify overlaps between the prioritised SNP set and cell type specific ATAC master peaks. A total of 35 master peaks for ATAC were identified across CD4, CD8, and CD14 cells in the fine-mapped window for the 96 prioritised SNPs (Chr17:45479446-45786452). ATAC peaks were more cell-type specific than ChIP peaks with only nine of 15 regions being shared across all three cell-types. Of the 15 regions, 10 were overlapped by prioritised SNPs (Table 4.5 and Figure 4.2).

A CD8-specific ATAC peak was located in the intergenic region between *NPEPPS* and *KPNB1*. This region overlapped *rs4375701*, a SNP identified in a known enhancer by VEP annotation. Two regions were identified as uniquely present in CD14 cells. Firstly, a region located in an intron of *NPEPPS* that is tagged by two SNPs *rs9635763* and *rs12603290* and overlap an enhancer regulatory feature identified by VEP analysis. Secondly, a promoter region identified by *rs8065669* upstream of *KPNB1*.

When comparing across both experimental datasets, 5 regions shared peaks from the ATAC-seq and H3K27ac ChIP-seq data. These align with the TSS of genes in this region (Figure 4.2).

ATAC master peak position on chr17	Cell-Type	SNP	SNP position on chr17	Overlap in public data annotation
45649599-45650665	CD14	*rs9635763* *rs12603290*	45649694 45650196	VEP: Enhancer VEP: Enhancer
45711499-45712945	CD8	*rs4375701*	45712218	VEP: Enhancer
45724665-45725717	CD14	*rs8065669*	45724996	VEP: Promoter
45736984-45738506 45736999-45738267 45736866-45738294	CD4 CD8 CD14	*rs4264433*	45737275	
45736984-45738506 45736999-45738267 45736866-45738294	CD4 CD8 CD14	*rs7219303*	45737701	RegulomeDB: 9th (Category = 2b)
45741229-45742383 45741340-45742377 45741203-45743760	CD4 CD8 CD14	*rs56226803* *rs67919208* *rs62074055*	45742233 45771873 45771933	VEP: Promoter flanking region VEP: Promoter / CADD: 1st (PHRED = 21.30) VEP: Promoter / CADD: 7th (PHRED = 10.04)
45780173-45781750 45780747-45782255	CD4 CD14	*rs4439799* *rs4561509*	45781599 45781799	 CADD: 4th (PHRED = 15.20)

Table 4.5: ATAC master peak overlaps with computational prioritised SNPs. Cell-type specific master peak position information is provided along with name and position of overlapping SNP. Overlap in public data annotation provides information for SNPs identified in VEP, RegulomeDB, and CADD.

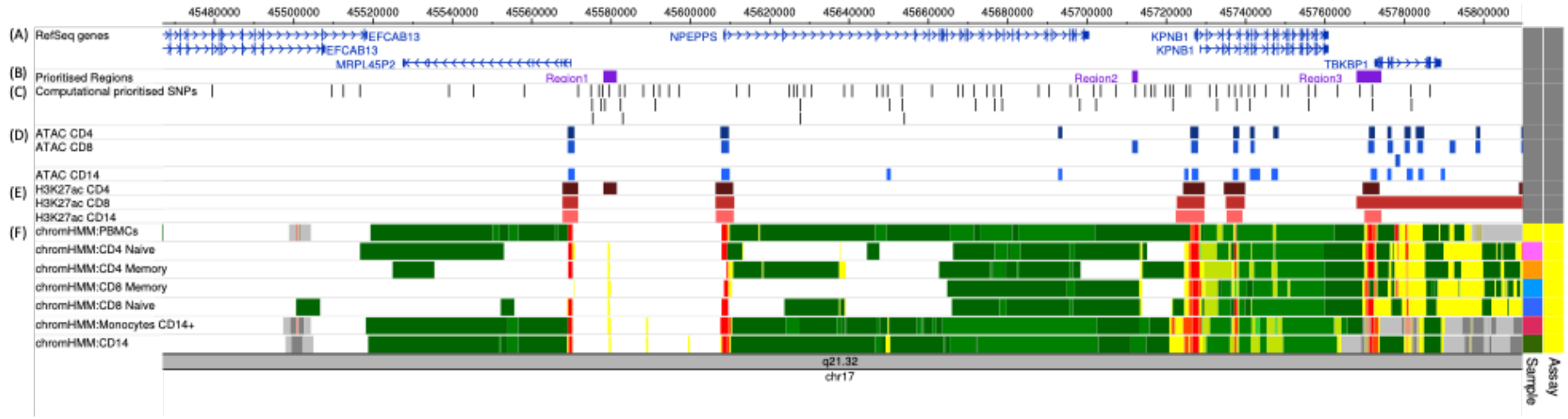


Figure 4.2: Chromatin landscape of the *NPEPPS* region and regions selected for further functional validation. Screenshot from the WashU epigenome browser displaying (A) RefSeq genes, (B) regions selected for further functional validation, (C) computationally prioritised SNPs, (D) in-house ATAC master peaks in CD4, CD8, and CD14 cells, (E) H3K27ac master peaks in CD4, CD8, and CD14 cells, (F) ChromHMM chromatin-states for relevant cell-types (green = strong transcription, dark green = weak transcription, yellow = enhancers, yellow green = generic enhancer, red = active TSS, orange red = flanking active TSS, grey = weak repressed PolyComb, dark grey = repressed PolyComb, white = quiescent/low). The prioritised regions selected for functional follow up are highlighted in purple.

4.2.3 Regions selected for functional follow-up

Ten ATAC peak and six H3K27ac peak regions were identified where variants in the prioritised SNP set overlapped with in-house master peak regions (Figure 4.2). Five regions had overlaps with prioritised SNPs in a single cell-type. Of these unique regions, three were selected as the focus of functional follow up experiments to confirm their functional role.

Region 1 (Chr17:45578178-45581570): A putative enhancer region upstream of *NPEPPS* (Figure 4.3).

This 3.4 kb region was identified by a cell-type specific H3K27ac mark in CD4 cells. It overlaps two prioritised SNPs (*rs11869940* and *rs62076549*) both identified as overlapping a CTCF binding site by VEP. This region also overlaps with a chromHMM enhancer mark in CD4, CD8, and CD14 cells from NIH Epigenomics Roadmap data, providing increased confidence that this region contains a functional element.

Region 2 (Chr17:45711499-45712945): A putative enhancer in the intergenic region between *NPEPPS* and *KPNB1* (Figure 4.4).

This 1.5 kb region was identified by the presence of a CD8-specific ATAC peak. The region overlaps one prioritised SNP (*rs4375701*) that was annotated as an enhancer by VEP and is in close proximity to a chromHMM enhancer mark seen in CD4 and CD8 cells.

Region 3 (Chr17:45768075-45774361): The promoter region of *TBKBP1* (Figure 4.5).

This 6.3 kb region was identified by its dense collection of ATAC and H3K27ac master peaks in all cell types. It overlaps multiple prioritised SNPs (*rs67919208*, *rs62074055*) identified as a promoter regulatory feature by VEP, the second rank SNP by both statistical methods (*rs8070463*), and a known eQTL for *TBKBP1* (*rs6503804*). This region is also heavily annotated with enhancer and promoter marks in chromHMM data.

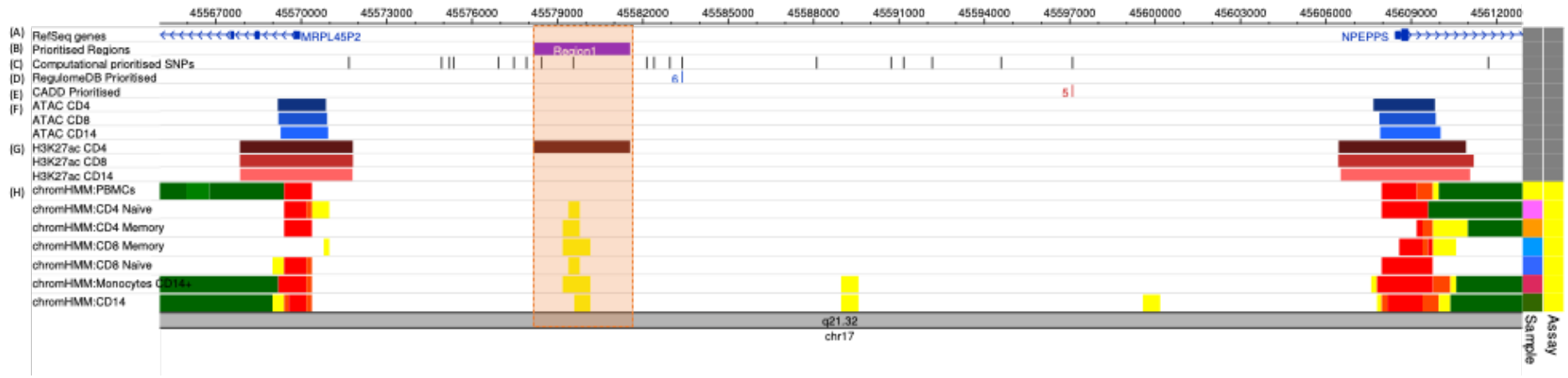


Figure 4.3: Annotation of *NPEPPS* region around prioritised Region 1. Screenshot from WashU epigenome browser display (A) RefSeq genes, (B) region selected for further functional validation containing two SNPs (*rs11869940* and *rs62076549*), (C) computationally prioritised SNPs, (D) RegulomeDB ranked SNPs, (E) CADD ranked SNPs, (F) in-house ATAC master peaks in CD4, CD8, and CD14 cells, (G) in-house H3K27ac master peaks in CD4, CD8, and CD14 cells, (H) ChromHMM chromatin-states for relevant cell-types (green = strong transcription, dark green = weak transcription, yellow = enhancers, red = active TSS, orange red = flanking active TSS, white = quiescent/low). The prioritised region 1 selected for functional follow up is highlighted in the orange box. Prioritised Region 1 overlaps two SNPs from the computationally prioritised SNP set, a CD4 specific H3K27ac ChIP-seq master peak, and enhancer chromHMM marks.

.

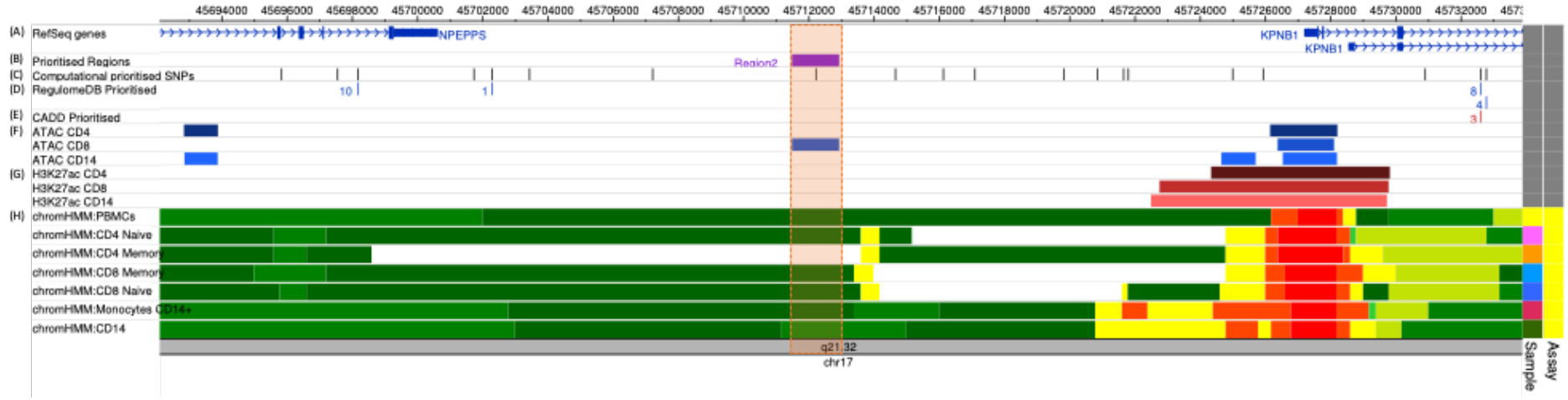


Figure 4.4: Annotation of *NPEPPS* region around prioritised Region 2. Screenshot from WashU epigenome browser display (A) RefSeq genes, (B) region selected for further functional validation containing one SNP (*rs4375701*), (C) computationally prioritised SNPs, (D) RegulomeDB ranked SNPs, (E) CADD ranked SNPs, (F) in-house ATAC master peaks in CD4, CD8, and CD14 cells, (G) in-house H3K27ac master peaks in CD4, CD8, and CD14 cells, (H) ChromHMM chromatin-states for relevant cell-types (green = strong transcription, dark green = weak transcription, yellow = enhancers, yellow green = generic enhancer, red = active TSS, orange red = flanking active TSS, white = quiescent/low). The prioritised region 2 selected for functional follow up is highlighted in the orange box. Prioritised Region 2 overlaps on SNPs from the computationally prioritised SNP set, a CD8 specific ATAC-seq master peak, and is proximal to enhancer chromHMM marks in CD4 and CD8 cells.

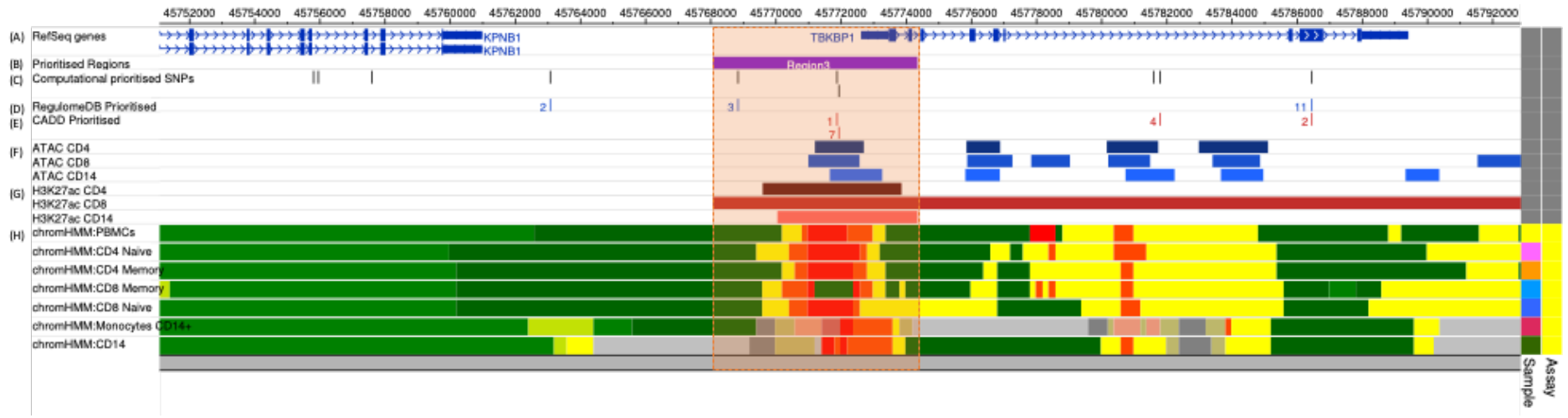


Figure 4.5: Annotation of *NPEPPS* region around prioritised Region 3. Screenshot from WashU epigenome browser display (A) RefSeq genes, (B) region selected for further functional validation containing three SNPs (*rs8070463*, *rs67919208* and *rs62074055*), (C) computationally prioritised SNPs, (D) RegulomeDB ranked SNPs, (E) CADD ranked SNPs, (F) in-house ATAC master peaks in CD4, CD8, and CD14 cells, (G) in-house H3K27ac master peaks in CD4, CD8, and CD14 cells, (H) ChromHMM chromatin-states for relevant cell-types (green = strong transcription, dark green = weak transcription, yellow = enhancers, yellow green = generic enhancer, red = active TSS, orange red = flanking active TSS, grey = weak repressed PolyComb, dark grey = repressed PolyComb, white = quiescent/low). The prioritised region 3 selected for functional follow up is highlighted in the orange box. Prioritised Region 3 overlaps three SNPs from the computationally prioritised SNP set, ATAC-seq and H3K27ac ChIP-seq master peaks in all cell-types, the top ranked CADD SNP, and promoter and enhancer chromHMM marks in all cell-types.

4.3 Discussion and conclusion

The *NPEPPS* region comprising SNPs prioritised by computational statistical methods was annotated using publicly available tools and databases and in-house cell-type and disease specific data to further prioritise causal SNPs and identify regions for functional investigation.

Publicly available tools, including Ensembl VEP, were used to categorise the 96 prioritised SNPs and identify overlaps with known functional regions using multiple public databases. The lead SNP for the *NPEPPS* region, *rs4794057*, was identified as the only variant to overlap a coding region and was categorised as a synonymous variant in exon 8 of the *TBKBP1* transcript ENST00000361722.3. The remaining 95 SNPs were located in intronic or intergenic regions. Of these non-coding variants, 20 overlapped with 15 known regulatory features as annotated by VEP.

Following the ranking of SNPs from the output of RegulomeDB and CADD, there was minimal correlation between the results of the two tools. As a result, these ranking approaches did not prove useful in further prioritisation as an independent approach. Both tools are built on large heterogenous datasets from many sources and cell-types and the output of the scoring system is indicative of any annotation within the dataset. As regulation is known to be cell-type, context, and disease specific (Andersson et al., 2014, Heintzman et al., 2009) these ranking systems are therefore not ideal for the prioritisation of SNPs that are associated in this disease. AS shows enrichment in immune and gut cell-types (Li

et al., 2017) and the use of these tools may be improved by limiting the data used for annotation to specific cell-types for the calculation of scores and ranking. The advantage of using cell-type specific data for annotation is demonstrated in the chromHMM tracks and in the in-house data used for annotation. These forms of annotation were of the greatest value when selecting regions for functional follow-up.

Overlaps were identified between the prioritised SNP set and in-house cell-type specific master peak data from ATAC-seq and H3K27ac ChIP-seq of AS patients and healthy controls. When this data was analysed from the context of disease status no differential peaks were identified. Additionally, capture-C experiments were performed from the viewpoint of the *NPEPPS* promoter and *TBKBP1* promoter in the Knight lab by Dr Ola Mielczarek. These experiments did not identify any clear interactions for functional follow up and have not been included in this study.

ATAC peaks map chromatin accessibility and the results across the three cell-types used in this study demonstrate areas of consistent open chromatin across all cell-types demonstrated by the nine share regions in the fine-mapping window. Unique cell-type specificity peaks were also observed in some areas of open chromatin supporting the idea of regulatory elements being cell-type specific. H3K27ac peaks are associated with active and poised enhancers. For the in-house dataset these marks remained relatively consistent across the three cell-types and were located proximal to TSSs of genes. A unique H3K27ac peak

was observed in CD4 cells in the intergenic region upstream of *NPEPPS* which overlapped with enhancer marks from chromHMM data.

Due to this cell-type and disease specificity of the in-house data more weight was put behind these results when selecting regions for functional follow-up. With the addition of the supporting evidence from cell-type specific chromHMM data and functional annotation from VEP prioritised SNPs that overlapped functional DNA were successfully identified.

Although this approach has led to the identification of putative regulatory regions for functional follow up this approach could be improved on. As mentioned above, selecting specific cell-types in the publicly available datasets would improve the specificity of the annotations leading to refinement of the output impact scores. These tools would then offer a convenient and accessible approach to further prioritise SNP sets following statistical fine-mapping. The ideal dataset for annotation is cell-type and disease specific, the use of in-house data in this annotation approach demonstrated the advantage of this data for the selection of putative causal variants. With the use of a mark for open chromatin and a mark for active enhancers the power of the in-house data to identify putative regulatory regions in a cell-type specific manner has been demonstrated and used to improve on the fine-mapping of this region. For AS the ideal sampling site would be at the point of inflammation at the sacroiliac joint and joints of the spine. This location is a challenge to sample due to the invasive nature of any procedure. A potential alternative would be to use *in vitro* activated cells from the AS patient cohort.

This prioritisation approach could be improved further with the inclusion of additional datasets that would allow for the confirmation of the enhancer regions, marked by H3K27ac, and the identification of alternative regulatory regions. H3K4me1 could be included to confirm enhancer regions, H3K4me3 could be used for the identification of promoters, and these marks could be used in combination to distinguish between enhancers and promoters in close proximity. Work is on-going in the Knight lab to produce additional data. With a sufficiently large sample set this data could be used to identify not only cell-type specific peaks but differential peaks between AS patients and healthy controls to provide a true disease and cell-type specific regulatory landscape ideal for annotation and fine-mapping.

This approach has led to the successful identification of multiple putative regulatory regions to follow up with functional experiments. By identifying overlaps with regions containing known regulatory elements this annotation has improved on the fine-mapping from statistical approaches alone enabling the SNP set to be prioritised further and the selection of regions that have a putative functional role in disease. Cell-type specific and disease specific data has proved most useful as it is of greater relevance in the context of AS. However, the annotation provided through publicly available datasets, specifically VEP annotation and chromHMM tracks, had merit in the initial categorisation of variants and to increase support for the selection of specific regions.

The approach applied in this chapter has enabled an original list of 96 potentially causative SNPs to be greatly reduced and three regions incorporating six SNPs, with putative evidence of function, to be selected for experimental follow up (Chapter 5).

5 Gene expression of *NPEPPS* and *TBKBP1*

5.1 Introduction

As discussed in Introduction section 1.3.1, gene expression is the process by which the information stored as DNA is processed through transcription and translation to produce a functioning protein. This process is regulated at the transcriptional level by TFs, enzymes, and proteins that bind and interact with the DNA to initiate transcription at areas of open chromatin. Variation in the specific DNA sequence at the regulatory elements where these factors bind can have an impact on the regulation of gene expression. This chapter will explore if the gene expression levels of two genes, *NPEPPS* and *TBKBP1*, differ with AS disease status. *NPEPPS* and *TBKBP1* have previously been implicated as the causal gene at this region (Ellinghaus et al., 2016, International Genetics of Ankylosing Spondylitis Consortium et al., 2013, Evans et al., 2011) and are captured within the fine-mapping genomic window identified by frequentist and Bayesian analysis in Chapter 3. It will also address whether the SNPs identified in Chapter 3 and prioritised in Chapter 4 are associated with any genotype-specific change in gene expression.

5.1.1 Assays for gene expression

Gene expression can be profiled to see if it varies under different conditions. Multiple experimental techniques have been developed to measure gene expression at the transcriptional level. These techniques allow for the quantification of gene expression and the comparison between different samples and/or conditions. Here, RT-qPCR will be used to measure expression of *NPEPPS* and *TBKBP1* in PBMCs from AS patients and healthy controls.

DNA microarrays can measure the expression of huge numbers of genes in parallel. DNA microarrays consist of a solid surface, or chip, which has multiple spots of specific DNA sequences, or probes, bound to it. These probes are specific to target regions of cDNA in the sample and if the target cDNA is present in a sample it will hybridize to the probe under specific conditions and release a signal that is used for detection. The strength of the signal is dependent on the amount of target DNA hybridised to a specific probe and therefore the amount of cDNA in the original sample. This technique relies on prior knowledge of the target genes/genome to design the probe sequences. Limitations of this approach include a high background level of detection which contributes to poor low-level detection and limited detection range. It is also challenging to compare expression levels across different experiments.

An alternative approach for genome wide quantification of gene expression is RNA-seq. In this approach, expression is quantified by the count of sequencing reads that align to a specific locus on the genome. This approach enables genome-wide

analysis and the ability to quantify for genes, specific transcripts and specific exons. It has the advantages of not being limited to prior knowledge of the genome enabling the investigation of unknown genomes through de novo transcriptome assembly as well as the detection of alternative splicing events. This approach improves on the previous technique through the accuracy and specificity of results, enabling single base resolution of transcriptional features. It also has higher sensitivity for genes expressed at low or very high levels.

RT-qPCR can be used to quantify gene expression for a single gene or handful of genes. In this process, RNA is converted to cDNA by reverse transcription. This is followed by a qPCR reaction in which a PCR product for a specific gene of interest is amplified and detected in real time at every cycle of the reaction. Two main detection methods have been developed for quantification of the PCR product at each cycle. Firstly, an intercalating dye method where a fluorescent dye such as SYBR green is incorporated into the double stranded PCR product and the fluorescent signal is detected. Secondly, a probe-based method where a sequence specific nucleotide probe, consisting of a reporter and quencher molecule, binds to complementary DNA in the PCR product and during elongation the probe is degraded and releases the reporter for fluorescent detection. For both methods, the amount of fluorescence detected in real-time at each cycle is proportional to the amount of PCR product at a specific cycle. The number of cycles required for a samples fluorescent signal to reach a set threshold can be used to calculate the amount of target specific material in the original sample. This approach has higher sensitivity and can detect lower

levels of RNA. Like microarrays, its major limitation is that the specific sequence of the target gene must be known to enable the design of specific PCR primers.

5.1.2 RT-qPCR using SYBR green technology

In this study the intercalating dye method was used in the RT-qPCR set-up using the SYBR Green PCR mastermix (Applied Biosystems, REF# 4309155). SYBR Green increases in fluorescence when bound within double-stranded DNA without inhibiting the PCR. During each PCR cycle, the dye will bind to newly formed amplicons and the fluorescence intensity will increase in proportion to the increase in PCR product. This fluorescence signal can be used to calculate the amount of target cDNA present in a sample which is proportional to the amount of mRNA.

The disadvantage of SYBR green technology is that the dye will bind to all double-stranded DNA produced in the reaction and is not specific to the target PCR product. Therefore, the PCR must be designed to be highly specific to the product of interest. This is achieved through the design of specific PCR amplification primers and the optimisation of the RT-qPCR reaction conditions to only generate the single target PCR product. Post-PCR QC checks, such as melt-curve analysis, can identify whether products are specific or due to primer dimerisation.

5.1.3 Primer Design for RT-qPCR

Specific primers were designed to amplify the protein coding transcripts of the genes of interest, *NPEPPS* and *TBKBP1*, and three reference genes (*GAPDH*, *HRPT1*,

RPL32). These three reference genes were selected due to their previous use in RT-qPCR experiments in the Brown group, and their being widely used 'housekeeper' reference genes for RT-qPCR experiments due to their stable expression in a wide range of cell types. As previously discussed in Material and Methods, primers were manually designed from Ensembl GRCh37 (hg19) transcript alignments and checked using primer-BLAST (Ye et al., 2012) to target all protein coding transcripts of *NPEPPS* (4 primer sets) and *TBKBP1* as well as the reference gene *HPRT1*. *NPEPPS* required a combination of 4 primer pairs to capture all protein coding transcripts as no consensus sequence was shared across them all. In-house primers for reference genes *GAPDH* and *RPL32* were also used. The genomic locations of *NPEPPS* and *TBKBP1* primers as well as primer sequences for all primers are provided in Material and Methods (section 2.4.4). Primer pairs were designed to span between exons to mitigate against any genomic DNA contamination. The following optimal primer design criteria were adhered to in order to allow for all RT-qPCR reactions to occur in the same reaction conditions on the same reaction plate:

- GC content: 50-60%
- Melting temperature (Tm): 50-65oC
- Avoid secondary structures
- Avoid 3` complementarity (primer-dimer)
- Avoid G or C repeats longer than 3 bases
- Include a G or C in the last 3 base pairs of primers
- Ensure predicted products are specific to target (primer-BLAST)

5.1.4 Data analysis using $2^{-\Delta\Delta Ct}$ method for RT-qPCR

For RT-qPCR there are two common methods for data analysis, absolute quantification and relative quantification. Absolute quantification determines the input copy number of the target RNA by comparing the PCR signal to a standard curve of known quantities, whereas relative quantification compares the expression of the gene of interest to that of a reference gene or group of reference genes. In this study the relative quantification approach was used to compare expression between different conditions, specifically AS patients and healthy controls.

Relative changes in gene expression were analysed from RT-qPCR experiments using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001) as described in Chapter 2 section 2.4.6. In brief, the raw C_t values for the target gene are normalised to the reference gene to calculate ΔC_t for each sample. The $\Delta\Delta C_t$ is then calculated for each target gene by normalising the mean ΔC_t for AS patients to the mean ΔC_t for healthy controls. This value is then used to calculate the fold expression change observed in AS patients compared to healthy controls.

5.1.5 Aim

Aim 3: To characterise the causative genetic variants through experimental approaches to validate their functional role in disease.

3.1 To use RT-qPCR to assess whether the expression levels of the genes in the *NPEPPS* locus differ between AS patients and healthy controls.

5.2 Results

5.2.1 AS case and healthy volunteer cohort

A total of 140 AS patients and 74 healthy volunteers were included in the RT-qPCR gene expression study (detailed in Materials and Methods Chapter 2.4). Peripheral venous blood was collected from AS patients and healthy volunteers. PBMCs were isolated and stored in liquid nitrogen prior to this study. RNA extractions for AS patient samples were performed by previous members of the Brown lab. For healthy volunteer samples, RNA was extracted from 2 million PBMCs per sample using the QIAsymphony RNA extraction programme “RNA_CT_400”. These samples were used as templates for cDNA synthesis. Briefly, 1 µg RNA was used in a 20 µL reaction using the Bioline Tetro cDNA synthesis kit. 1 µL cDNA was used as template for each RT-qPCR reaction.

5.2.2 RT-qPCR experimental optimisation and primer efficiencies

An assumption of the $2^{-\Delta\Delta Ct}$ method used for data analysis of the RT-qPCR data is that amplification efficiencies of the target genes are equal to the amplification efficiencies of the reference genes. To test whether this assumption was true for this RT-qPCR experiment the efficiency for each primer pair was calculated. A dilution series of cell-line cDNA was used as template material in the optimised RT-qPCR reaction and the mean C_t value at each dilution point calculated. This was then plotted against concentration and a linear regression generated for the dilution series

for each primer set. The amplification efficiency was calculated using the PCR efficiency calculation recommended in the MIQE guidelines (PCR efficiency = $10^{-1/slope}$-1) (Bustin et al., 2009). All primer sets for target and reference genes had a primer efficiency between 90 and 110% (Supplementary Table 5.1) so were suitable for use with the $2^{-\Delta\Delta Ct}$ method.

5.2.3 Sample QC during RNA extraction and cDNA synthesis

RNA quantity and quality checks were performed by spectrophotometry using the NanoDrop (NanoDrop 2000, ThermoFisher Scientific). A 260/280 ratio of ~2.0 is generally accepted as optimal for RNA, with a range from 1.8 to 2.2 used as the cut off in this study. No RNA samples were excluded due to falling outside this range. The recommended input volume for cDNA synthesis was 1 µg RNA. If sufficient material was not available, the volume required for 500 ng RNA was calculated and used. Three AS samples were removed from the study due to insufficient material for cDNA synthesis. All samples, where cDNA synthesis was performed, were carried forward to RT-qPCR.

5.2.4 Sample QC following RT-qPCR

Following all RT-qPCR experiments, initial quality control steps were performed on the raw output data using the QuantStudio Design and Analyse software (ThermoFisher Scientific). Amplification curves were assessed to identify any non-conforming triplicates or unusual curves which may indicate failed reactions. Melt

curves were checked to make sure no secondary products or primer dimers were present in the reactions.

Subsequent to this initial data quality check, the raw C_t values were analysed to check uniformity in triplicates. An acceptable range for variance in triplicate C_t values was established as 2 x standard deviation observed in the cell line across all plates. This was applied to each sample as the percentage of overall raw C_t the 2 x standard deviations represents. A total of 191 samples successfully generated usable data that passed the initial quality control steps. The total consisted of 125 AS case samples and 66 healthy control samples.

During an initial review of the data, a bimodal distribution in the normalised ΔC_t values for AS cases was observed (see Figure 5.1). This correlated with two batches of cDNA synthesis, using the same protocol but Batch 2 using new reagents and being prepared by a different researcher. Strategies were discussed to attempt to include both batches using meta-analysis to combine the datasets. However, since all healthy control samples were prepared in the second batch of cDNA synthesis there were no matched control samples for comparison. Consequently, Batch 1 was removed from further analysis.

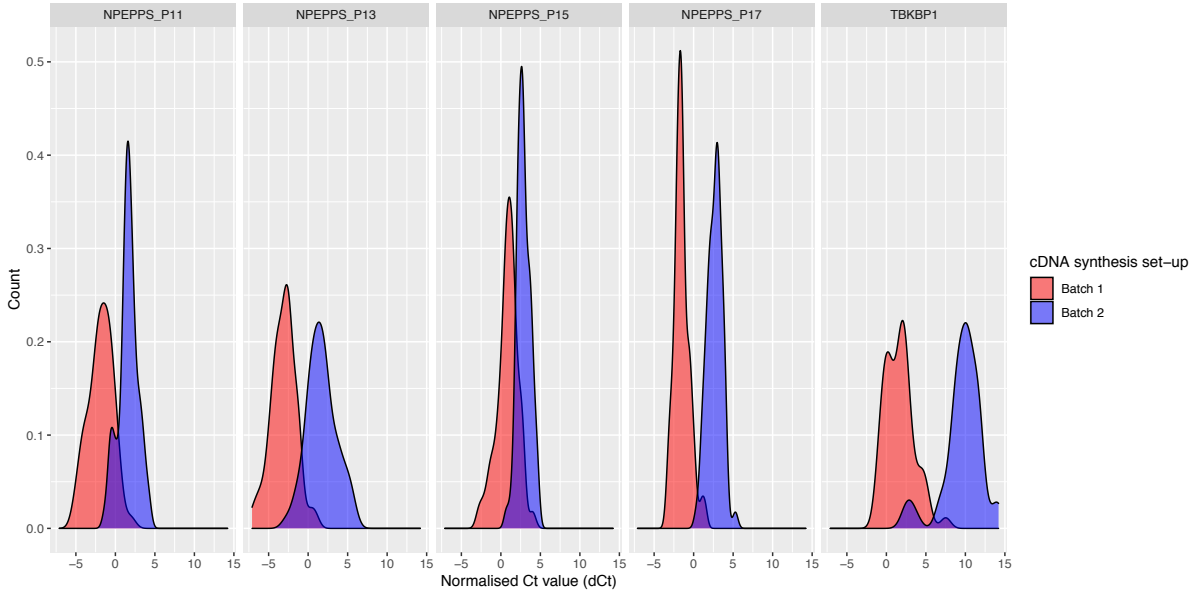


Figure 5.1: Batch effects in AS cases due to cDNA synthesis preparations differences. (A) Distribution of normalised ΔC_t values showing bimodal spread when categorised by cDNA synthesis batch. Bimodal distribution was not observed in healthy controls where samples were all prepared in the same cDNA synthesis batch (Batch 2).

This led to 134 samples (70 AS cases and 66 healthy controls) being carried forward to final analysis.

5.2.5 Selection of appropriate reference genes for normalisation

Three reference genes were included in the experimental design to be used in data analysis for normalisation. Reference genes should be stably expressed and their abundance should have strong correlation with the total amount of RNA present in the sample (Bustin et al., 2009). If all three genes are acting as good reference genes it is expected that the general trend in raw C_t for each sample would be mirrored across all reference genes. The three genes selected were; *GAPDH*, *RPL32*, and *HPRT1*.

Analysis of reference gene expression across all samples (AS = 70, HV = 66) was performed in order to determine which reference genes to use in later analysis (Figure 5.2).

Mean C_t values of *RPL32* were most consistent across samples. *RPL32* has the smallest range in C_t values (range = 6.4) and consistency between median and mean, as well as the lowest standard deviation. Both *GAPDH* and *HPRT1* perform well but have a larger spread in their data (range = 10.1 and 9.6, respectively). When the dataset is split by disease status the mean and median C_t values remain consistent. This is indicative of good reference genes. This pattern can be seen in the summary statistics (Table 5.1).

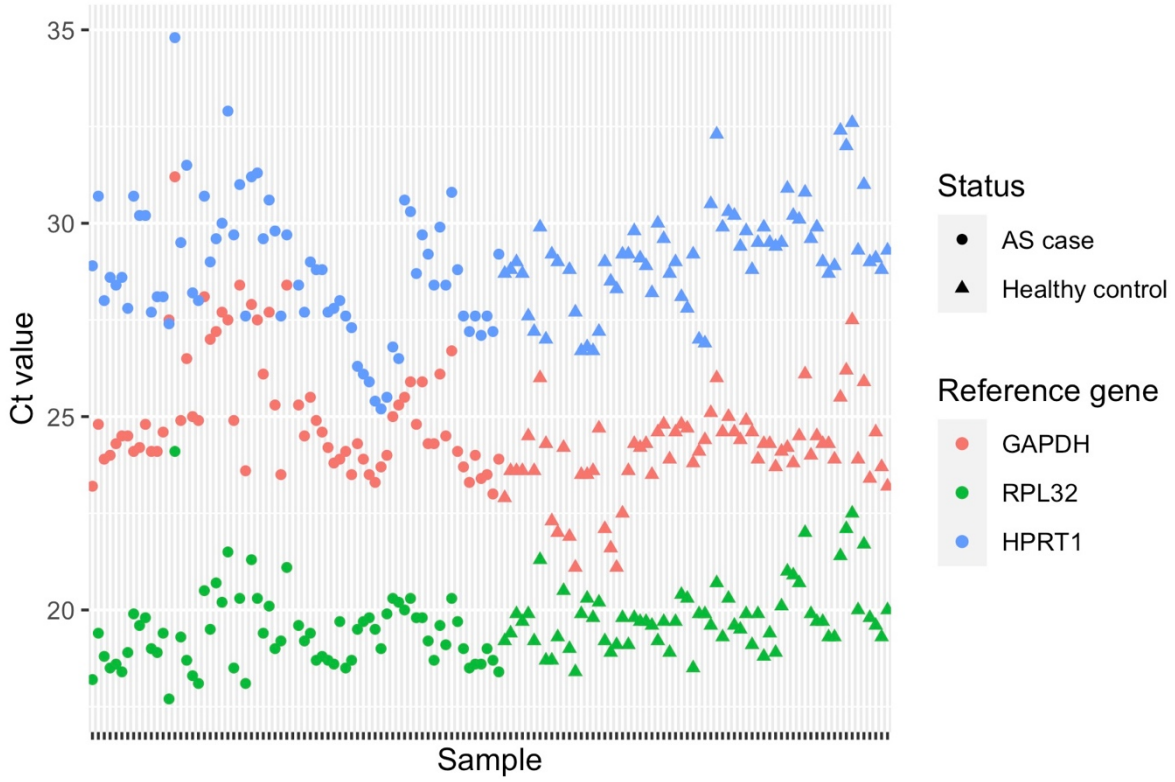


Figure 5.2: Scatterplot showing the mean C_t values for three reference genes across all sample. *RPL32* C_t values are the most consistent across all samples. AS cases are plotted on the left of the graph as circles and healthy volunteers on the right as triangles.

	GAPDH			RPL32			HPRT1		
	ALL	**Case**	**Control**	**ALL**	**Case**	**Control**	**ALL**	**Case**	**Control**
Count	136	70	66	136	70	66	136	70	66
Min	21.1	23.0	21.1	17.7	17.7	18.4	25.2	25.2	26.7
Max	31.2	31.2	27.5	24.1	24.1	22.5	34.8	34.8	32.6
Range	10.1	8.2	6.4	6.4	6.4	4.1	9.6	9.6	5.9
Median	24.3	24.5	24.2	19.6	19.3	19.7	29.0	28.6	29.1
Mean	24.6	25.1	24.1	19.6	19.4	19.8	28.9	28.7	29.2
SD	1.50	1.62	1.17	0.94	0.98	0.85	1.57	1.76	1.32

Table 5.1: Summary statistics for C_t values of reference genes across all samples.

Pairwise comparisons were made between each combination of reference genes. Better correlation was observed between *RPL32-HPRT1* and *GAPDH-HPRT1* than between *GAPDH-RPL32* (see figure 5.3(C), (B), and (A) respectively). PCA was performed on the reference gene to see if the data would separate based on disease state (see figure 5.3(D)). There was no clear segregation by status, although there was a slight trend to separate on the vertical axis. *RPL32* has the smallest influence on separation in this axis whereas *HPRT1* and *GAPDH* appear to have an equal and opposing effect. This indicates that *RPL32* is more stable across all samples regardless of disease status.

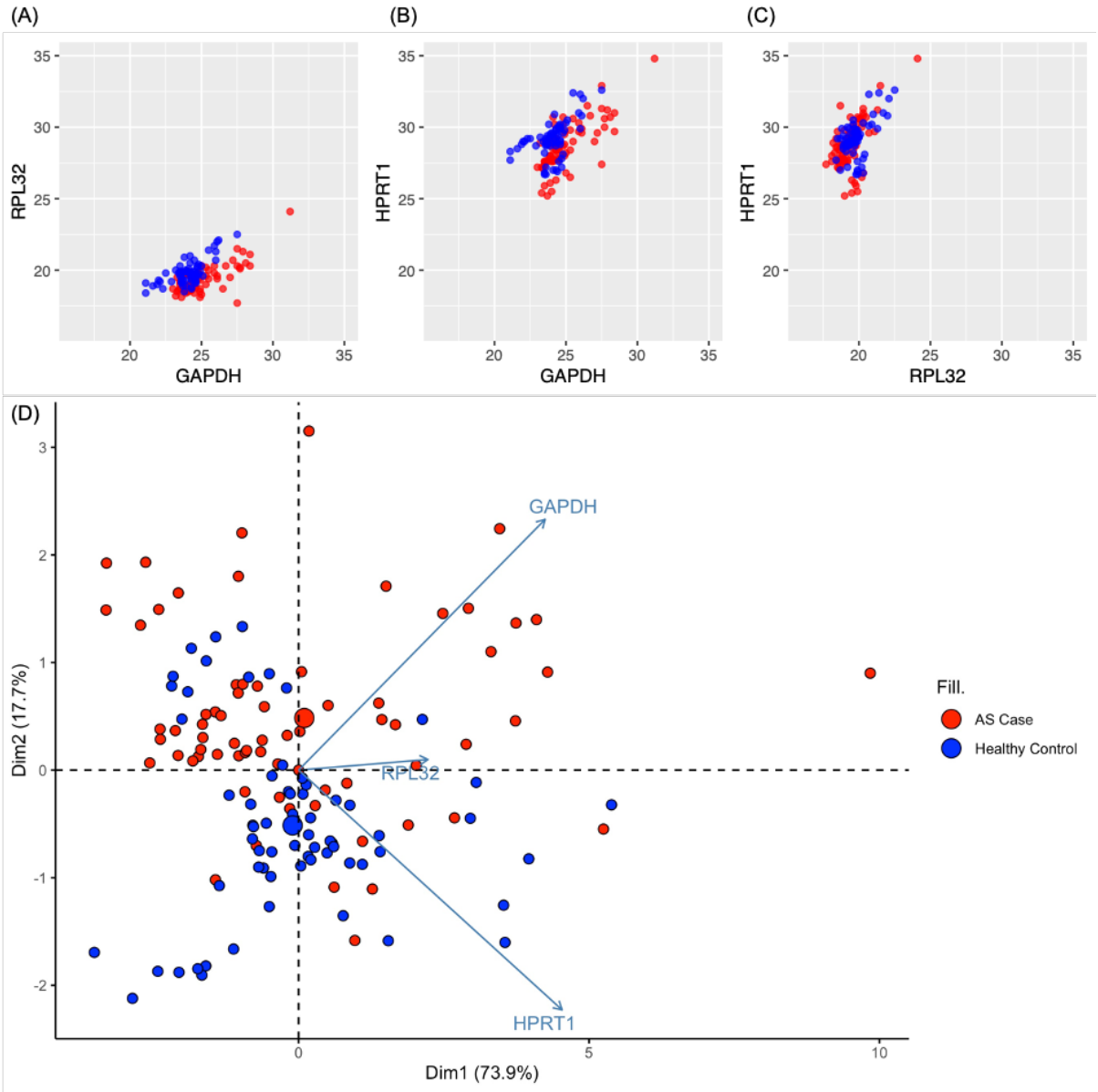


Figure 5.3: Comparison of correlation observed in reference gene data. (A-C) Comparison of the AS cases and healthy control C_t values for each sample in pairs of reference gene. Greater correlation is observed for the combination of *RPL32* and *HRPT1*. (D) PCA of reference genes displaying the separation for the 1st and 2nd dimensions. AS case and healthy controls cluster well with some separation in the 2nd dimension.

Following this analysis, *RPL32* and *HPRT1* were selected as reference genes as this combination included a high expressor (*RPL32*) in combination with a low expressor (*HPRT1*). *GAPDH* was excluded as a reference gene due to its greater range in data. It is critical for all samples to have data points for both reference genes. One AS case sample and one healthy control were removed from the dataset due to missing data points, resulting in 134 samples in the final analysis (AS cases = 69, healthy controls = 65).

5.2.6 Gene expression of *NPEPPS* and *TBKBP1* in AS cases and healthy controls

The $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001) for RT-qPCR data analysis was used to analyse the gene expression of *NPEPPS* and *TBKBP1* relative to the geometric mean of the reference gene set (*RPL32* and *HPRT1*). For each sample the ΔC_t value and a one-way ANOVA were performed to compare the mean ΔC_t for all AS cases to the mean ΔC_t for all healthy controls. The fold expression change was calculated using the $\Delta\Delta C_t$ value for each primer set using the equation $2^{-\Delta\Delta Ct}$, as described in the methods.

5.2.6.1 *NPEPPS* gene expression is significantly lower in AS cases

NPEPPS expression was detected using four separate primer sets that span different combinations of exons (Figure 5.4(A) and Figure 2.2(A) for transcript specific locations). Lower expression of *NPEPPS* (i.e. higher ΔC_t values) was observed in AS cases compared to healthy controls across all 4 primer sets (Figure

5.4(B) and (C)). Mean ΔC_t values for each primer set were compared between AS cases and healthy controls at each primer set by way of one-way ANOVA. The difference between groups was significant at each primer set (see Figure 5.4(B)). At primer NPEPPS_P13, AS cases had a mean ΔC_t of 2.15 (SD = 1.83) compared to the healthy control mean ΔC_t of -0.21 (SD = 1.43, P = 5.6×10^{-13}). Mean ΔC_t for AS cases was 3.19 (SD = 0.99) compared with healthy control mean ΔC_t of 0.97 (SD = 0.50, P < 2.2×10^{-16}) for NPEPPS_P17, and 2.08 (SD = 1.17) compared to 0.74 (SD = 1.13, P= 2.9×10^{-6}) for NPEPPS_P11. Finally, at primer NPEPPS_P15 the mean ΔCt for AS cases was 3.35 (SD = 0.84) compared to a mean ΔC_t of 2.69 (SD = 0.91, P = 1.2×10^{-9}) in healthy controls.

The fold expression change was calculated from healthy controls to AS cases at each primer pair with a 0.20 (0.07-0.53) fold decrease for NPEPPS_P13, 0.23 (0.16-0.32) for NPEPPS_P17, 0.40 (0.18-0.88) for NPEPPS_P11, and 0.64 (0.34-1.20) for NPEPPS_P15 (see Figure 5.4(C) and Supplementary Figure 5.2 for log2(fold change)).

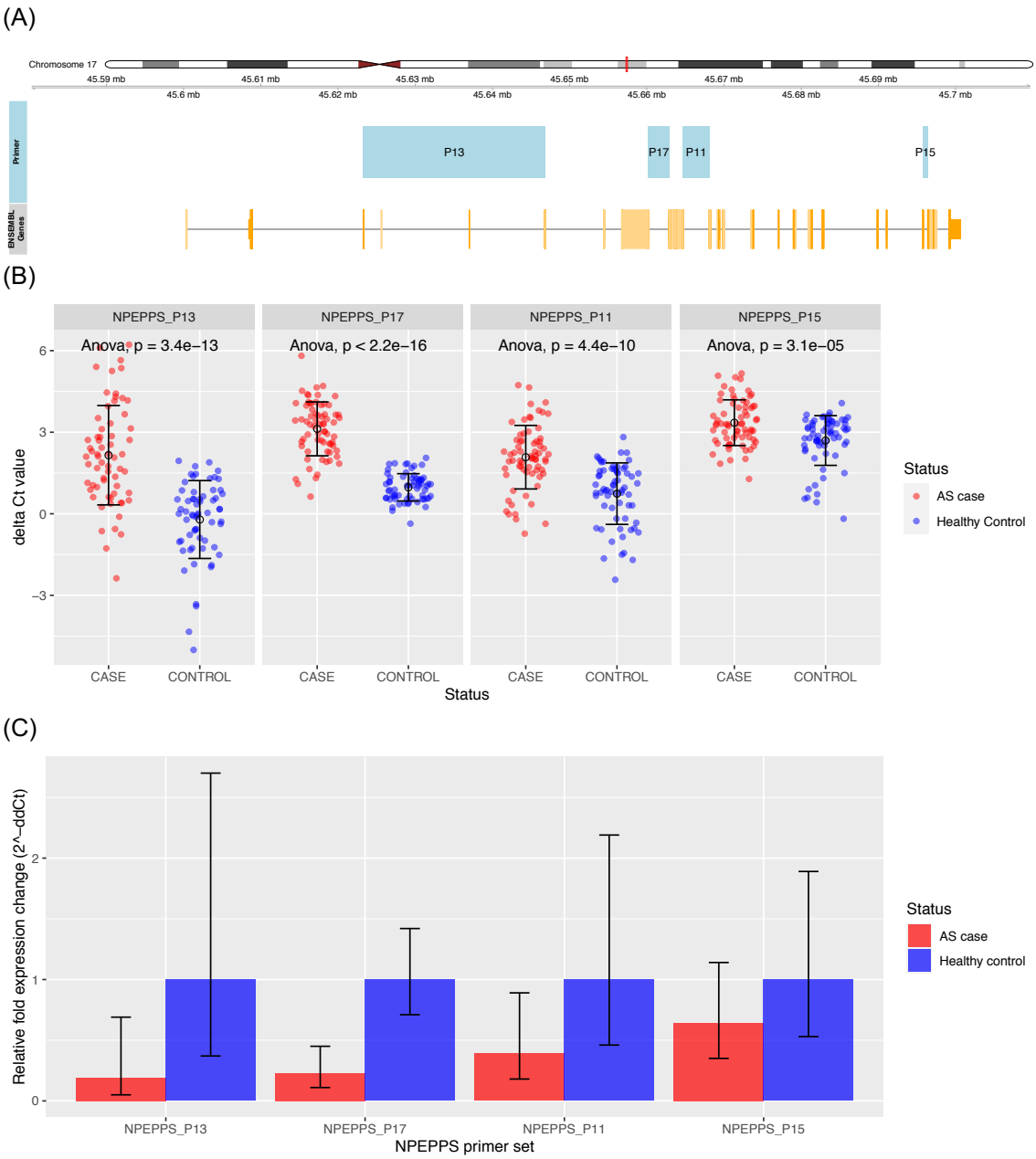


Figure 5.4: *NPEPPS* gene expression is lower in AS cases (n = 69) compared to healthy controls (n = 65). *NPEPPS* mRNA expression was quantified by RT-qPCR using four primer pairs at different locations to capture all protein-coding transcripts. (A) Location of amplicons from four primer sets (blue) on *NPEPPS* gene structure (orange). (B) Mean ΔC_t values for AS cases and healthy controls normalised to reference genes (*RPL32* and *HPRT1*) for each primer pair. Status-specific ΔC_t mean and standard deviation shown. The p values were determined by one-way ANOVA, with a significant difference between groups observed at all primer pairs. (C) Relative fold expression change of mean AS cases compared to mean healthy controls per primer set as calculated by $2^{-\Delta\Delta Ct}$. Standard deviation shown.

5.2.6.2 *TBKBP1* gene expression is significantly lower in AS cases

When comparing *TBKBP1* expression levels between AS cases and healthy controls, considerably lower expression is observed in AS cases (Figure 5.5). When comparing the mean ΔC_t values between AS cases (mean ΔC_t = 10.12, SD = 2.39) and healthy controls (mean ΔC_t = 4.58, SD = 1.15) by way of one-way ANOVA, the difference between groups was significant ($P < 2.2 \times 10^{-16}$, see Figure 5.5(B)). The expression fold change was calculated from healthy controls to AS cases with a 0.02 (0.01-0.05) fold decrease (see Figure 5.5(C) and Supplementary Figure 5.2 for log2(fold change)).

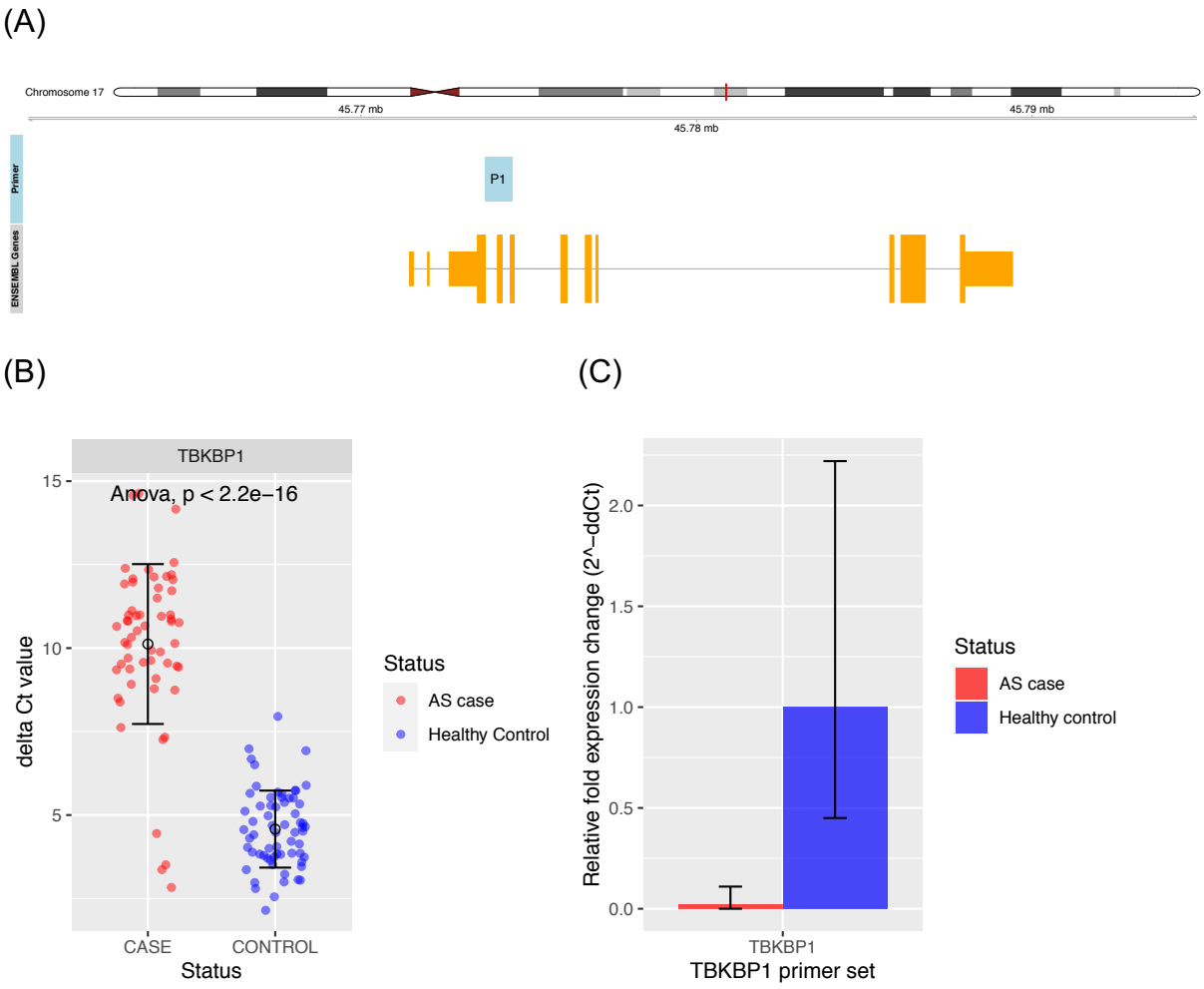


Figure 5.5: *TBKBP1* gene expression is significantly lower in AS cases (n = 69) compared to healthy controls (n = 65). *TBKBP1* mRNA expression was quantified by RT-qPCR. (A) Location of primer pair on *TBKBP1* gene structure. (B) Mean ΔC_t values of AS cases and healthy controls normalised to reference genes (*RPL32* and *HPRT1*). Status-specific ΔC_t mean and standard deviation shown. The p values were determined by one-way ANOVA, with a significant difference between groups observed ($P < 2.2x10^{-16}$). (C) Relative fold expression change of mean AS cases compared to mean healthy controls as calculated by $2^{-\Delta\Delta Ct}$. Standard deviation shown.

5.2.7 Genotype specific analysis of genome-wide significant SNPs

Genotype data was available from Illumina CoreExome data for tested SNPs from 65 AS cases and 26 healthy control samples (four AS cases and 41 healthy controls had no genotype data available at the time of analysis). Expression quantitative trait loci (eQTL) analysis was performed to determine the effect of the genotype of all 96 SNPs, prioritised in Chapter 3, on expression of the proximal genes *NPEPPS* and *TBKBP1*. Covariables including age and sex were not controlled for in this eQTL analysis.

5.2.7.1 No genotype-specific eQTL effect was observed for *NPEPPS* mRNA levels

The effect of genotype was assessed by way of eQTL analysis to see if there was correlation between a specific genotype and expression levels. AS cases and healthy controls were grouped by genotype and the difference between groups assessed by one-way ANOVA. Due to the strong LD observed across all 96 SNPs analysed ($R^2 \geqq 0.68$), the trends observed across genotypes were consistent across all SNPs with only minor variation seen between SNPs. The prioritised SNP, *rs11869940* ($p = 4.68 \times 10^{-10}$, OR = 1.12, risk allele = T), in the putative enhancer region upstream of *NPEPPS* was used to assess any potential genotype-specific eQTL effect on *NPEPPS* gene expression.

When comparing the genotype-specific expression of *NPEPPS* for all samples at *rs11869940* (TT = 15, TA = 54, AA = 20), no significant difference was observed between groups although samples appear to group by status (see Figure 5.6 (A)). This may be due to opposing effects in AS cases and healthy controls, or effects seen only in either cases or controls (as has previously been reported for the nearby gene *TBX21* in AS (Lau et al., 2017)) masking any detectable signal, and therefore the analysis was repeated in AS cases and healthy controls separately. When analysis was restricted to AS cases only (rs11869940; TT = 12, TA = 42, AA = 10), a trend was observed for higher expression (lower ΔC_t) in samples homozygous for the AS risk allele compared to samples homozygous for the protective allele (see Figure 5.6 (B) and Supplementary Table 5.2). The opposite trend was observed when the dataset was restricted to healthy controls (rs11869940; TT = 3, TA = 12, AA = 10). Lower expression (higher ΔC_t) was observed in samples homozygous for the AS risk allele compared to those homozygous for the protective allele ($P>0.05$). These non-significant trends were observed for the *NPEPPS* primers sets NPEPPS_P13, NPEPPS_P17, and NPEPPS_P11 but were not as clear for NPEPPS_P15. The lack of significant results is potentially due to these analyses being underpowered due to small cohort sizes.

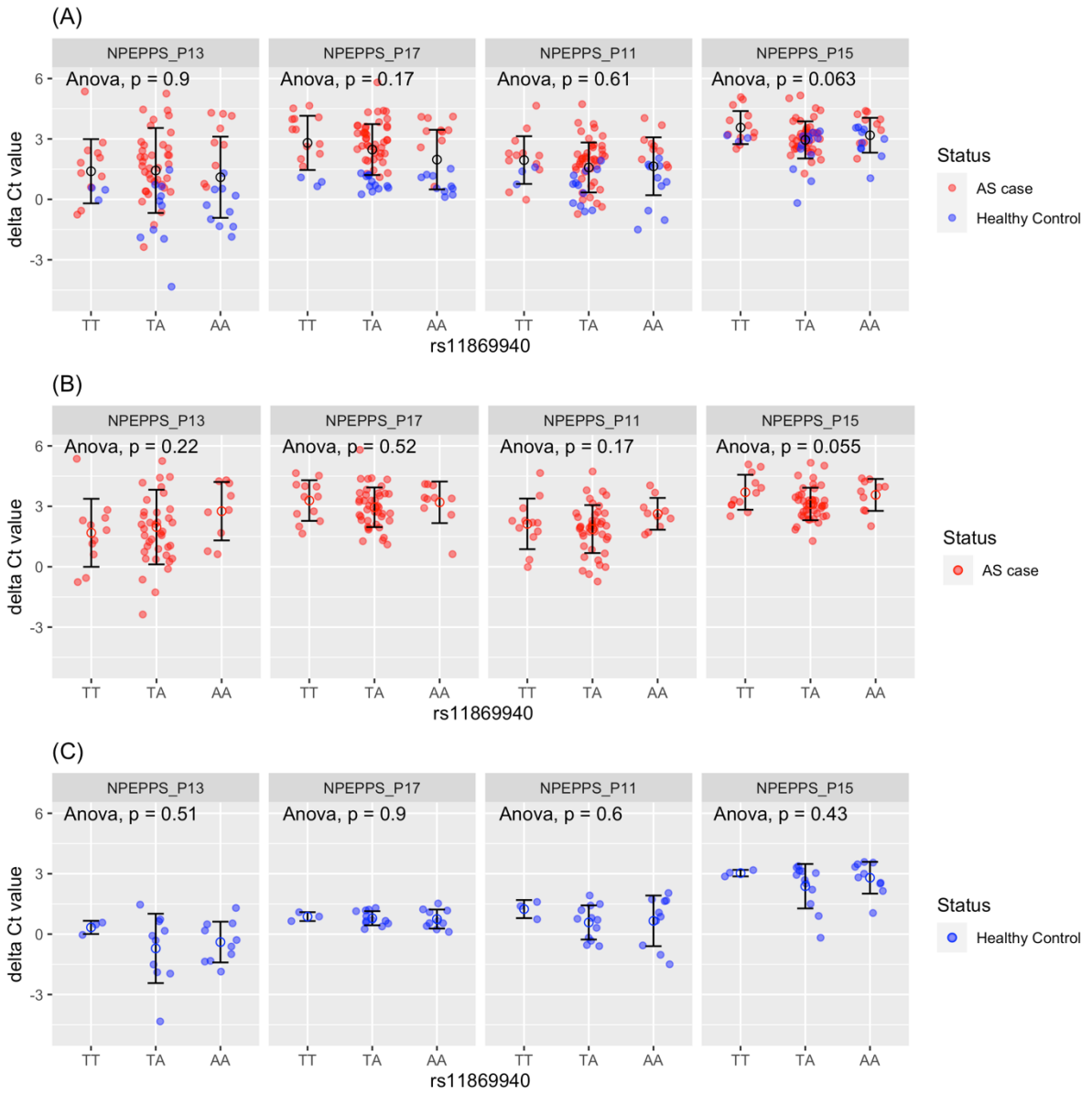


Figure 5.6: eQTL analysis of gene expression of *NPEPPS* show no significant difference between genotypes for *rs11869940* at all four *NPEPPS* primer sets. (A) *rs11869940*-specific mean ΔC_t values for AS cases and healthy controls (risk allele = T) across all 4 *NPEPPS* primer sets. (B) *rs11869940*-specific mean ΔC_t values restricted to AS cases showing trend for higher expression (lower ΔC_t) in AS cases homozygous for risk allele (TT) compared to protective allele (AA). No significant difference observed. (C) *rs11869940*-specific mean ΔC_t values restricted to healthy controls showing trend for higher expression (lower ΔC_t) in samples homozygous for protective allele (AA) compared to risk allele (TT) the opposite trend to that observed in AS cases. No significant difference observed. Mean ΔC_t and standard deviation provided. P values calculated using one-way ANOVA.

5.2.7.2 No genotype-specific eQTL effect was observed for *TBKBP1* mRNA levels

LD had the same impact when comparing the results of any eQTL effect on *TBKBP1* expression. The prioritised SNP, *rs67919208* ($p = 3.28x10^{-11}$, OR = 1.13, risk allele = C), in the *TBKBP1* putative promoter region was used to exemplify the genotype-specific eQTL effect on *TBKBP1* gene expression. When comparing the genotype-specific expression of *TBKBP1* for all samples (rs67919208; CC = 15, TC = 50, TT = 15), no significant difference was observed between groups although a clear separation between AS cases and healthy controls is observed (see Figure 5.7 (A) and Supplementary Table 5.3). As with *NPEPPS*, this lack of eQTL signal may be due to opposing effects in AS cases and healthy controls masking any detectable signal, so analysis was repeated in status specific cohorts. When analysis was restricted to AS cases only (rs67919208; CC = 12, TC = 34, TT = 9), no pattern of differential expression was observed between genotype groups (see Figure 5.7 (B) and Supplementary Table 5.3). No significant difference was observed. When the dataset was restricted to healthy controls (rs67919208; CC = 3, TC = 16, TT = 6), lower expression (higher ΔC_t) was observed in samples homozygous for the AS risk allele compared to those homozygous for the protective allele. This trend was not shown to result in significant differences between groups. Again, the reduced cohort size may have reduced the ability to detect any significant results.

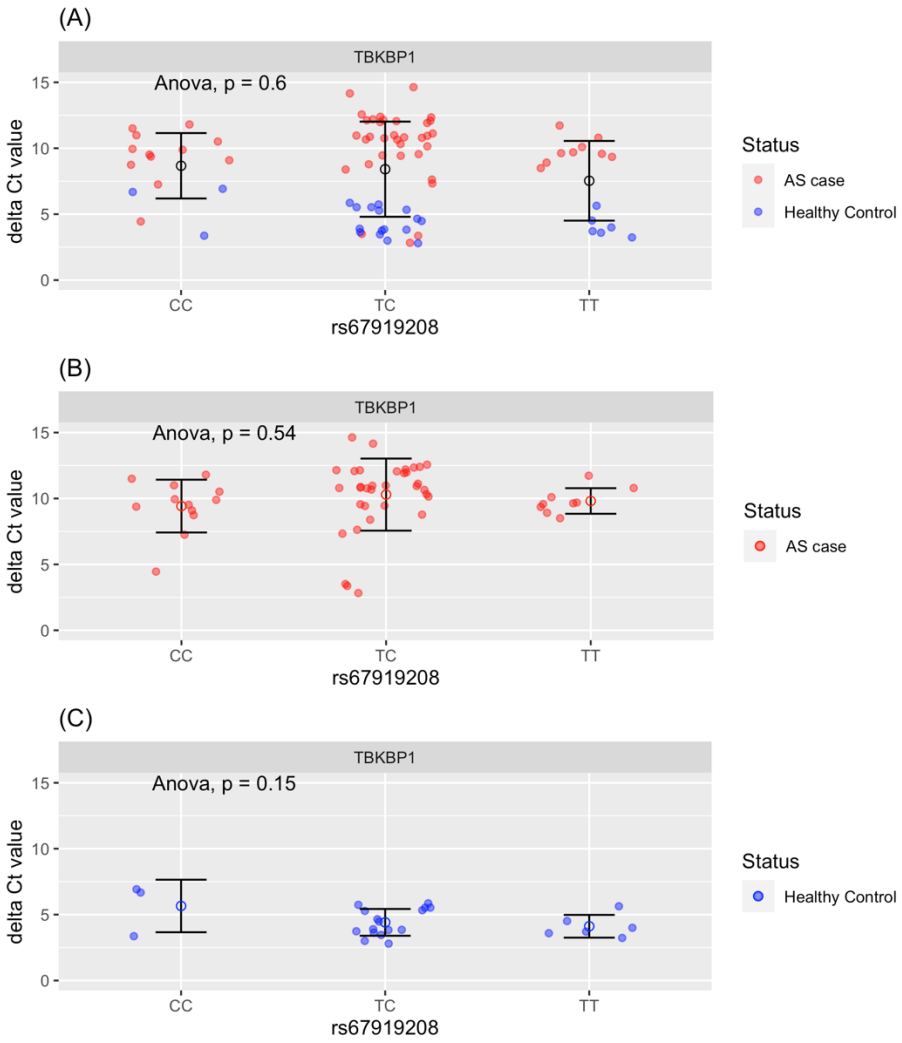


Figure 5.7: eQTL analysis of gene expression of *TBKBP1* shows no significant difference between genotypes for *rs67919208*. (A) *rs67919208*-specific mean ΔC_t values for AS cases and healthy controls (risk allele = T) for *TBKBP1* primer set. (B) *rs67919208*-specific mean ΔC_t values restricted to AS cases shows no obvious trend across genotypes. No significant difference observed. (C) *rs67919208*-specific mean ΔC_t values restricted to healthy controls showing trend for higher expression (lower ΔC_t) in samples homozygous for protective allele (TT) compared to risk allele (CC). No significant difference observed. Mean ΔC_t and standard deviation provided. P values calculated using one-way ANOVA.

5.3 Discussion and conclusion

RT-qPCR experiments were performed to quantify the relative level of *NPEPPS* and *TBKBP1* gene expression levels in AS cases and healthy controls in PBMCs. Significantly lower relative mRNA expression was observed in AS cases for both genes. This further implicates these two genes in AS and motivates further research to identify the regions that regulate their expression, as these regulatory regions are likely to harbour causal variants.

Lower expression of *NPEPPS* and *TBKBP1* in AS cases compared to healthy controls may indicated potential functional roles in the mechanism of AS. Reduced *NPEPPS* expression may lead to a loss of active puromycin-sensitive aminopeptidase in the cytoplasm available to trim amino acids from its substrates. This could lead to a change in the levels and diversity of peptides, of a specific length, crossing into the ER to be processed further by ERAP1 and ERAP2 to be presented on the cell surface by MHC class I molecules. Reduced *TBKBP1* expression in AS cases is likely to impacted on TBK1 activity and its role in the TNF/NFKB inflammatory pathway. Reduction in *TBKBP1* expression is likely to reduce the availability of its encoded adaptor protein and therefore reduce the activity of TBK1 as a protein kinase or drive its interaction with alternative adaptor proteins including TANK and NAP1 (Zhu et al., 2019). It is also plausible that this observation is specific to the sampling cell-type, PBMCs, and not implicated in disease mechanism.

The ability to accurately quantify the levels of *NPEPPS* mRNA in a sample was limited by the large number of predicted *NPEPPS* transcripts (n = 25) without a consensus region to design primers. This challenge was met by the use of four primer pairs designed to target all protein-coding transcripts which theoretically captured 20 of the 25 predicted transcripts. This approach proved successful as the same pattern for lower relative expression in AS cases was observed across all four primer pairs. Variation was observed across the four primers, this is likely a result of the positioning of the primers capturing different combinations of *NPEPPS* transcripts and their abundance differing in the samples. Primer pair NPEPPS_P15 is located at the 3`end of *NPEPPS*, at this site of quantification the highest mean ΔC_t values, corresponding to the lowest relative mRNA expression, is observed for both AS cases and healthy controls suggesting transcripts captured by this primer set as less abundant in both cohorts. An alternative approach that would remove this limitation would be to use RNA-seq to assess gene expression. This approach is a lot more expensive per sample but enables genome-wide analysis and the ability to quantify genes, transcripts and specific exons from the same dataset. This was not feasible for this current study but would be of great benefit if utilising a similar fine-mapping strategy to investigate multiple disease-associated loci.

A major limitation of this RT-qPCR study was the removal of 45% of AS case samples during the QC steps of the experiment due to inconsistency in cDNA synthesis step. Efforts were made to batch-correct for these samples but as no healthy control samples were prepared by the same cDNA synthesis procedure,

as all were processed with the second batch, it was not possible to analyse the two batches separately and combine them in post-analysis. The loss of a large proportion of the test data could have had a substantial negative impact on the power of this study to identify any difference between the two groups. However, the removal of 45% of the samples resulted in a ratio of roughly 1:1 with 69 AS cases and 65 healthy controls (total = 134). This was a sufficient sample size to detect a mean ΔC_t change of 1 between cases and control with 95% confidence and 80% power.

No eQTLs for *NPEPPS* or *TBKBP1* were identified for any of the 96 genome-wide significant SNPs when all samples were analysed in combination. In AS cases, a pattern was observed for samples homozygous for the risk allele to have higher expression of *NPEPPS* (lower ΔC_t). An opposite pattern was observed for healthy controls, with lower expression (higher ΔC_t) in samples homozygous for the AS risk allele. Neither of these patterns were significant. For *TBKBP1* expression, no pattern was observed in AS cases, but in healthy controls lower expression was observed in samples homozygous for the AS risk allele. The lack of a significant result may be due to this analysis being under-powered when restricted by disease status (AS cases = 65, healthy controls = 24) with low numbers of samples in groups limiting the ability to detect the difference between genotypes. To detect a medium effect size (0.25) with 95% confidence and 80% power a sample size of more than 50 samples per group is required. This estimation is confirmed by Huang and colleagues (Huang et al., 2018) who suggest typical eQTL studies have a sample size of between 50 and

1000 samples with the Genotype-Tissue Expression (GTEx) project have a sample size between 70 and 361 for its dataset of 44 tissues. A further potential reason for the lack of significant results is that any eQTL effect maybe cell-type specific and masked in this data in the combined cell-types in PBMCs. Replication of this experiment in disease-specific cell-types, as used for SNP prioritisation, may enable the detection of eQTLs in this AS cohort.

The RT-qPCR data is based on RNA extracted from PBMCs and not cell-type specific RNA. RNA from PBMCs was used in this study as material from a large cohort of AS cases was available to use from previous work performed by the Brown group. This is a limitation of this study as RNA levels can differ between cell-types and putative causal SNPs may only correlate with differential gene expression in a cell-type specific manner as has been reported previously (Dimas et al., 2009). A better approach to the use of RNA extracted from PBMCs to assess gene expression would be the use of RNA extracted from specific cell types that have been implicated in disease. Furthermore, methods in Chapter 4 prioritised SNPs based on cell-type specific annotation and these SNPs may only have an impact on gene expression in the cell-type they were identified in. The quantification of gene expression in CD4, CD8, and CD14 cells would be of initial interest for future work as these cell-types were used in the prioritisation of putative causal regions in chapter 4 and may offer insight into whether prioritised SNPs correlate in a genotype-specific manner. Multiple studies investigating putative regulatory elements in AS disease-associated loci have investigated the

expression of their SNPs in a cell-type specific context (Vecellio et al., 2016, Roberts et al., 2015).

Following on from these findings, a next step is to investigate the prioritised regions and assess whether these putative regulatory elements can have a regulatory effect on gene expression.

6 Functional follow up by luciferase reporter assays

6.1 Introduction

Following the identification and prioritisation of potentially causative SNPs by computational methods (Chapter 3) and genomic annotation (Chapter 4), three regions were identified as having putative regulatory roles in the *NPEPPS* region. These regions were selected for functional follow up experiments with the aim of confirming their regulatory role. In this chapter, the luciferase assays was used to investigate whether Region 1, a putative enhancer region upstream of *NPEPPS*, Region 2, a putative enhancer in the intergenic region between *NPEPPS* and *KPNB1*, and Region 3, a putative promoter region upstream of *TBKBP1*, exhibit regulatory activity. It then addresses whether this regulatory activity is influenced by the specific alleles of the prioritised SNP(s) in these regulatory regions.

6.1.1 Luciferase reporter assay

The luciferase reporter assay is a common form of transient transfection assay that can be used to define regulatory elements involved at the transcriptional level of gene expression. This approach utilises luciferase, an oxidative enzyme that produces light as it breaks down its substrate luciferin. The putative regulatory element is inserted into the vector upstream of the reporter gene using standard

cloning techniques. On transfection into cultured cells, this putative regulatory element will regulate the expression of the luciferase gene and therefore the amount of luciferase produced. Following cell lysis, the amount of luciferase produced can be quantified through the introduction of luciferin which it will breakdown producing light. The luminescence produced is measured using a luminometer. The luciferase activity can be directly correlated with the activity of the regulatory element as the regulatory element and luciferase gene are physically linked in the vector construct.

Advantages of this approach are high sensitivity with low background measurements, a wide linear range for quantification, relative ease and timeframe to perform, no endogenous activity in the host cells, and interpretation of results is simple as the amount of protein produced is directly proportional to transcriptional activity (Himes and Shannon, 2000). A key limitation of this approach is that the regulatory element is tested in an artificial context of plasmid DNA and not in its native genomic context. This mean the regulatory element under investigation is free from the specific TFs, accessory proteins, and chromatin structure and modifications that act in concert for the sequence to fulfil its function. Consequently, the results of the luciferase assay will not truly reflect the activity of the regulatory element in its specific genomic context but give an indication of it potential to act in a regulatory way.

6.1.2 The Dual-Glo luciferase assay system

In the dual-reporter assay system, activity of two luciferases is sequentially measured in a single reaction, in order to normalise for transfection efficiency. This is

possible using two different luciferase constructs that use different substrates to generate signal.

The Promega Dual-Glo luciferase reporter assay (#E2920) was used in this study. Firefly (*Photinus pyralis*) luciferase signal was used to measure the regulatory activity in the test vector. Renilla (*Renilla reniformis*) luciferase signal was measured as an internal control for data normalisation. Independent vectors containing the two luciferases were co-transfected into cells enabling the measurement of both experimental and control vectors from the same sample.

6.1.3 Reporter constructs for investigating putative promoters and enhancers

Different types of putative regulatory elements can be investigated using specific reporter constructs. For example, putative promoters are inserted into a reporter construct lacking a promoter region. For putative enhancers, a reporter construct with minimal promoter activity should be used. This enables the investigation in changes in level of transcription due to the putative enhancer sequence above that observed in the minimal promoter sequence alone.

Promega have developed commercially available reporter constructs (pGL4 series) for the investigation of putative promoters and enhancers using the Dual-Glo luciferase reporter assay (see Figure 6.1). The pGL4.10 [luc2] is a promoterless luciferase reporter vector and was selected for use to investigate putative promoter regions. An experimental approach was designed to insert putative promoter Region

3 at the *Nhe*I/*Xho*I cut-site in the multiple cloning region (see Figure 6.1 (A)). The pGL4.23 [luc2/minP] reporter vector was selected for the investigation of putative enhancers. This vector contains a minimal promoter region, and the putative enhancer region will be inserted at the *Bam*HI/*Sal*I restriction enzyme site (see Figure 6.1 (B)). This vector also contains sequence for a polyA signal to prevent any transcription initiation of the luc2 gene from the putative enhancer region.

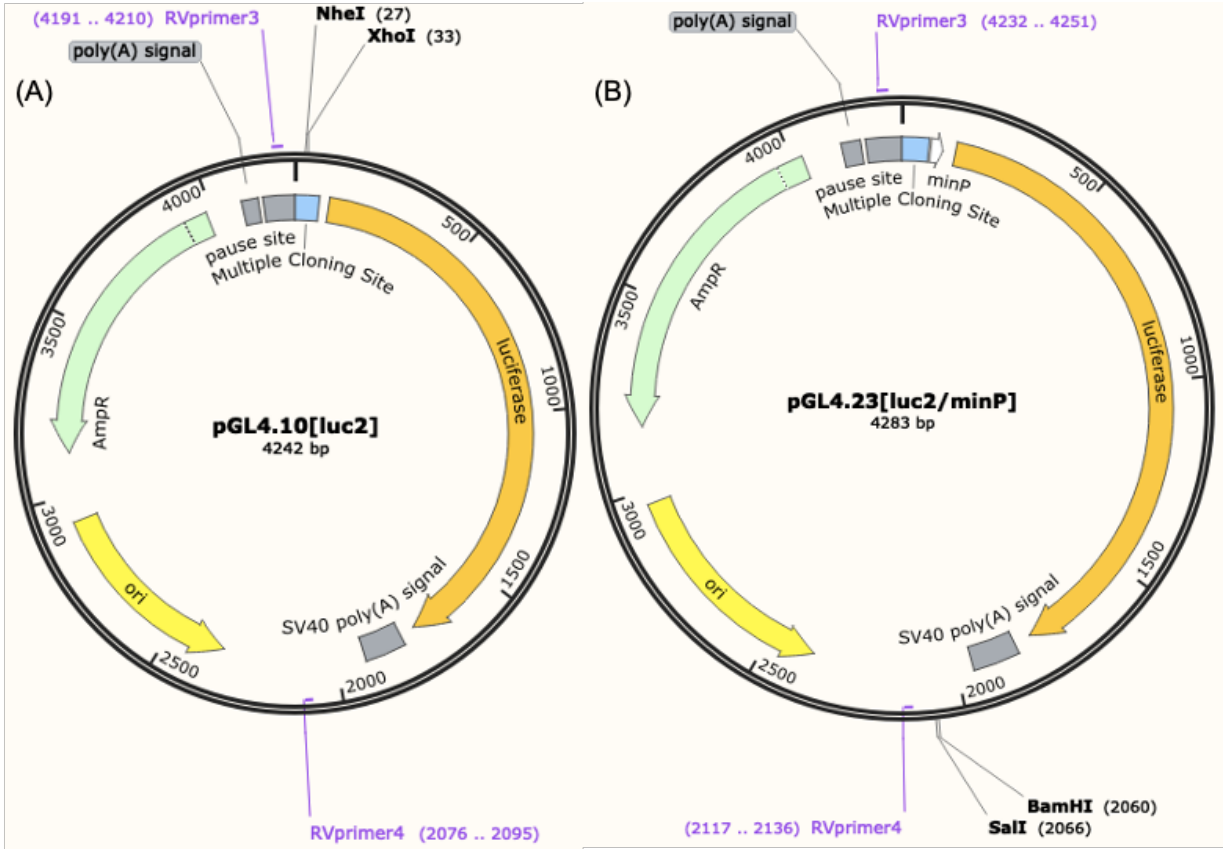


Figure 6.1: Vector maps of Promega PGL4-series reporter constructs used for luciferase assays. (A) Map of pGL4.10 [luc2] promoterless vector highlighting *Nhe*I/*Xho*I cut-site for putative promoter insertion. (B) Map of pGL4.23 [luc2/minP] minimal promoter vector highlighting the *Bam*HI/*Sal*I cut-site for putative enhancer insertion. Multiple cloning site shown as blue box. Minimal promoter shown as white arrow. luc2 reporter gene shown as orange arrow. Synthetic β-lactamase (AmpR) coding region shown as green arrow. SV40 late poly(A) signal and synthetic poly(A) signal/transcriptional pause site shown as grey box. Reporter vector primer 3 and 4 binding regions shown as purple lines. Images generated in SnapGene Viewer.

6.1.4 Prioritised regions for investigation

Three regions were identified in Chapter 4 as putative regulatory elements and were selected for functional investigation using luciferase assays. Region 1: *NPEPPS* putative enhancer (Chr17:45578178-45581570) is a 3.4 kb region defined by the CD4-specific H3K27ac mark contains two prioritised SNPs, *rs11869940* and *rs62076549* (see Figure 4.3). Region 2: *NPEPPS-KPNB1* intergenic putative enhancer (Chr17:45711499-45712945) was defined by a 1.5 kb CD8-specific ATAC peak that overlaps the prioritised SNP, *rs4375701* (see Figure 4.4). This ATAC mark was in close proximity to a chromHMM enhancer mark seen in CD4 and CD8 cells. A 3.0 kb region that captured both the ATAC peak, its prioritised SNP, and neighbouring chromHMM enhancer mark was selected as the Region 2 target insert. Region 3: *TBKBP1* putative promoter (Chr17:45768075-45774361) is a 6.3 kb region was identified by its dense collection of ATAC and H3K27ac master peaks in all cell types (see Figure 4.5). This region overlapped three prioritised SNPs and is heavily annotated with enhancer and promoter marks in chromHMM data. For cloning, a region of 2.7 kb surrounding the smaller ATAC peak was selected as the Region 3 insert to align experimental set-up across the three regions. This region captures two SNPs (*rs67919208* and *rs62074055*) and the transcription start site of *TBKBP1*.

6.1.5 Aim

Aim 3: To characterise the causative genetic variants through experimental approaches to validate their functional role in disease.

3.2 To investigate the regulatory role of specific regions of the *NPEPPS* locus, identified through computational analysis, using luciferase reporter assays.

6.2 Results

Three genomic regions containing prioritised SNPs identified in Chapter 4 were selected for functional follow up using a luciferase reporter assay. These regions were selected because the evidence from annotations was suggestive of a putative enhancer or promoter role.

6.2.1 Haplotype analysis of the *NPEPPS* region

AS patient DNA samples from the GENExpress ID study cohort were used as template to PCR amplify genomic regions of interest to be used for subsequent cloning steps.

Haplotype analysis was performed using LDhap (Machiela and Chanock, 2015). Due to the strong LD across the *NPEPPS* region, discussed in Chapter 3, two haplotypes are present across all three regions under investigation. Genotypes of AS-associated SNPs in these regions are shown in Table 6.1. Two AS patient DNA samples were selected that were homozygous for either the risk (sample ID = AS050) or protective (sample ID = AS051) haplotype, in order to be able to determine the effect of haplotype on enhancer activity in the luciferase assays. In-house genotyping data from the GENExpress ID study were used to call the genotype of AS patient samples. Since *rs11869940* and *rs62076549* SNPs were not directly genotyped, two SNPs in perfect LD to the test SNPs (*rs7222851* and *rs8067286*) were used to infer the test SNPs genotype.

(A)

SNP	Position on chr17	Prioritised region	REF allele	REF allele Freq.	ALT allele	ALT allele Freq.	AS050	AS051
rs7222851	45592169	1	T	0.462	C	0.538	T|T	C|C
rs8067286	45594590	1	A	0.462	G	0.538	A|A	G|G
rs4375701	45712218	2	C	0.543	T	0.457	T|T	C|C
rs8070463	45768836	3	T	0.501	C	0.499	C|C	T|T
rs67919208	45771873	3	T	0.501	C	0.499	C|C	T|T
rs62074055	45771933	3	G	0.513	C	0.487	C|C	G|G

(B)

Haplotype	Frequency	Risk/Protective
C_G_C_T_T_G	0.4841	Protective
T_A_T_C_C_C	0.4463	Risk

Table 6.1: Genotype information for samples used as template to amplify regions of interest. (A) Genotype information for AS050 and AS051 identifying the risk allele and frequency for the SNPs in the three prioritised regions. For region 1, SNPs *rs7222851* and *rs8067286* act as surrogate SNPs for *rs11869940* and *rs62076549*. (B) Frequency of risk and protective haplotypes. Frequency data from LDhap using European reference populations (Machiela and Chanock, 2015).

6.2.2 Generation of pGL4-based constructs containing putative regulatory regions

Region 2: *NPEPPS-KPNB1* intergenic putative enhancer insert was successfully generated by PCR amplification of patient genomic DNA using sequence specific primers (Material & Methods section 2.5.3 and Table 2.3). A single product was obtained as visualised by gel electrophoresis and was ligated into pCR-Blunt II-TOPO. The orientation of the insert and genotype at *rs4375701* was confirmed by Sanger sequencing. Subsequently, these TOPO vectors were used to clone four pGL4.23-based constructs to use in luciferase assays. These pGL4.23 constructs contained Region 2 inserts, combining two haplotypes with forward and reverse orientations of insert (Table 6.1).

The sequence of Region 1 (3.0 kb) has 98% sequence identity to an intergenic region at chr17:36440982-36444381. This shared sequence is a result of *MRPL45P2*, the gene upstream of *NPEPPS*, being a pseudogene of *MRPL45* (mitochondrial ribosomal protein L45) and sharing highly similar sequence upstream of their gene position. To attempt to capture the target region whilst maintaining specificity to the *NPEPPS* upstream region, a nested PCR strategy was designed, employing an initial long-range PCR amplification to produce a product specific to the region (~18kb). The long-range product was then used as the template material for a nested PCR, amplifying the Region 1 specific 3.0 kb product. This approach proved unsuccessful as the long-range product could not be reliably generated.

The PCR amplification of the *TBKBP1* promoter region also proved unsuccessful. The target region is GC rich and therefore difficult to amplify using PCR based methods.

As a result of these challenges to amplify the desired target region, all subsequent work focuses of Region 2 the putative enhancer located in the intergenic region between *NPEPPS* and *KPNB1*.

Test vector name	Template DNA	Risk/protection Genotype	Size (bp)	5` restriction enzyme tag	3` restriction enzyme tag	Reporter vector construct
Region 2: Risk-Fwd	AS050	Risk	7306	*Bam*HI	*Sal*I	pGL4.23 [luc2/minP]
Region 2: Protect-Fwd	AS051	Protective	7306	*Bam*HI	*Sal*I	pGL4.23 [luc2/minP]
Region 2: Risk-Rev	AS050	Risk	7306	*Sal*I	*Bam*HI	pGL4.23 [luc2/minP]
Region 2: Protect-Rev	AS051	Protective	7306	*Sal*I	*Bam*HI	pGL4.23 [luc2/minP]

Table 6.1: pGL4.23-based vectors used in luciferase dual-reporter assay. Template DNA indicates the AS sample genomic DNA used as template material for insert. Risk/Protective genotype status provided. Size (bp) indicates the length of the vector sequence in base pairs. 5` and 3` restriction enzyme tag ascribes restriction enzyme tag sequence. Reporter vector construct used to generate test vector indicated.

6.2.3 Luciferase assay experimental design and data QC

The four pGL4.23-based vectors containing the Region 2: *NPEPPS-KPNB1* intergenic putative enhancer were used in luciferase assays (see Materials and Methods section 2.5.4). Briefly, pGL4 vector appropriate controls (see below) were

transiently co-transfected with pRL-TK into HEK-293T cells using lipofectamine. After a 48-hr incubation, luciferase activity was measured using the Dual-Glo luciferase reporter assay kit (Promega #E2920). Enhancer activity was calculated as the ratio of firefly (from pGL4) to Renilla (from pRL-TK) activity. Data was generated from nine independent experiments. Each run contained the four insert DNA test plasmids (Table 6.1) and six controls (Table 6.2) in triplicate.

Following a quality control check of the raw data, two complete runs were removed from the study. This was due to leakage of luciferase signal from the pGL4-SV40 high expression control into surrounding wells leading to artificially increased RLU readings in surrounding wells that contained other control samples. This prevented accurate analysis of test sample data as normalisation could not be performed to a reliable control reading. Thereafter, the pGL4-SV40 control was positioned in the 96-well plate so that it could not impact any further samples with leakage of signal. One further sample was removed from analysis (Region 2: Risk-Fwd in experiment 2) because it transpired that the incorrect sample had been loaded.

In total 7 sets of triplicate datapoints were generated for all sample conditions except Region 2: Risk-Fwd where 6 sets of triplicates were produced.

Control vector name	Size (bp)	Description
pGL4.23 [luc2/minP]	4283	Firefly luciferase reporter gene *luc2* under control of minimal promoter. Used as a blank vector control for baseline activity.
pGL4.23 [luc2/minP] 1.6kb λDNA	5883	Firefly luciferase reporter gene *luc2* under control of minimal promoter plus stuffer sequence of ~1.6kb λDNA inserted at the *Bam*HI/*Sal*I restriction enzyme site. Used as a size-matched baseline control.
pGL4.23 [luc2/minP] 2.4kb λDNA	6683	Firefly luciferase reporter gene *luc2* under control of minimal promoter plus stuffer sequence of ~2.4kb λDNA inserted at the *Bam*HI/*Sal*I restriction enzyme site. Used as a size-matched baseline control.
pGL4.23 [luc2/minP] 4.0kb λDNA	8283	Firefly luciferase reporter gene *luc2* under control of minimal promoter plus stuffer sequence of ~4.0kb λDNA inserted at the *Bam*HI/*Sal*I restriction enzyme site. Used as a size-matched baseline control.
pGL4.23 [luc2/SV40]	4641	Vector encodes the firefly luciferase reporter gene *luc2* and is designed for high expression. Used as positive control vector.
pRL-TK	4045	Vector encodes the Renilla luciferase *Rluc*. Used as transfection control.

Table 6.2: Control vectors used in luciferase dual-reporter assay. Size (bp) indicates the length of the vector sequence in base pairs. Description ascribes the reporter gene encoded and vector function.

6.2.4 Selection of appropriate control vector: effect of insert size on luciferase activity

It has previously been reported that there is an inverse relationship between relative luciferase activity and the size of the construct (Yin et al., 2005). As the region 2 insert is 3003bp it increases the plasmid DNA length from 4283bp for the blank pGL4.23 [luc2/minP] plasmid to 7286bp when containing the insert. Therefore, the activity of this 7.3kb plasmid should be compared with a size-matched control rather than the empty 4.3kb blank pGL4.23 [luc2/minP] vector.

A series of size control vectors were generated by Dr Carla Cohen as part of a wider project in the Knight group. These were generated by inserting 1.6kb, 2.4kb, and 4.0kb fragments of λ DNA into the pGL4.23 [luc2/minP] vector at the *Bam*HI/*Sal*I restriction enzyme cut-site. This produced plasmid DNA 5883bp, 6683bp, and 8283bp in length, respectively. All three controls, along with the blank pGL4.23 [luc2/minP] vector, were included in the experiments.

The effect of plasmid size on luciferase activity was confirmed in the experimental system used here. The average fold change was calculated from the blank pGL4.23 [luc2/minP] vector to the three controls vectors containing different length stuffer sequence (see Figure 6.2). As expected, inclusion of 1.6kb, 2.4kb, and 4.0kb inserts resulted in a linear and significant decrease in their average relative luciferase activity (0.34, 0.29, and 0.17 respectively). In subsequent experiments, and to ensure that all conclusions made in this study are conservative, the pGL4.23 [luc2/minP] λ2.4kb vector was used for fold change comparisons with test constructs. This results in a comparison of a 7.3kb test vector to a 6.7kb control.

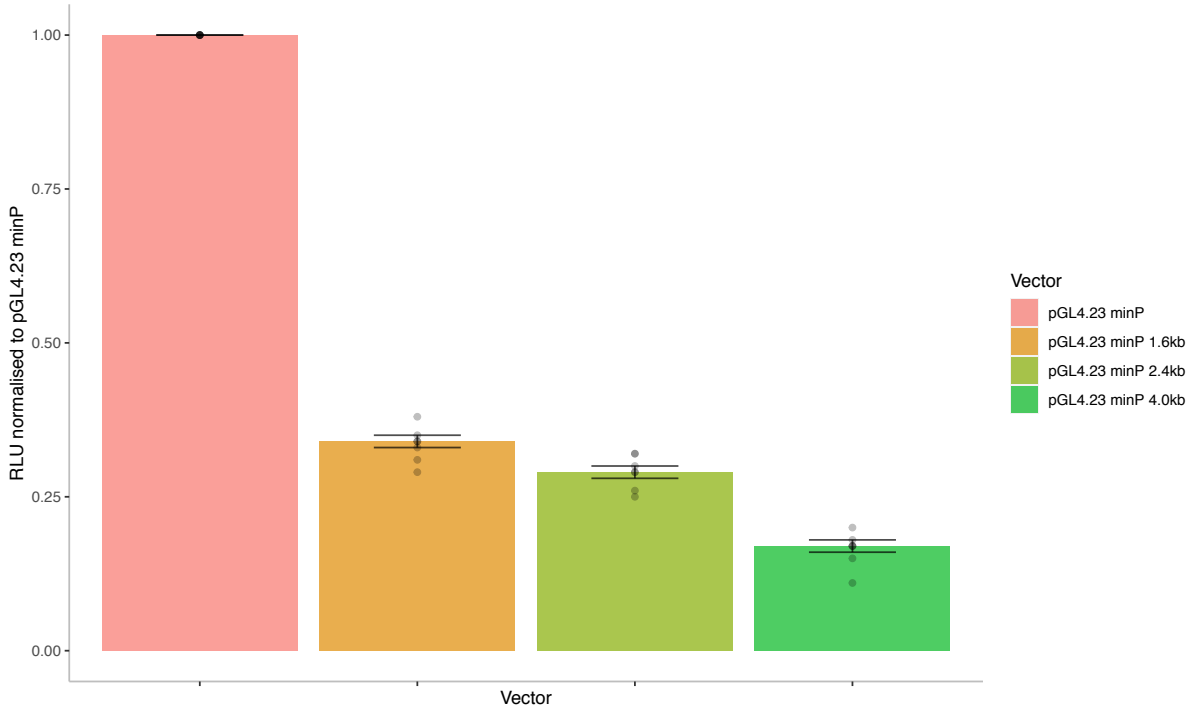


Figure 6.2: Relative luciferase activity of pGL4.23 [luc2/minP] control vectors containing different size inserts. Relative luciferase units are normalised to the pGL4.23 [luc2/minP] vector results for comparison. Standard error of mean shown.

6.2.5 Region 2 insert DNA displays enhancer activity

Region 2, spanning 3.0kb that captures the *rs4375701* in the intergenic region between *NPEPPS* and *KPNB1* exhibits >2.5-fold enhancer activity when compared to a size matched minimal promoter control (pGL4.23 [luc2/minP] λ2.4kb) (Figure 6.3). A one-way ANOVA was performed and showed there was a significant ($p<2.6x10^{-16}$) difference between groups. The fold change increase from pGL4.23 [luc2/minP] λ2.4kb control to test constructs was calculated. Paired t-tests were performed to specifically test if the significant difference held true when comparing each insert to the size matched minimal promoter control. Samples were grouped by experimental run in the paired t-test comparison.

A 3.1 fold increase was observed for Region 2: Risk-Fwd insert (p-value = $9.2x10^{-5}$), 3.5 for Region 2: Protect-Fwd insert (p-value = $3.6x10^{-5}$), 2.7 for Region 2: Risk-Rev insert (p-value = $1.9x10^{-6}$), and 2.8 for Region 2: Protect-Rev insert (p-value = $2.1x10^{-4}$). All four inserts had a greater than 2.5-fold increase in relative luciferase activity when compared to the sized matched control indicating that this region of DNA acts to increase expression in an enhancer type mechanism.

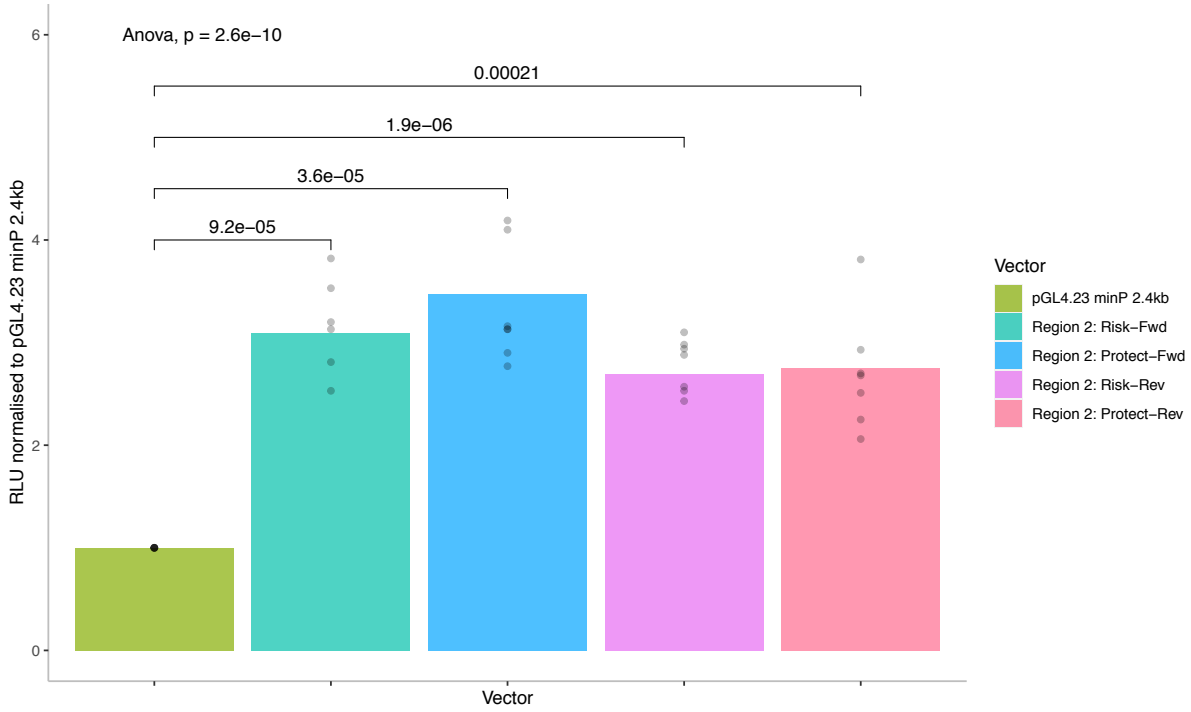


Figure 6.3: Relative luciferase activity of Region 2 inserts compared to size-matched control. One-way ANOVA was performed to compare the mean relative luciferase units (RLU) for each test construct normalised to pGL4.23 [luc2/minP] 2.4kb control vector. Paired t-tests were performed to compare each Region 2 insert to the pGL4.23 [luc2/minP] 2.4kb control vector (p-values shown for each pairwise comparison).

6.2.6 The genotype at *rs4375701* does not significantly affect enhancer activity

The genotype of *rs4375701*, that lies within region 2, may alter the level of enhancer activity. To test this, Region 2 constructs were generated from AS patient DNA templates containing either the risk allele or the protective allele. If this SNP is exerting a substantial effect on the activity of the putative enhancer it should be observed in the experiment as a significant difference in relative luciferase units between the inserts containing the different alleles.

No difference in relative luciferase activity of Region 2: Risk-Fwd compared with Region 2: Protect-Fwd was observed (fold change 0.89, paired t-test p-value = 0.83). Similarly, no difference was observed in the reverse orientation (Region 2: Risk-Rev to Region 2: Protect-Rev, fold change 0.98, p-value = 0.71). No significant difference in enhancer activity was observed between constructs containing the risk and protective allele at *rs4375701* regardless of insert orientation.

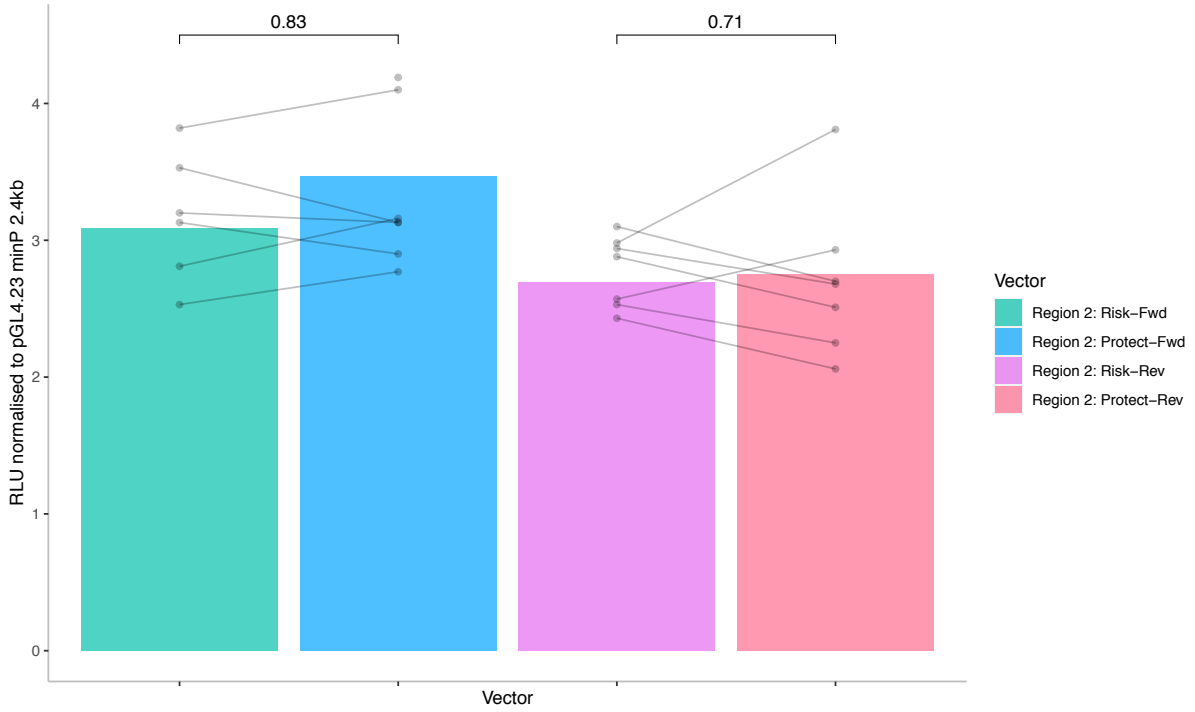


Figure 6.4: Comparison of luciferase activity of genotype specific Region 2 inserts. Relative luciferase units were normalised to pGL4.23 [luc2/minP] 2.4kb vector. Paired t-tests were performed to compare constructs containing *rs4375701* risk genotype (Risk-Fwd and Risk-Rev) with the protective genotype (Protect-Fwd and Protect-Rev).

6.2.7 The orientation of the insert significantly affects enhancer activity

Enhancer activity is generally reported to function independently of sequence orientation (Pennacchio et al., 2013, Grosschedl, 2001) and the direction of the enhancer sequence should not significantly affect the ability of sequence to exert an enhancer type role. To test this, pGL4 constructs were generated with the Region 2 sequence in both potential orientations, as well as for each allele of *rs4375701* (Table 6.1). This was achieved by the positioning of *Bam*HI and *Sal*I cut-sites in the amplification primers allowing for orientational control of the insert when ligating into the pGL4.23 [luc2/minP] vector.

Relative luciferase activity from constructs with the same genotype but opposing orientations were compared using a paired t-test. Region 2: Risk-Fwd was compared to Risk-Rev and showed a fold change of 0.87 (p = 0.084). Region 2: Protect-Fwd was compared to Protect-Rev, a fold change of 0.79 (p = 0.0014). This result showed that the orientation of the insert did in fact have an effect on enhancer activity.

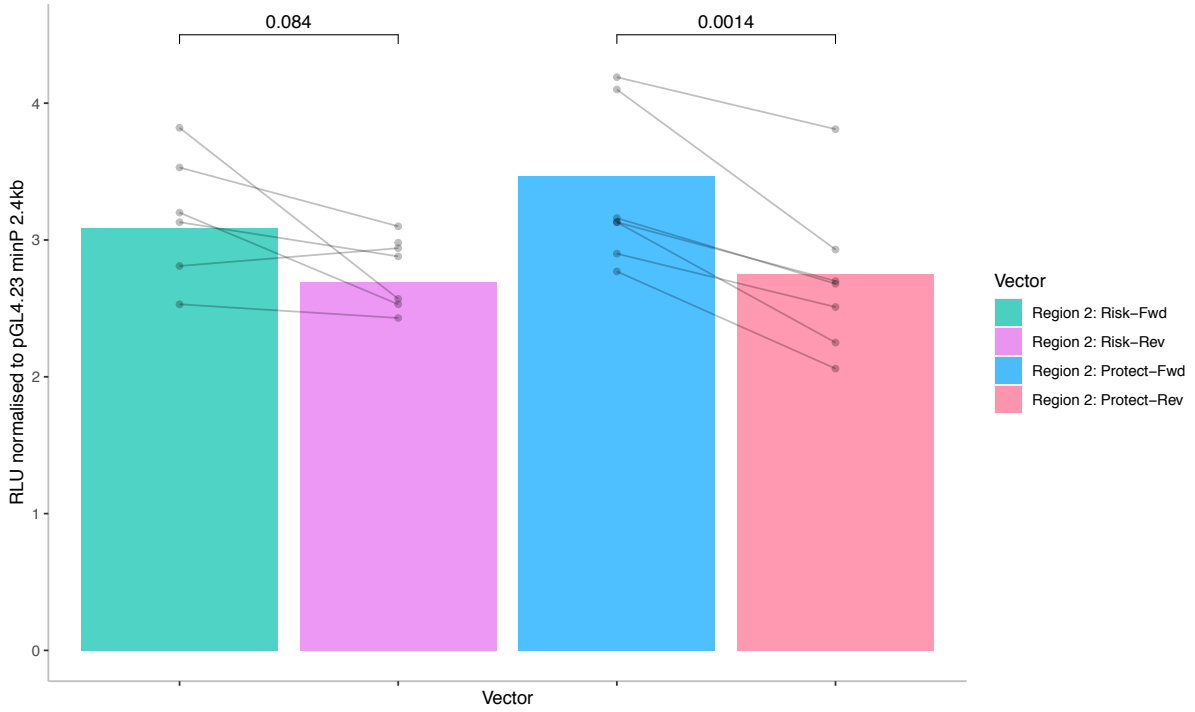


Figure 6.5: Comparison of luciferase activity of orientation-specific Region 2 inserts. Relative luciferase units were normalised to pGL4.23 [luc2/minP] 2.4kb vector. Paired t-tests were performed to compare different oriented Region 2 inserts of the same genotype.

6.3 Discussion and conclusion

Region 2 was demonstrated to contain a regulatory element that has an enhancer-type effect on gene expression in a luciferase assay. The presence of the inserted sequence increased expression more than 2.7 times above that observed in a size-matched control vector. This finding supports the presence of a putative enhancer region in the *NPEPPS-KPNB1* intergenic region that was identified in Chapter 4.

This putative regulatory region is tagged by the SNP *rs4375701* that is associated at genome-wide significance with AS ($p = 6.44x10^{-9}$). When the risk and protective alleles, at *rs4375701*, were compared to assess their effect on gene expression, no significant difference was observed. In both insert orientations, gene expression was slightly reduced in the insert containing the risk genotype. Although not significant, this is consistent with the result seen in the RT-qPCR gene expression results for *NPEPPS* and *TBKBP1* for AS cases, where lower expression was observed compared to healthy controls. Further work is required to investigate if this pattern might become significant under different experimental conditions (discussed below).

A limitation of this study was that the construct was transfected into the cell-line HEK-293T and not a cell-type implicated in AS or specific to the prioritised region. HEK-293T cells are derived from human embryonic kidney cells. This cell-line was used due to its relative ease to transfect and its previous use for reporter studies in ankylosing spondylitis (Vecellio et al., 2016, Roberts et al., 2015). This approach allowed for the identification of a regulatory element with the aim to follow up with

experiments in a more relevant cell-type. As discussed previously, gene expression and its regulation are context and cell-type specific and although the effects of this regulatory region were observed in HEK-293T cells the genotype-specific effect may require the conditions provided in a cell-type specific environment preventing the detection of any differential expression. Work was begun to replicate this study in JURKAT cells, a cell-line derived from human T lymphocyte cells. This is a more relevant cell-type for the investigation of Region 2 as this region was identified by a CD8-specific ATAC peak. Transient transfection was attempted using lipofectamine but insufficient transfection efficiency was obtained. The next step would have been to test electroporation methods of transfection. Unfortunately, due to the recent laboratory closure because of COVID-19, work could not continue on this line of investigation.

A further limitation of this approach is the size of Region 2 insert fragment. The size of the construct is an important factor to consider when measuring enhancer activity. As demonstrated in the effect of size on the gene expression of the control vector, increase in vector size results in decreased activity observed. This holds true for constructs containing enhancers with larger constructs displaying reduced activity (Yin et al., 2005). The large fragment size of the Region 2 construct is likely to reduce the signal observed and inhibit the ability to distinguish between small but significant changes. The insert for Region 2 was 3.0 kb in length and was designed to capture the observed CD8-specific ATAC peak from in-house data as well as the neighbouring enhancer region in cell-type specific chromHMM data. By starting with a larger fragment, the full potential for the regulatory element to influence

transcriptional activity was captured in the insert but this may have hindered the comparison between the risk and protective allele inserts. Further experiments were designed to generate smaller inserts as subsections of the original Region 2 insert would enable further characterisation of the exact sequence of the regulatory element and the contribution of the *rs4375701* genotype.

7 General discussion

This study has achieved its primary goal to fine-map the ankylosing spondylitis disease-associated *NPEPPS* region on chromosome 17. Through statistical approaches it has improved on previously reported fine-mapping of this region. Using publicly available datasets and in-house AS-specific data, the disease-associated region was successful annotated and the identification of regulatory regions with a putative role in disease was achieved. Functional experiments demonstrated that *NPEPPS* and *TBKBP1*, two genes in the prioritised region, had lower expression in AS cases compared to healthy controls. It was also shown that a putative regulatory region identified during the fine-mapping process has the ability to acts as an enhancer.

Although this study has improved on our previous knowledge of this disease-associated region there were limitations to these findings. This chapter will discuss the key findings in the context of AS research, its limitations, and suggest future work that would build on these results.

7.1 Successes and limitations of the fine-mapping approach

This study improved on previously reported fine-mapping of the *NPEPPS* region because it used CoreExome genotyping data which covers 5080 SNPs in the 1Mb genomic. As expected, these statistical approaches were unable to pinpoint

a specific causal gene or variant in disease, but rather identified a 99% credible set of 62 SNPs. This was an improvement on the previously reported 99% credible set at the *NPEPPS* region which contained 141 SNPs (van de Bunt et al., 2015) and reduced the overlapping fine-mapping genomic window by over 62 kb. The inability of this approach to pinpoint a specific gene or variant in disease is due to the strong LD structure observed in the region. Linkage disequilibrium is both a strength and limitation of GWAS studies. It enables their performance through the use of tag SNPs to predict the genotype of larger networks of SNPs (Spain and Barrett, 2015) but when LD is strong, and its structure extends over a large genomic region, it can prevent the fine level identification of specific genes or variants. When this limits the resolution of the fine-mapping results, like it has for the *NPEPPS* region, alternative approaches are required.

7.2 Cell-type and disease specific annotation is key to successful prioritisation

The annotation of the *NPEPPS* region was a critical step to transition from the statistically identified putative causal variants to experimental validation through the prioritisation of causal variants. Annotation of the region enabled the visualisation of the genomic landscape in combination with the location of putative causal SNPs identified by statistical methods. Thus, it was possible to further prioritise the variants and identify those with a plausible role in disease. In this study, the value of publicly available datasets in the annotation and prioritisation of variants was limited. With the majority of variants in non-coding

regions of the genome, the tools used to predict causality struggled to agree on annotation of function and subsequent ranking of variants. This poor correlation between results hindered interpretation and the use of these findings in prioritisation.

Greater weight was placed on in-house data when prioritising SNPs due to this datasets cell-type and disease specificity. These are highly important factors when using data for annotation and prioritisation as they are a better reflection of the situation in disease. As a result, prioritisation of SNPs and their selection for functional follow-up experiments is more likely to capture variants with a role in disease when compared to annotation using data that utilises highly heterogenous cell-types and disease contexts.

7.3 Functional experiments and the remaining challenge of identifying disease mechanism

Experimental approaches fell short of their ultimate goal to validate the functional role of the prioritised variants in disease. However, lower expression was identified in *NPEPPS* and *TBKBP1* in AS cases compared to controls. This could not be linked to the regulatory control of a specific region or influence of a specific set of variants through eQTL analysis, possible due to the study being underpowered. Despite this, homozygosity of the risk allele was observed to have some effect on gene expression when the cohort was limited by disease status. Through luciferase reporter assays, a putative regulatory region in the

intergenic region between *NPEPPS* and *KPNB1* was shown to display enhancer activity. This enhancer activity was not shown to be influenced by the specific *rs4375701* allele present in the DNA sequence under the particular experimental conditions used.

It was the aim of functional experiments to to link the lower expression of *NPEPPS* and *TBKBP1* observed in AS cases to the influence of a specific regulatory region and the activity of a disease-associated SNP as has been demonstrated for *rs4648889* in *RUNX3* (Vecellio et al., 2016). This was not possible from the current experiments and as the identified enhancer in the *NPEPPS-KPNB1* intergenic region did not demonstrate differential activity through the influence of the *rs4375701* allele it cannot be demonstrated to be linked to the observed RT-qPCR results. This can be interpreted as this enhancer region is not responsible for the differential expression observed and other regulatory elements may be responsible. However, as previously discussed in Chapter 6, the disease-associated SNP in this regulatory region may require the specific context of a specific cell-type or AS-disease conditions to influence differential regulatory active which are not provided in the current experimental conditions. For this reason, it is not possible to exclude this enhancer region using the current experimental approaches. Until further work is performed it is not possible to explain the reduced expression observed in AS cases in RT-qPCR data or which regulatory mechanisms or specific elements are driving this difference and whether this has an effect on disease progression.

Both methods offered a glimpse of the potential role that regulatory elements may play in disease through regulating gene expression but could not extend to the identification of specific variants controlling this mechanism. Further work is required to identify how gene regulation is specifically controlled in AS cases compare with healthy controls and what mechanism is responsible for the differential expression observed in *NPEPPS* and *TBKBP1*. Through addressing this limitation of the study, it may be possible to explain the *NPEPPS* regions association with disease and identify the mechanism this region contributes to ankylosing spondylitis.

7.4 Future work

7.4.1 Fine-mapping of all AS associated regions using CoreExome data

This study has demonstrated the improvement in fine-mapping results for the *NPEPPS* region when using CoreExome genotyping data. Through a Bayesian analysis approach GWAS loci can be fine-mapped to identify credible sets of SNPs in clearly defined genomic windows. When credible sets contain many SNPs in strong LD blocks, like the *NPEPPS* region, further steps may prove useful to reduce the number of putative causal SNPs.

Trans-ethnic studies offer one approach that may yield positive outcomes in regions where strong LD is present (Spain and Barrett, 2015). By using the naturally occurring difference in LD between distinct populations, the LD structure

can be broken down into smaller sections to further reduce the number of associated SNPs or genes in a region (Li and Keating, 2014). The *NPEPPS* region has been associated with AS in the East Asian population by the IGAS study (International Genetics of Ankylosing Spondylitis Consortium et al., 2013) and may offer a suitable population for a trans-ethnic study. This approach would be useful to improve the fine-mapping at the *NPEPPS* region where the LD block in the European population extends across three genes, *NPEPPS*, *KPNB1*, and *TBKBP1*, in a 211,521 bp genomic region.

An alternative approach would be to deep-sequence the prioritised region identified from GWAS analysis to investigate the fine-scale genomic structure of the region (Salomon et al., 2016). This approach would address an essential component of fine-mapping by capturing all common SNPs in the analysis and provide information regarding variants not previously genotyped. This approach would also allow the investigation of rare variants (MAF <1%) not considered in GWAS methodologies and address the contribution of rare variants in the disease-associated region.

The analytical process used in this study has the potential to be automated in a computational pipeline and applied to all genome-wide AS-associated regions of the CoreExome dataset. In the Knight group, similar work has been performed on the UK cohort of the IGAS study (International Genetics of Ankylosing Spondylitis Consortium et al., 2013) but the increased coverage of the CoreExome dataset may contribute improved and potential new findings as has been shown for the *NPEPPS* region. This has the potential to improve on the current knowledge of

the known AS associated regions and implicate new putative regions and variants for further investigation.

7.4.2 Cell-type and disease specific data are essential for annotation

The cell-type and disease specific epigenomic data used to annotate the *NPEPPS* region enabled the prioritisation of putative causal variants overlapping regulatory elements. This AS-specific dataset outperformed the annotation provided by publicly available dataset, highlighting a current limitation when prioritising putative causal variants in a fine-mapping strategy.

The epigenomic dataset was of great value and should be used to further extend the fine-mapping of the AS disease-associated loci and prioritise variants for functional experimental follow-up. With data in the master peak coordinate format it would be possible to bring this aspect of annotation into an automated pipeline to identify overlapping regions. This could aid in reducing the time and effort required to manually assess variants by directing the researcher to specific sets of SNPs. The Knight group is currently developing a chromHMM version of its AS-specific epigenomic data using CD4, CD8, and CD14 cells and this resource will provide a leap forward in disease-specific annotation in future fine-mapping projects.

7.4.3 Replication of functional experiments in disease-specific cell-types

The relevance and interpretation of data is dependent on the cell-type in which the data has been generated. Gene expression assays were performed using PBMCs providing a broad insight into the difference in expression between AS cases and healthy controls but the mechanism behind this differential expression could not be identified. Cell-type specificity must be a consideration for functional experiments when validating a causal variants mechanism in disease as understanding the cell-type and context in which a disease risk variant produces its function can offer novel insight into pathogenic mechanism (Al-Mossawi et al., 2019). This is of greater importance when variants have been prioritised using cell-type specific data, as for the selection of putative regulatory regions in this study. By replicating this experiment using RNA extracted from disease-specific cell-types it may be possible to determine whether the differential expression observed is driven by gene regulation in a specific cell subset.

A similar limitation must also be considered for the luciferase reporter assays to investigate function. In the current study, the cell-line HEK-293T was used as a transfection protocol had successfully been established in the Knight group and there was precedent for the use of this cell-line in AS studies (Vecellio et al., 2016, Roberts et al., 2015). By replicating these findings in a cell-type implicated in disease the results generated will more closely reflect what occurs in the disease context. This is not without its challenges, as more disease-appropriate cell-lines are not as simple to transfect and alternative methods may be required

to replicate any experiments. In addition, assays performed in cell-lines are not a true reflection of a disease-specific environment and may miss the functional mechanism or provide misleading results, complicating interpretation. Performing luciferase assays in AS patient-derived samples would provide the most realistic context for additional studies. The benefit of achieving the replication of these luciferase experiments in an appropriate model system is that any regulatory effect will be more relevant in the context of disease, simplifying interpretation. This will ultimately aid the investigation of the mechanism in which disease-associated variants play a role in ankylosing spondylitis.

7.4.4 Alternative and additional functional experiments to investigate putative regulatory variants

Further functional approaches could be utilised in an attempt to identify and validate the mechanism linking the *NPEPPS* region to disease. Current approaches have implicated causal variants with a role in gene regulation reducing the expression of *NPEPPS* and *TBKBP1* in AS cases. Through alternative experiments it may be possible to identify the regulatory region or causal variant driving this mechanistic effect.

RNA-seq has the potential to improve on the findings of gene expression analysis performed by RT-qPCR in this study. The *NPEPPS* gene was specifically challenging to investigate by RT-qPCR due to its large number of predicted transcripts which required multiple primer sets to capture. An RNA-seq approach would specifically address this issue by enabling the capture of all transcripts with

the additional advantage of being genome-wide. This approach allows for the detection of more genomic features including; the capture of specific transcript abundance, unknown genes, splice variants, and non-coding RNAs all of which could offer further insight into disease mechanism.

Another approach would be to investigate the effect of variants on transcription factor interaction with the DNA sequence in the putative regulatory elements implicated in this region. Through electrophoretic mobility gel shift assays (EMSA) it is possible to identify which TFs are capable of binding to the DNA sequence at these putative regulatory elements and whether the affinity of this binding is altered by the presence of different alleles of a putative causal variant. Vecellio and colleagues identified differential allelic TF binding at the DNA sequence around *rs4265380* in the AS-associated *RUNX3* region using EMSAs (Vecellio et al., 2018). This approach could be an additional experiment to further investigate the putative regulatory elements in the *NPEPPS* regions.

Genome editing is another approach that offers an opportunity to address some of the limitation in the current strategy used to investigate putative variants and regulatory elements. Through the use of clustered regularly interspaced short palindromic repeats (CRISPR) in combination with the Cas9 nuclease and specific guide RNAs, it is possible to create precise targeted double stranded breaks at specific DNA sites (Mali et al., 2013, Cong et al., 2013). This technology can be used to investigate putative regulatory elements by either by the introduction of base pair edits to replicate specific SNP polymorphisms or knockout complete regulatory elements within appropriate cell types. Thus,

CRISPR/Cas9 can enable the investigation of the regulatory impact of specific putative causal variants within their true genomic and cellular environment addressing the limitations of cell-type and disease specificity, the two major limitations of the approaches used in this study.

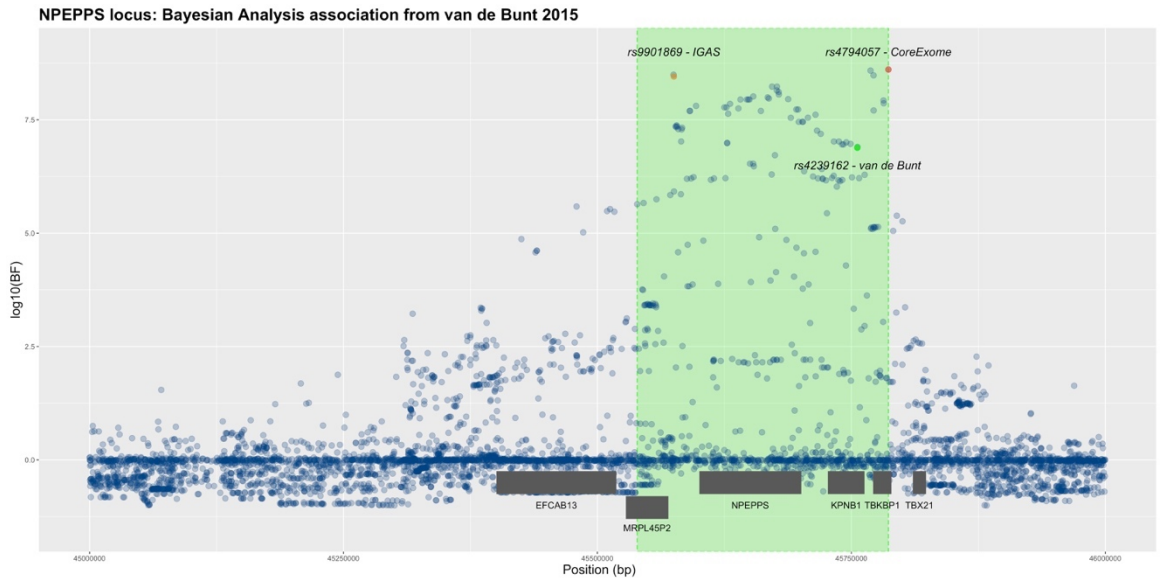
7.5 Conclusion

The overarching aim of this study was to fine-map the disease-associated GWAS signal at the *NPEPPS* region on chromosome 17. Through computational fine-mapping approaches followed by region specific annotation putative causal variants were identified and prioritised. The results of this process were then investigated using functional experiments in an attempt to validate their role in ankylosing spondylitis. This study design demonstrates an example workflow for the initial fine-mapping investigation of a disease-associated region identified through GWAS. This computational fine-mapping strategy could be readily employed as an initial analytical step of post-GWAS analysis to identify and prioritise causal variants to undergo functional experiments to validate their mechanism in disease. Specifically, in the field of AS research this approach should be utilised to investigate the known disease-associated loci. In doing so, this work has the potential to lead to the discovery of new disease pathways as a first step to designing new therapies for AS. With current treatments only effective in subsets of patients and having to be taken life-long there is still a real need to identify and develop new treatments for this debilitating condition.

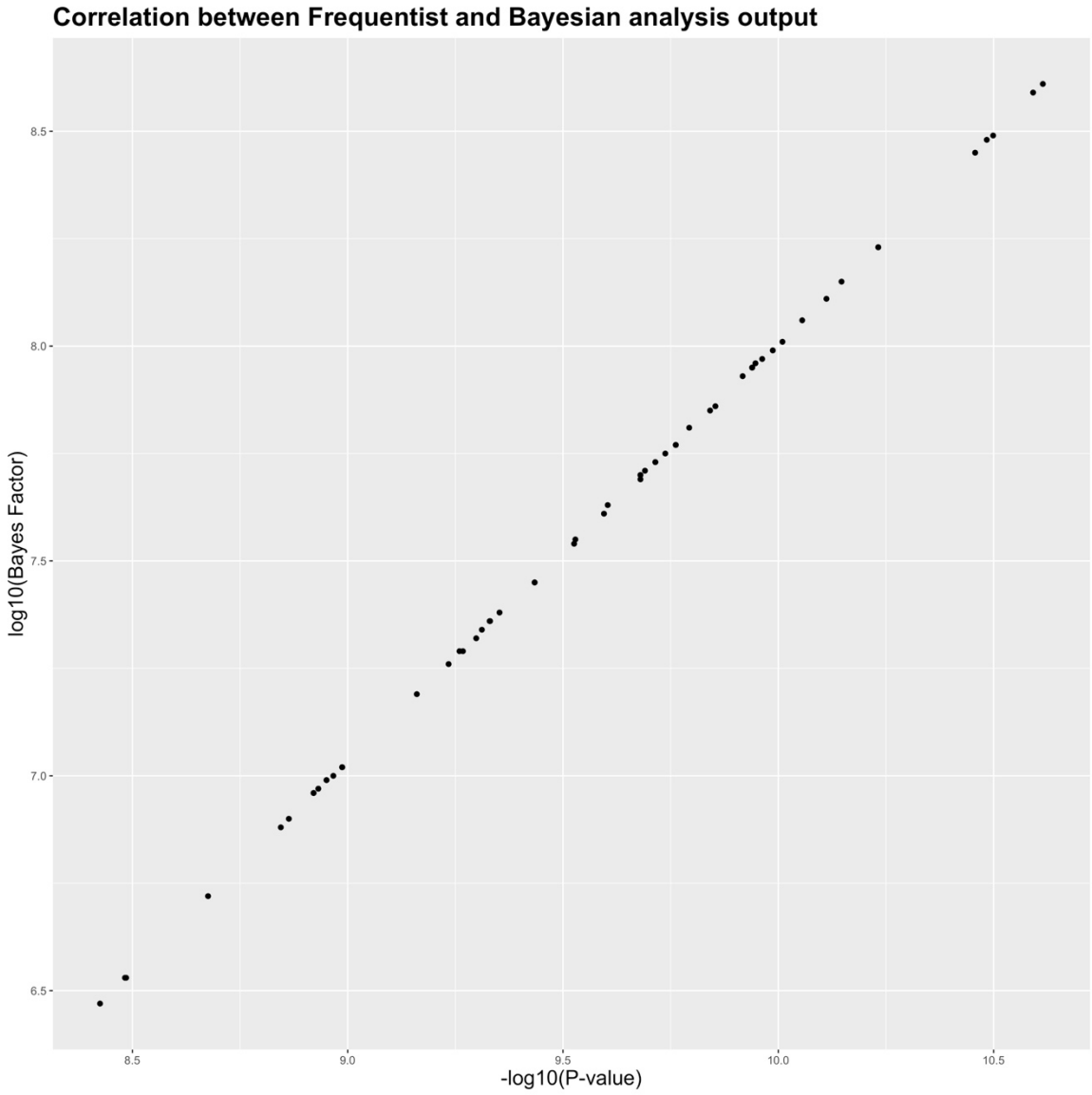
8 Appendices

8.1 Supplementary figures

8.1.1 Chapter 3 supplementary figures

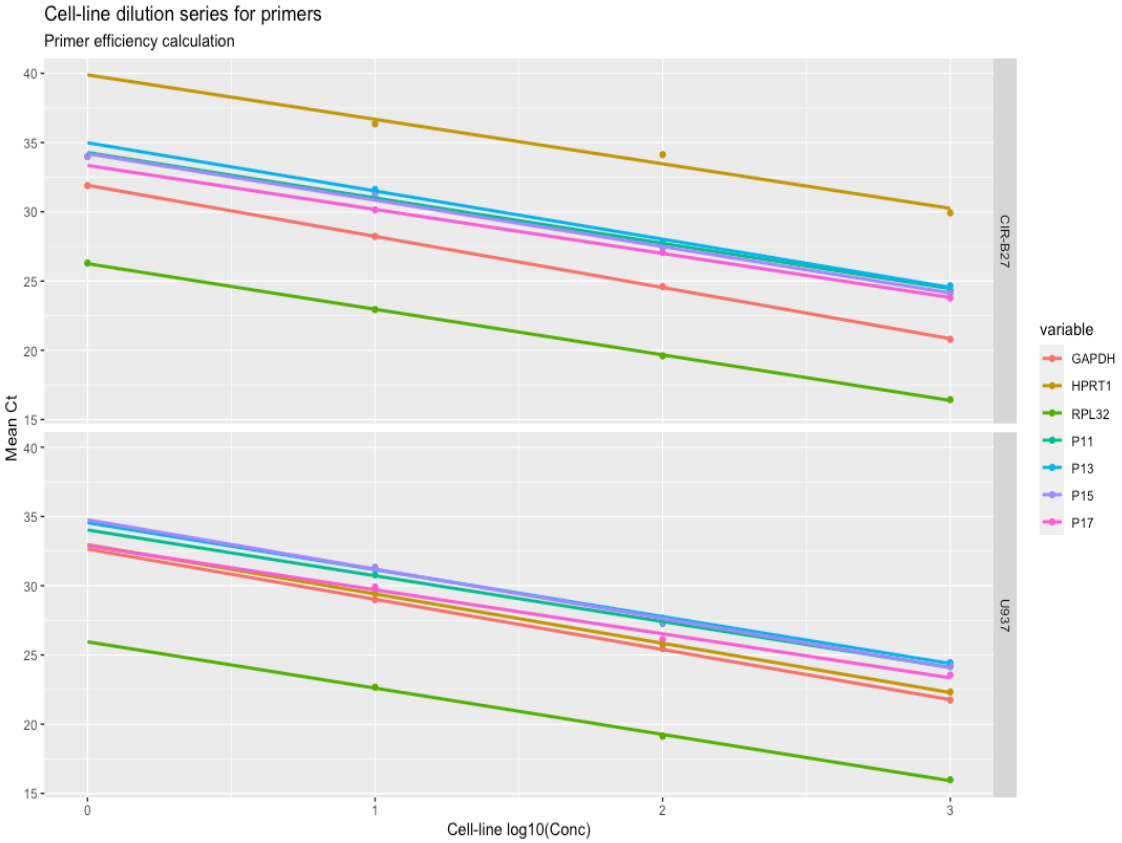


Supplementary Figure 3.1: CoreExome Bayesian analysis association plot for the *NPEPPS* region highlighting the prioritised region from van de Bunt 2015 (van de Bunt et al., 2015). The most significant SNP identified, *rs4239162* (log10(BF) = 9.95, PP = 0.04, OR = 1.15), has been highlighted in green with the 247,336 bp window capturing the 99% and 95% credible sets highlighted in green. Also highlighted is the lead SNP from the original IGAS 2013 dataset (orange) and from the current CoreExome study (red).

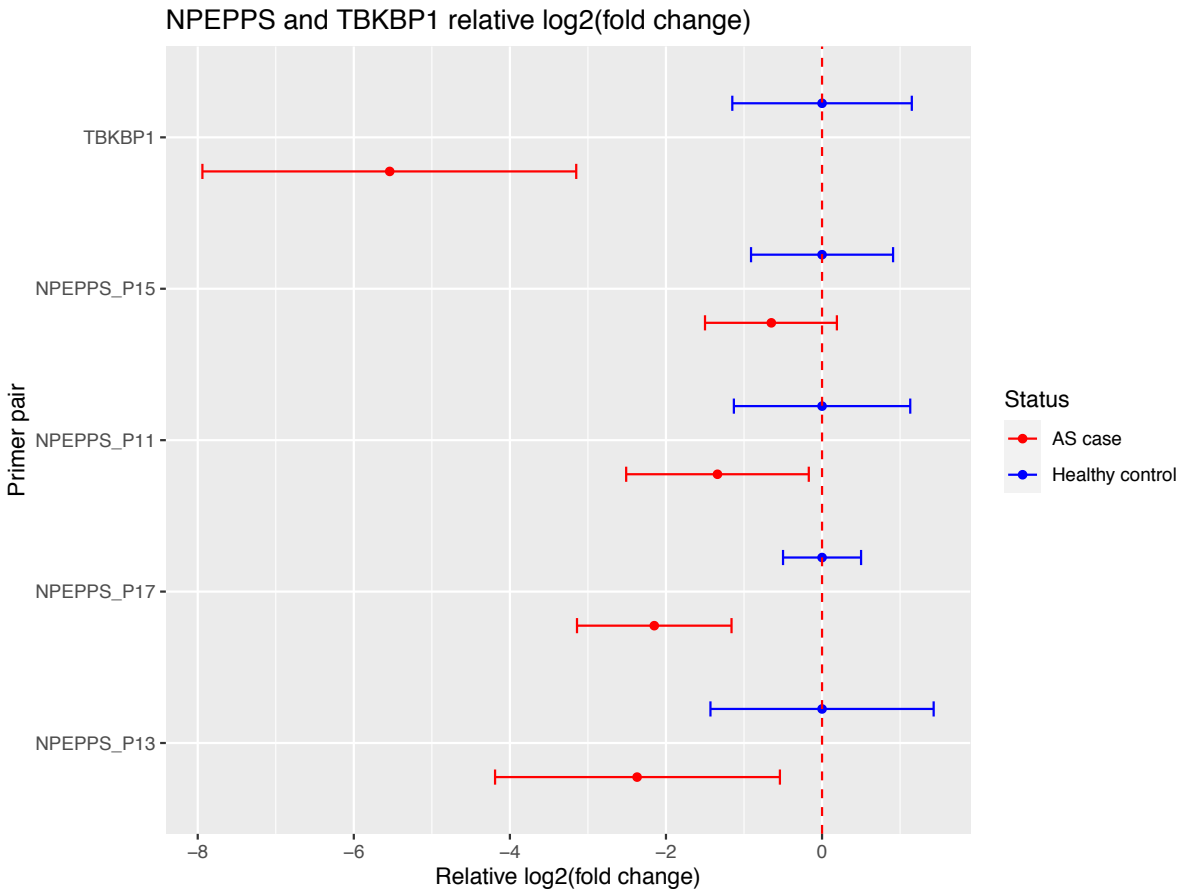


Supplementary Figure 3.2: Correlation between Frequentist -log10(P-value) and log10(Bayes Factor) for 62 SNPs in 99% credible set.

8.1.2 Chapter 5 supplementary figures



Supplementary Figure 5.1: RT-qPCR mean Ct values for each primer set across two cell-line dilution series. Dilution series for C1R-B27 cell-line (top) and U937 cell-line (bottom). Data used to calculate the primer efficiencies. As show, gradients of trend lines are highly similar indicating that the reaction efficiency is approximately equal for each primer.



Supplementary Figure 5.2: Log2(fold change) for *NPEPPS* and *TBKBP1* from healthy controls to AS cases showing a decrease in expression in AS cases across all primers.

8.2 Supplementary tables

8.2.1 Chapter 2 supplementary tables

Primer set	202	204	213	215	207	218	201	220	208	221	225	210	224	205	223	212	222	217	206	211	216	209	219	203	214
NPEPPS_P11		+		+			+		+		+		+							+	+	+	+	+	
NPEPPS_P13	+	+	+	+	?	+	+	+	+	+															
NPEPPS_P15				+			+									+		+			+				
NPEPPS_P17		+		+			+		+	+	+	+					+								

Protein coding	
Nonsense-mediated decay	
Processed transcript	
retained intron	

Supplementary Table 2.1: *NPEPPS* primer capture table for *NPEPPS* transcripts. Table lists all 25 *NPEPPS* transcripts and identifies their known and predicted transcript status using colour. A (+) identifies if a primer is predicted to capture this transcript from its primer locations.

8.2.2 Chapter 3 supplementary tables

Chromosome	SNP name	Position (bp)	Allele A	Allele B	Odds Ratio (OR)	L95	U95	p-value
17	rs4794057	45786452	C	T	1.13	1.08	1.17	2.43E-11
17	exm-rs8070463	45768836	T	C	1.13	1.08	1.17	2.56E-11
17	rs11651766	45574931	G	T	1.13	1.09	1.17	3.17E-11
17	rs67919208	45771873	T	C	1.13	1.08	1.17	3.28E-11
17	rs9901869	45575206	G	A	1.13	1.09	1.17	3.49E-11
17	rs12937560	45671869	T	C	1.13	1.08	1.17	5.86E-11
17	rs8075411	45676760	T	C	1.13	1.08	1.17	5.86E-11
17	rs8078880	45676526	C	G	1.13	1.08	1.17	7.13E-11
17	rs4793908	45678408	A	G	1.13	1.08	1.17	7.73E-11
17	rs4793913	45678579	T	C	1.12	1.08	1.17	8.80E-11
17	rs8072644	45653364	T	A	1.12	1.08	1.17	9.78E-11
17	rs4092460	45667550	A	G	1.12	1.08	1.17	1.03E-10
17	rs62074015	45668760	T	A	1.12	1.08	1.17	1.09E-10
17	rs10163469	45687767	A	G	1.12	1.08	1.17	1.13E-10
17	rs11079779	45640770	T	C	1.12	1.08	1.17	1.15E-10
17	rs8082284	45647067	C	T	1.12	1.08	1.17	1.15E-10
17	rs9635762	45648446	C	A	1.12	1.08	1.17	1.15E-10
17	rs9635763	45649694	A	C	1.12	1.08	1.17	1.15E-10
17	rs4439799	45781599	C	T	1.12	1.08	1.16	1.21E-10
17	rs4561509	45781799	G	A	1.12	1.08	1.16	1.40E-10
17	rs55687600	45630497	T	C	1.12	1.08	1.17	1.44E-10
17	rs4794058	45597098	T	C	1.12	1.08	1.17	1.61E-10
17	rs10432035	45624913	C	G	1.12	1.08	1.17	1.73E-10
17	rs79443387	45626961	T	C	1.12	1.08	1.17	1.73E-10
17	rs8070759	45638760	T	C	1.12	1.08	1.17	1.83E-10
17	rs3760370	45695832	T	C	1.12	1.08	1.17	1.93E-10
17	rs7207542	45697549	G	C	1.12	1.08	1.17	1.93E-10
17	rs62074055	45771933	G	C	1.12	1.08	1.16	2.04E-10
17	rs2125873	45660887	G	T	1.12	1.08	1.17	2.09E-10
17	rs4794007	45590722	A	G	1.12	1.08	1.17	2.09E-10
17	rs8065990	45591166	A	G	1.12	1.08	1.17	2.09E-10
17	rs4794392	45628693	T	A	1.12	1.08	1.17	2.49E-10
17	rs11656855	45714655	T	C	1.12	1.08	1.16	2.54E-10
17	rs4289035	45707211	C	T	1.12	1.08	1.16	2.96E-10
17	rs12943464	45690351	T	A	1.12	1.08	1.16	2.98E-10
17	rs4793978	45698175	A	G	1.12	1.08	1.16	3.68E-10
17	rs7223029	45701731	C	T	1.12	1.08	1.16	3.68E-10
17	rs11079784	45702280	C	T	1.12	1.08	1.16	3.68E-10
17	rs4794051	45577914	A	G	1.12	1.08	1.16	4.44E-10
17	rs4060785	45576936	C	G	1.12	1.08	1.16	4.67E-10
17	rs11869940	45578435	A	T	1.12	1.08	1.16	4.68E-10
17	rs12103836	45577469	G	A	1.12	1.08	1.16	4.88E-10
17	rs8074149	45583375	C	A	1.12	1.08	1.16	5.03E-10
17	rs62076549	45579555	G	T	1.12	1.08	1.16	5.40E-10
17	rs7221548	45582952	C	T	1.12	1.08	1.16	5.50E-10
17	rs10853090	45716124	T	C	1.12	1.08	1.16	5.83E-10
17	rs55633793	45719816	G	A	1.12	1.08	1.16	6.91E-10
17	rs12051716	45582395	A	G	1.12	1.07	1.16	1.03E-09
17	exm2264630	45732774	A	C	1.12	1.07	1.16	1.03E-09
17	rs7219303	45737701	G	A	1.12	1.07	1.16	1.03E-09
17	rs73315878	45744993	A	G	1.12	1.07	1.16	1.08E-09
17	rs4794372	45627760	A	G	1.12	1.07	1.16	1.12E-09
17	rs4794371	45627693	C	G	1.12	1.07	1.16	1.12E-09
17	rs4399567	45749172	G	C	1.12	1.07	1.16	1.17E-09
17	rs9905922	45740805	A	G	1.12	1.07	1.16	1.20E-09
17	rs56226803	45742233	T	A	1.12	1.07	1.16	1.20E-09
17	rs4239163	45755954	C	T	1.11	1.07	1.16	1.37E-09
17	rs4239162	45755810	G	A	1.11	1.07	1.16	1.43E-09
17	rs8072100	45674687	A	T	0.90	0.86	0.93	2.11E-09
17	rs8077106	45653291	T	G	0.90	0.87	0.94	3.27E-09
17	rs12603290	45650196	T	C	0.90	0.87	0.94	3.29E-09
17	rs8081717	45653921	T	A	0.90	0.87	0.94	3.76E-09
17	rs4794001	45720849	T	C	1.11	1.07	1.15	4.38E-09
17	rs7220935	45703433	C	T	0.90	0.87	0.94	4.87E-09
17	rs4793842	45671506	A	G	1.11	1.07	1.15	5.83E-09
17	rs4794048	45763073	C	A	0.90	0.87	0.94	5.87E-09
17	rs11870935	45732605	G	A	0.90	0.87	0.94	6.17E-09
17	rs4375701	45712218	T	C	0.90	0.87	0.94	6.44E-09
17	rs8067286	45594590	A	G	0.90	0.87	0.94	6.66E-09
17	rs3809868	45750596	G	A	0.90	0.87	0.94	6.73E-09
17	rs2317826	45730899	T	C	0.90	0.87	0.94	6.97E-09
17	rs78551919	45626002	A	C	0.90	0.87	0.94	7.04E-09
17	rs7210738	45757598	G	A	0.90	0.87	0.94	7.04E-09
17	rs12150231	45717091	G	A	0.90	0.87	0.94	7.10E-09
17	rs11869730	45614763	T	A	0.90	0.87	0.94	7.12E-09
17	rs8065437	45721653	T	C	0.90	0.87	0.94	7.14E-09
17	rs7222851	45592169	T	C	0.90	0.87	0.94	7.15E-09
17	rs7350906	45588112	G	A	0.90	0.87	0.94	7.26E-09
17	rs11650372	45721795	A	G	0.90	0.87	0.94	7.36E-09
17	rs8079821	45611704	T	C	0.90	0.87	0.94	7.66E-09
17	rs9905583	45741016	T	C	0.90	0.87	0.94	7.70E-09
17	rs4264433	45737275	A	T	0.90	0.87	0.94	7.80E-09
17	rs8065669	45724996	A	G	0.90	0.87	0.94	7.87E-09
17	rs4580230	45738861	A	G	0.90	0.87	0.94	8.26E-09
17	rs8078686	45735706	C	T	1.11	1.07	1.15	1.10E-08
17	rs9903898	45575354	C	T	1.11	1.06	1.15	1.41E-08
17	rs12602134	45582136	G	C	1.11	1.06	1.15	1.63E-08
17	rs7405889	45571676	T	C	0.90	0.87	0.94	1.71E-08
17	rs62076542	45558090	T	C	0.91	0.87	0.94	2.14E-08
17	rs2085755	45545378	G	T	0.91	0.87	0.94	2.60E-08
17	rs9895274	45539117	T	C	0.91	0.87	0.94	2.77E-08
17	rs12452315	45479446	C	A	0.91	0.87	0.94	3.12E-08
17	rs4338849	45512483	G	A	0.91	0.87	0.94	3.58E-08
17	rs77454872	45509496	A	G	0.91	0.87	0.94	3.97E-08
17	rs11079775	45516667	T	C	0.91	0.87	0.94	4.10E-08
17	rs4794016	45725931	T	A	0.91	0.87	0.94	4.46E-08

Supplementary Table 3.1: Frequentist Analysis genome-wide significant SNPs (n = 96).

Chromosome	SNP name	Position (bp)	Allele A	Allele B	Odds ratio (OR)	L95	U95	p-value
17	rs4794057	45786452	C	T	1.13	1.08	1.17	2.43E-11
17	exm-rs8070463	45768836	T	C	1.13	1.08	1.17	2.56E-11
17	rs11651766	45574931	G	T	1.13	1.09	1.17	3.17E-11
17	rs67919208	45771873	T	C	1.13	1.08	1.17	3.28E-11
17	rs9901869	45575206	G	A	1.13	1.09	1.17	3.49E-11
17	rs12937560	45671869	T	C	1.13	1.08	1.17	5.86E-11
17	rs8075411	45676760	T	C	1.13	1.08	1.17	5.86E-11
17	rs8078880	45676526	C	G	1.13	1.08	1.17	7.13E-11
17	rs4793908	45678408	A	G	1.13	1.08	1.17	7.73E-11
17	rs4793913	45678579	T	C	1.12	1.08	1.17	8.80E-11
17	rs8072644	45653364	T	A	1.12	1.08	1.17	9.78E-11
17	rs4092460	45667550	A	G	1.12	1.08	1.17	1.03E-10
17	rs62074015	45668760	T	A	1.12	1.08	1.17	1.09E-10
17	rs10163469	45687767	A	G	1.12	1.08	1.17	1.13E-10
17	rs11079779	45640770	T	C	1.12	1.08	1.17	1.15E-10
17	rs8082284	45647067	C	T	1.12	1.08	1.17	1.15E-10
17	rs9635762	45648446	C	A	1.12	1.08	1.17	1.15E-10
17	rs9635763	45649694	A	C	1.12	1.08	1.17	1.15E-10
17	rs4439799	45781599	C	T	1.12	1.08	1.16	1.21E-10
17	rs4561509	45781799	G	A	1.12	1.08	1.16	1.40E-10
17	rs55687600	45630497	T	C	1.12	1.08	1.17	1.44E-10
17	rs4794058	45597098	T	C	1.12	1.08	1.17	1.61E-10
17	rs10432035	45624913	C	G	1.12	1.08	1.17	1.73E-10
17	rs79443387	45626961	T	C	1.12	1.08	1.17	1.73E-10
17	rs8070759	45638760	T	C	1.12	1.08	1.17	1.83E-10
17	rs3760370	45695832	T	C	1.12	1.08	1.17	1.93E-10
17	rs7207542	45697549	G	C	1.12	1.08	1.17	1.93E-10
17	rs62074055	45771933	G	C	1.12	1.08	1.16	2.04E-10
17	rs2125873	45660887	G	T	1.12	1.08	1.17	2.09E-10
17	rs4794007	45590722	A	G	1.12	1.08	1.17	2.09E-10
17	rs8065990	45591166	A	G	1.12	1.08	1.17	2.09E-10
17	rs4794392	45628693	T	A	1.12	1.08	1.17	2.49E-10
17	rs11656855	45714655	T	C	1.12	1.08	1.16	2.54E-10
17	rs4289035	45707211	C	T	1.12	1.08	1.16	2.96E-10
17	rs12943464	45690351	T	A	1.12	1.08	1.16	2.98E-10
17	rs4793978	45698175	A	G	1.12	1.08	1.16	3.68E-10
17	rs7223029	45701731	C	T	1.12	1.08	1.16	3.68E-10
17	rs11079784	45702280	C	T	1.12	1.08	1.16	3.68E-10
17	rs4794051	45577914	A	G	1.12	1.08	1.16	4.44E-10
17	rs4060785	45576936	C	G	1.12	1.08	1.16	4.67E-10
17	rs11869940	45578435	A	T	1.12	1.08	1.16	4.68E-10
17	rs12103836	45577469	G	A	1.12	1.08	1.16	4.88E-10
17	rs8074149	45583375	C	A	1.12	1.08	1.16	5.03E-10
17	rs62076549	45579555	G	T	1.12	1.08	1.16	5.40E-10
17	rs7221548	45582952	C	T	1.12	1.08	1.16	5.50E-10
17	rs10853090	45716124	T	C	1.12	1.08	1.16	5.83E-10
17	rs55633793	45719816	G	A	1.12	1.08	1.16	6.91E-10
17	rs12051716	45582395	A	G	1.12	1.07	1.16	1.03E-09
17	exm2264630	45732774	A	C	1.12	1.07	1.16	1.03E-09
17	rs7219303	45737701	G	A	1.12	1.07	1.16	1.03E-09
17	rs73315878	45744993	A	G	1.12	1.07	1.16	1.08E-09
17	rs4794372	45627760	A	G	1.12	1.07	1.16	1.12E-09
17	rs4794371	45627693	C	G	1.12	1.07	1.16	1.12E-09
17	rs4399567	45749172	G	C	1.12	1.07	1.16	1.17E-09
17	rs9905922	45740805	A	G	1.12	1.07	1.16	1.20E-09
17	rs56226803	45742233	T	A	1.12	1.07	1.16	1.20E-09
17	rs4239163	45755954	C	T	1.11	1.07	1.16	1.37E-09
17	rs4239162	45755810	G	A	1.11	1.07	1.16	1.43E-09
17	rs8072100	45674687	A	T	0.90	0.86	0.93	2.11E-09
17	rs8077106	45653291	T	G	0.90	0.87	0.94	3.27E-09
17	rs12603290	45650196	T	C	0.90	0.87	0.94	3.29E-09
17	rs8081717	45653921	T	A	0.90	0.87	0.94	3.76E-09
17	rs4794001	45720849	T	C	1.11	1.07	1.15	4.38E-09
17	rs7220935	45703433	C	T	0.90	0.87	0.94	4.87E-09
17	rs4793842	45671506	A	G	1.11	1.07	1.15	5.83E-09
17	rs4794048	45763073	C	A	0.90	0.87	0.94	5.87E-09
17	rs11870935	45732605	G	A	0.90	0.87	0.94	6.17E-09
17	rs4375701	45712218	T	C	0.90	0.87	0.94	6.44E-09
17	rs8067286	45594590	A	G	0.90	0.87	0.94	6.66E-09
17	rs3809868	45750596	G	A	0.90	0.87	0.94	6.73E-09
17	rs2317826	45730899	T	C	0.90	0.87	0.94	6.97E-09
17	rs78551919	45626002	A	C	0.90	0.87	0.94	7.04E-09
17	rs7210738	45757598	G	A	0.90	0.87	0.94	7.04E-09
17	rs12150231	45717091	G	A	0.90	0.87	0.94	7.10E-09
17	rs11869730	45614763	T	A	0.90	0.87	0.94	7.12E-09
17	rs8065437	45721653	T	C	0.90	0.87	0.94	7.14E-09
17	rs7222851	45592169	T	C	0.90	0.87	0.94	7.15E-09
17	rs7350906	45588112	G	A	0.90	0.87	0.94	7.26E-09
17	rs11650372	45721795	A	G	0.90	0.87	0.94	7.36E-09
17	rs8079821	45611704	T	C	0.90	0.87	0.94	7.66E-09
17	rs9905583	45741016	T	C	0.90	0.87	0.94	7.70E-09
17	rs4264433	45737275	A	T	0.90	0.87	0.94	7.80E-09
17	rs8065669	45724996	A	G	0.90	0.87	0.94	7.87E-09
17	rs4580230	45738861	A	G	0.90	0.87	0.94	8.26E-09
17	rs8078686	45735706	C	T	1.11	1.07	1.15	1.10E-08
17	rs9903898	45575354	C	T	1.11	1.06	1.15	1.41E-08
17	rs12602134	45582136	G	C	1.11	1.06	1.15	1.63E-08
17	rs7405889	45571676	T	C	0.90	0.87	0.94	1.71E-08
17	rs62076542	45558090	T	C	0.91	0.87	0.94	2.14E-08
17	rs2085755	45545378	G	T	0.91	0.87	0.94	2.60E-08
17	rs9895274	45539117	T	C	0.91	0.87	0.94	2.77E-08
17	rs12452315	45479446	C	A	0.91	0.87	0.94	3.12E-08
17	rs4338849	45512483	G	A	0.91	0.87	0.94	3.58E-08
17	rs77454872	45509496	A	G	0.91	0.87	0.94	3.97E-08
17	rs11079775	45516667	T	C	0.91	0.87	0.94	4.10E-08
17	rs4794016	45725931	T	A	0.91	0.87	0.94	4.46E-08

Chromosome	SNP name	Position (bp)	Allele A	Allele B	Odds ratio (OR)	L95	U95	p-value
17	rs1808192	45794706	A	G	1.11	1.07	1.16	5.64E-08
17	rs7219420	45800299	T	C	1.11	1.07	1.16	7.64E-08
17	rs9913503	45776330	T	G	1.10	1.06	1.15	9.57E-08
17	rs8078309	45774072	T	C	1.10	1.06	1.15	9.58E-08
17	rs62074054	45771816	T	C	1.10	1.06	1.15	9.62E-08
17	rs6503804	45772447	G	A	1.10	1.06	1.15	9.62E-08
17	rs4794053	45769551	G	A	1.10	1.06	1.15	9.89E-08
17	rs6503807	45772932	T	C	1.10	1.06	1.15	1.01E-07
17	rs6505048	45674998	A	G	1.10	1.06	1.15	1.05E-07
17	rs4794052	45769236	T	C	1.10	1.06	1.15	1.05E-07
17	rs8075566	45771037	G	A	1.10	1.06	1.15	1.06E-07
17	rs2013383	45791123	T	G	1.10	1.06	1.15	1.18E-07
17	rs9894905	45485982	T	C	0.91	0.88	0.95	1.23E-07
17	rs11656056	45659161	C	G	1.10	1.06	1.15	1.63E-07
17	rs7206971	45425115	G	A	1.10	1.06	1.14	1.74E-07
17	rs11079783	45686016	A	G	1.10	1.06	1.15	1.88E-07
17	rs2611998	45604290	C	T	1.10	1.06	1.14	1.95E-07
17	rs6504963	45588864	A	G	1.10	1.06	1.14	2.43E-07
17	rs3760371	45695914	C	T	1.10	1.06	1.14	2.83E-07
17	rs62076461	45440269	C	T	1.09	1.05	1.14	3.23E-07
17	rs2317631	45440385	C	T	1.09	1.05	1.14	3.23E-07
17	rs6504872	45438952	C	T	1.09	1.05	1.14	3.53E-07
17	rs11656856	45714671	T	C	1.10	1.06	1.14	3.53E-07
17	rs11650210	45579398	T	G	1.10	1.06	1.14	3.60E-07
17	rs6505062	45703264	T	C	1.10	1.06	1.14	3.81E-07
17	17:45744652[b37]C,T	45744652	T	C	1.10	1.05	1.14	7.27E-07
17	rs6505049	45675781	A	G	1.09	1.05	1.14	1.03E-06
17	rs12601144	45565592	T	C	1.09	1.05	1.14	1.30E-06
17	rs7222225	45693497	G	A	1.09	1.05	1.14	1.31E-06
17	rs4793836	45670312	A	G	1.09	1.05	1.13	1.60E-06
17	rs1138193	45651056	T	C	1.09	1.05	1.13	1.74E-06
17	rs11870585	45618812	A	G	1.09	1.05	1.13	1.91E-06
17	rs72829689	45593661	C	T	1.09	1.05	1.13	1.97E-06
17	rs1962307	45707871	T	C	1.09	1.05	1.13	1.98E-06
17	rs11870057	45588703	A	G	1.09	1.05	1.13	2.18E-06
17	rs35159316	45590257	T	C	1.09	1.05	1.13	2.18E-06
17	rs7214799	45702099	T	C	1.09	1.05	1.13	2.47E-06
17	rs7208530	45544107	C	T	1.09	1.05	1.13	2.57E-06
17	rs12451409	45544856	A	G	1.09	1.05	1.13	2.66E-06
17	rs6503796	45765249	T	C	1.09	1.05	1.13	3.55E-06
17	rs8079744	45554571	A	G	1.09	1.04	1.13	5.42E-06
17	rs9907159	45549741	T	G	1.09	1.04	1.13	5.54E-06
17	rs9912592	45556945	G	A	1.09	1.04	1.13	5.54E-06
17	17:45548958[b37]A,C	45548958	C	A	1.09	1.04	1.13	5.73E-06
17	rs9905581	45547185	C	T	1.09	1.04	1.13	5.81E-06
17	rs9908016	45551361	G	A	1.09	1.04	1.13	5.81E-06
17	rs9892497	45545919	G	A	1.09	1.04	1.13	6.00E-06
17	rs9912101	45547078	T	C	1.09	1.04	1.13	6.00E-06
17	rs9902574	45550662	C	T	1.09	1.04	1.13	6.00E-06
17	rs28507314	45551402	C	T	1.09	1.04	1.13	6.00E-06
17	rs28458768	45552140	A	G	1.09	1.04	1.13	6.00E-06
17	rs12944103	45553267	G	C	1.09	1.04	1.13	6.00E-06
17	rs59231671	45557594	T	C	1.08	1.04	1.13	6.06E-06
17	rs8076778	45554786	G	A	1.08	1.04	1.13	6.13E-06
17	rs62076540	45557784	C	T	1.08	1.04	1.13	6.65E-06
17	rs3809863	45385012	T	C	0.92	0.89	0.96	6.66E-06
17	rs9910408	45802874	G	A	1.09	1.05	1.14	6.75E-06
17	rs9303533	45386540	G	A	0.92	0.89	0.96	6.92E-06
17	rs9674670	45386157	A	G	0.92	0.89	0.96	7.08E-06
17	rs7213086	45752310	C	G	1.08	1.04	1.13	7.18E-06
17	rs11655943	45385640	C	T	0.92	0.89	0.96	7.82E-06
17	rs9913206	45528884	C	T	1.08	1.04	1.13	1.24E-05
17	rs72823464	45318090	A	G	1.12	1.07	1.19	1.30E-05
17	rs4968259	45528177	T	C	1.08	1.04	1.13	1.47E-05
17	17:45781264[b37]A,G	45781264	A	G	1.08	1.04	1.13	1.50E-05
17	rs12600603	45390951	C	G	0.93	0.89	0.96	1.50E-05
17	17:45527574[b37]G,T	45527574	T	C	1.18	1.09	1.27	1.53E-05
17	rs757060	45794204	T	G	1.08	1.04	1.13	1.55E-05
17	rs9891111	45709375	C	G	1.08	1.04	1.12	1.55E-05
17	rs4794047	45763005	A	T	1.08	1.04	1.12	1.79E-05
17	rs9900123	45549055	T	C	1.08	1.04	1.12	2.15E-05
17	rs9904967	45536875	G	A	1.08	1.04	1.13	2.19E-05
17	rs11079776	45546893	C	T	1.08	1.04	1.12	2.32E-05
17	rs9895509	45549618	A	T	1.08	1.04	1.12	2.32E-05
17	rs7223246	45759767	T	C	0.91	0.87	0.95	2.35E-05
17	rs9904664	45559659	A	G	1.08	1.04	1.12	2.43E-05
17	rs9905073	45468296	G	A	1.08	1.04	1.12	2.83E-05
17	rs7221196	45374994	G	A	1.08	1.04	1.12	2.99E-05
17	rs8075058	45482857	A	G	1.08	1.04	1.12	3.08E-05
17	exm1332165	45451894	G	A	1.08	1.04	1.12	3.23E-05
17	rs7220606	45374795	A	G	1.08	1.04	1.12	3.35E-05
17	rs1912485	45515678	T	G	1.08	1.04	1.12	3.54E-05
17	rs9915222	45448615	G	T	1.08	1.04	1.12	3.60E-05
17	rs11079787	45813225	C	A	1.07	1.03	1.12	3.84E-05
17	rs1912486	45515167	A	G	1.08	1.04	1.12	3.97E-05
17	rs9916007	45390472	C	T	1.08	1.04	1.12	3.98E-05
17	rs9807106	45457102	G	T	1.08	1.04	1.12	4.15E-05
17	rs2136750	45456100	A	G	1.08	1.04	1.12	4.30E-05
17	rs2292346	45455473	G	A	1.08	1.04	1.12	4.36E-05
17	rs9894594	45461383	A	G	1.08	1.04	1.12	4.41E-05
17	rs9894179	45565093	C	T	1.07	1.03	1.12	4.42E-05
17	17:45477573[b37]C,T	45477573	C	T	1.08	1.04	1.12	4.53E-05
17	rs8078974	45815525	T	A	1.07	1.03	1.12	4.54E-05
17	rs67688344	45546841	G	A	1.07	1.03	1.12	4.61E-05
17	rs2074190	45811210	G	A	1.08	1.04	1.13	4.67E-05
17	rs11652969	45817367	A	G	1.07	1.03	1.12	4.73E-05
17	rs56356751	45557025	A	G	1.07	1.03	1.12	4.80E-05

Supplementary Table 3.2: Frequentist Analysis suggestive significant SNPs (n = 193).

Chromosome	SNP name	Position (bp)	Allele A	Allele B	Odds Ratio (OR)	L95	U95	log10(Bayes Factor)	Posterior Probability	∑(Posterior Probability)
17	rs4794057	45786452	C	T	1.13	1.08	1.17	8.61	0.09	0.09
17	exm-rs8070463	45768836	T	C	1.13	1.08	1.17	8.59	0.09	0.17
17	rs11651766	45574931	G	T	1.13	1.09	1.17	8.49	0.07	0.24
17	rs67919208	45771873	T	C	1.13	1.08	1.17	8.48	0.07	0.31
17	rs9901869	45575206	G	A	1.13	1.09	1.17	8.45	0.06	0.37
17	rs12937560	45671869	T	C	1.13	1.08	1.17	8.23	0.04	0.41
17	rs8075411	45676760	T	C	1.13	1.08	1.17	8.23	0.04	0.45
17	rs8078880	45676526	C	G	1.13	1.08	1.17	8.15	0.03	0.48
17	rs4793908	45678408	A	G	1.13	1.08	1.17	8.11	0.03	0.51
17	rs4793913	45678579	T	C	1.12	1.08	1.17	8.06	0.03	0.53
17	rs8072644	45653364	T	A	1.12	1.08	1.17	8.01	0.02	0.56
17	rs4092460	45667550	A	G	1.12	1.08	1.17	7.99	0.02	0.58
17	rs62074015	45668760	T	A	1.12	1.08	1.17	7.97	0.02	0.60
17	rs10163469	45687767	A	G	1.12	1.08	1.17	7.96	0.02	0.62
17	rs11079779	45640770	T	C	1.12	1.08	1.17	7.95	0.02	0.64
17	rs8082284	45647067	C	T	1.12	1.08	1.17	7.95	0.02	0.66
17	rs9635762	45648446	C	A	1.12	1.08	1.17	7.95	0.02	0.68
17	rs9635763	45649694	A	C	1.12	1.08	1.17	7.95	0.02	0.70
17	rs4439799	45781599	C	T	1.12	1.08	1.16	7.93	0.02	0.72
17	rs4561509	45781799	G	A	1.12	1.08	1.16	7.86	0.02	0.73
17	rs55687600	45630497	T	C	1.12	1.08	1.17	7.85	0.02	0.75
17	rs4794058	45597098	T	C	1.12	1.08	1.17	7.81	0.01	0.76
17	rs10432035	45624913	C	G	1.12	1.08	1.17	7.77	0.01	0.78
17	rs79443387	45626961	T	C	1.12	1.08	1.17	7.77	0.01	0.79
17	rs8070759	45638760	T	C	1.12	1.08	1.17	7.75	0.01	0.80
17	rs3760370	45695832	T	C	1.12	1.08	1.17	7.73	0.01	0.81
17	rs7207542	45697549	G	C	1.12	1.08	1.17	7.73	0.01	0.83
17	rs62074055	45771933	G	C	1.12	1.08	1.16	7.71	0.01	0.84
17	rs4794007	45590722	A	G	1.12	1.08	1.17	7.70	0.01	0.85
17	rs8065990	45591166	A	G	1.12	1.08	1.17	7.70	0.01	0.86
17	rs2125873	45660887	G	T	1.12	1.08	1.17	7.69	0.01	0.87
17	rs4794392	45628693	T	A	1.12	1.08	1.17	7.63	0.01	0.88
17	rs11656855	45714655	T	C	1.12	1.08	1.16	7.61	0.01	0.89
17	rs4289035	45707211	C	T	1.12	1.08	1.16	7.55	0.01	0.90
17	rs12943464	45690351	T	A	1.12	1.08	1.16	7.54	0.01	0.90
17	rs4793978	45698175	A	G	1.12	1.08	1.16	7.45	0.01	0.91
17	rs7223029	45701731	C	T	1.12	1.08	1.16	7.45	0.01	0.92
17	rs11079784	45702280	C	T	1.12	1.08	1.16	7.45	0.01	0.92
17	rs4794051	45577914	A	G	1.12	1.08	1.16	7.38	0.01	0.93
17	rs4060785	45576936	C	G	1.12	1.08	1.16	7.36	0.01	0.93
17	rs11869940	45578435	A	T	1.12	1.08	1.16	7.36	0.01	0.94
17	rs12103836	45577469	G	A	1.12	1.08	1.16	7.34	0.00	0.94
17	rs8074149	45583375	C	A	1.12	1.08	1.16	7.32	0.00	0.95
17	rs62076549	45579555	G	T	1.12	1.08	1.16	7.29	0.00	0.95
17	rs7221548	45582952	C	T	1.12	1.08	1.16	7.29	0.00	0.96
17	rs10853090	45716124	T	C	1.12	1.08	1.16	7.26	0.00	0.96
17	rs55633793	45719816	G	A	1.12	1.08	1.16	7.19	0.00	0.96
17	rs12051716	45582395	A	G	1.12	1.07	1.16	7.02	0.00	0.97
17	exm2264630	45732774	A	C	1.12	1.07	1.16	7.02	0.00	0.97
17	rs7219303	45737701	G	A	1.12	1.07	1.16	7.02	0.00	0.97
17	rs73315878	45744993	A	G	1.12	1.07	1.16	7.00	0.00	0.97
17	rs4794372	45627760	A	G	1.12	1.07	1.16	6.99	0.00	0.97
17	rs4794371	45627693	C	G	1.12	1.07	1.16	6.99	0.00	0.98
17	rs4399567	45749172	G	C	1.12	1.07	1.16	6.97	0.00	0.98
17	rs9905922	45740805	A	G	1.12	1.07	1.16	6.96	0.00	0.98
17	rs56226803	45742233	T	A	1.12	1.07	1.16	6.96	0.00	0.98
17	rs4239163	45755954	C	T	1.11	1.07	1.16	6.90	0.00	0.98
17	rs4239162	45755810	G	A	1.11	1.07	1.16	6.88	0.00	0.99
17	rs8072100	45674687	A	T	0.90	0.86	0.93	6.72	0.00	0.99
17	rs8077106	45653291	T	G	0.90	0.87	0.94	6.53	0.00	0.99
17	rs12603290	45650196	T	C	0.90	0.87	0.94	6.53	0.00	0.99
17	rs8081717	45653921	T	A	0.90	0.87	0.94	6.47	0.00	0.99

Supplementary Table 3.3: Bayesian Analysis 99% credible set (n = 62).

Chromosome	SNP name	Position (bp)	Allele A	Allele B	Odds Ratio (OR)	L95	U95	log10(Bayes Factor)	Posterior Probability	∑(Posterior Probability)
17	rs4794057	45786452	C	T	1.13	1.08	1.17	8.61	0.09	0.09
17	exm-rs8070463	45768836	T	C	1.13	1.08	1.17	8.59	0.09	0.17
17	rs11651766	45574931	G	T	1.13	1.09	1.17	8.49	0.07	0.24
17	rs67919208	45771873	T	C	1.13	1.08	1.17	8.48	0.07	0.31
17	rs9901869	45575206	G	A	1.13	1.09	1.17	8.45	0.06	0.37
17	rs12937560	45671869	T	C	1.13	1.08	1.17	8.23	0.04	0.41
17	rs8075411	45676760	T	C	1.13	1.08	1.17	8.23	0.04	0.45
17	rs8078880	45676526	C	G	1.13	1.08	1.17	8.15	0.03	0.48
17	rs4793908	45678408	A	G	1.13	1.08	1.17	8.11	0.03	0.51
17	rs4793913	45678579	T	C	1.12	1.08	1.17	8.06	0.03	0.53
17	rs8072644	45653364	T	A	1.12	1.08	1.17	8.01	0.02	0.56
17	rs4092460	45667550	A	G	1.12	1.08	1.17	7.99	0.02	0.58
17	rs62074015	45668760	T	A	1.12	1.08	1.17	7.97	0.02	0.60
17	rs10163469	45687767	A	G	1.12	1.08	1.17	7.96	0.02	0.62
17	rs11079779	45640770	T	C	1.12	1.08	1.17	7.95	0.02	0.64
17	rs8082284	45647067	C	T	1.12	1.08	1.17	7.95	0.02	0.66
17	rs9635762	45648446	C	A	1.12	1.08	1.17	7.95	0.02	0.68
17	rs9635763	45649694	A	C	1.12	1.08	1.17	7.95	0.02	0.70
17	rs4439799	45781599	C	T	1.12	1.08	1.16	7.93	0.02	0.72
17	rs4561509	45781799	G	A	1.12	1.08	1.16	7.86	0.02	0.73
17	rs55687600	45630497	T	C	1.12	1.08	1.17	7.85	0.02	0.75
17	rs4794058	45597098	T	C	1.12	1.08	1.17	7.81	0.01	0.76
17	rs10432035	45624913	C	G	1.12	1.08	1.17	7.77	0.01	0.78
17	rs79443387	45626961	T	C	1.12	1.08	1.17	7.77	0.01	0.79
17	rs8070759	45638760	T	C	1.12	1.08	1.17	7.75	0.01	0.80
17	rs3760370	45695832	T	C	1.12	1.08	1.17	7.73	0.01	0.81
17	rs7207542	45697549	G	C	1.12	1.08	1.17	7.73	0.01	0.83
17	rs62074055	45771933	G	C	1.12	1.08	1.16	7.71	0.01	0.84
17	rs4794007	45590722	A	G	1.12	1.08	1.17	7.70	0.01	0.85
17	rs8065990	45591166	A	G	1.12	1.08	1.17	7.70	0.01	0.86
17	rs2125873	45660887	G	T	1.12	1.08	1.17	7.69	0.01	0.87
17	rs4794392	45628693	T	A	1.12	1.08	1.17	7.63	0.01	0.88
17	rs11656855	45714655	T	C	1.12	1.08	1.16	7.61	0.01	0.89
17	rs4289035	45707211	C	T	1.12	1.08	1.16	7.55	0.01	0.90
17	rs12943464	45690351	T	A	1.12	1.08	1.16	7.54	0.01	0.90
17	rs4793978	45698175	A	G	1.12	1.08	1.16	7.45	0.01	0.91
17	rs7223029	45701731	C	T	1.12	1.08	1.16	7.45	0.01	0.92
17	rs11079784	45702280	C	T	1.12	1.08	1.16	7.45	0.01	0.92
17	rs4794051	45577914	A	G	1.12	1.08	1.16	7.38	0.01	0.93
17	rs4060785	45576936	C	G	1.12	1.08	1.16	7.36	0.01	0.93
17	rs11869940	45578435	A	T	1.12	1.08	1.16	7.36	0.01	0.94
17	rs12103836	45577469	G	A	1.12	1.08	1.16	7.34	0.00	0.94
17	rs8074149	45583375	C	A	1.12	1.08	1.16	7.32	0.00	0.95

Supplementary Table 3.4: Bayesian Analysis 95% credible set (n = 43).

8.2.3 Chapter 4 supplementary tables

SNP	Position on Chr17	RefSeq Gene	Functional annotation	Reference allele	Alternative allele	AFR	AMR	ASN	EUR	SiPhy conservation score	GERP conservation score	Chromatin marks	eQTL	GWAS
rs12452315	45479446	*C17orf57*	Intronic	A	C	0.43	0.4	0.27	0.47	0	0	H3K4me3	TRUE	.
rs77454872	45509496	*C17orf57*	Intronic	G	A	0.42	0.4	0.27	0.47	0	0		TRUE	.
rs4338849	45512483	*C17orf57*	Intronic	A	G	0.42	0.4	0.27	0.47	0	0	H3K4me3	TRUE	.
rs11079775	45516667	*C17orf57*	Intronic	C	T	0.42	0.4	0.27	0.47	0	0	H3K4me3	TRUE	.
rs9895274	45539117	*MRPL45P2*	Intronic	C	T	0.43	0.41	0.28	0.47	0	0		TRUE	.
rs2085755	45545378	*MRPL45P2*	Intronic	T	G	0.42	0.41	0.27	0.47	0	0	H3K4me1	TRUE	.
rs62076542	45558090	*MRPL45P2*	Intronic	C	T	0.4	0.38	0.26	0.45	0	0		TRUE	.
rs7405889	45571676	*MRPL45P2*	.	C	T	0.42	0.4	0.34	0.46	0	0	H3K27ac	TRUE	.
rs11651766	45574931	*MRPL45P2*	.	G	T	0.65	0.51	0.64	0.5	0	0	H3K4me1	TRUE	.
rs9901869	45575206	*MRPL45P2*	.	G	A	0.64	0.51	0.63	0.49	0	0	H3K4me1	TRUE	.
rs9903898	45575354	*MRPL45P2*	.	C	T	0.48	0.4	0.33	0.46	0	0	H3K4me1	TRUE	.
rs12103836	45577469	*MRPL45P2*	.	G	A	0.61	0.47	0.42	0.46	0	0	H3K27ac	TRUE	.
rs4794051	45577914	*MRPL45P2*	.	A	G	0.62	0.48	0.44	0.46	0	0	H3K9ac	TRUE	.
rs11869940	45578435	*MRPL45P2*	.	T	A	0.28	0.49	0.54	0.5	0	0	H3K4me1	TRUE	.
rs62076549	45579555	*MRPL45P2*	.	T	G	0.31	0.51	0.54	0.51	0	0	H3K4me1	TRUE	.
rs12602134	45582136	*MRPL45P2*	.	C	G	0.48	0.6	0.66	0.54	0	0	H3K27ac	TRUE	.
rs12051716	45582395	*MRPL45P2*	.	G	A	0.37	0.54	0.55	0.53	0	0	H3K27ac	TRUE	.
rs7221548	45582952	*MRPL45P2*	.	T	C	0.31	0.51	0.55	0.52	0	0	H3K4me1	TRUE	.
rs8074149	45583375	*MRPL45P2*	.	A	C,G,T	0.33	0.51	0.55	0.52	0	0	H3K4me1	TRUE	.
rs7350906	45588112	*MRPL45P2*	.	G	A	0.49	0.6	0.67	0.54	0	0	H3K27ac	TRUE	.
rs4794007	45590722	*NPEPPS*	.	G	A	0.3	0.48	0.53	0.51	0	0	H3K4me1	TRUE	.
rs8065990	45591166	*NPEPPS*	.	G	A	0.31	0.51	0.53	0.52	0	0	H3K4me1	TRUE	.
rs7222851	45592169	*NPEPPS*	.	T	C	0.5	0.56	0.64	0.52	0	0		TRUE	.
rs8067286	45594590	*NPEPPS*	.	A	G	0.47	0.6	0.65	0.53	0	0	H3K27ac	TRUE	.
rs4794058	45597098	*NPEPPS*	.	C	T	0.35	0.53	0.55	0.53	0	0	H3K4me1	TRUE	.
rs8079821	45611704	*NPEPPS*	Intronic	T	C	0.43	0.55	0.62	0.49	0	0	H3K4me1	TRUE	.
rs11869730	45614763	*NPEPPS*	Intronic	T	A	0.44	0.55	0.62	0.49	0	0	H3K27ac	TRUE	.
rs10432035	45624913	*NPEPPS*	Intronic	G	C	0.28	0.49	0.53	0.51	0	0		TRUE	.
rs78551919	45626002	*NPEPPS*	Intronic	A	C	0.46	0.59	0.64	0.54	0	0		TRUE	.
rs79443387	45626961	*NPEPPS*	Intronic	C	T	0.28	0.49	0.53	0.49	0	0		TRUE	.
rs4794371	45627693	*NPEPPS*	Intronic	G	C	0.32	0.51	0.53	0.51	1	0		TRUE	.
rs4794372	45627760	*NPEPPS*	Intronic	G	A	0.32	0.52	0.54	0.53	0	0		TRUE	.
rs55687600	45630497	*NPEPPS*	Intronic	C	T	0.28	0.5	0.54	0.51	0	0	H3K4me1	TRUE	.
rs8070759	45638760	*NPEPPS*	Intronic	C	T	0.35	0.53	0.55	0.53	0	0	H3K4me1	TRUE	.
rs11079779	45640770	*NPEPPS*	Intronic	C	T	0.28	0.48	0.52	0.5	0	0	H3K4me1	TRUE	.
rs8082284	45647067	*NPEPPS*	Intronic	T	C	0.28	0.5	0.53	0.5	0	0		TRUE	.
rs9635762	45648446	*NPEPPS*	Intronic	A	C	0.28	0.5	0.53	0.51	0	1		TRUE	.
rs9635763	45649694	*NPEPPS*	Intronic	C	A	0.28	0.49	0.53	0.51	0	0	H3K27ac	TRUE	.
rs12603290	45650196	*NPEPPS*	Intronic	T	C	0.45	0.59	0.64	0.53	0	0	H3K27ac	TRUE	.
rs8077106	45653291	*NPEPPS*	Intronic	T	G	0.5	0.6	0.66	0.54	0	0	H3K4me1	TRUE	.
rs8072644	45653364	*NPEPPS*	Intronic	A	T	0.29	0.5	0.53	0.51	0	0	H3K4me1	TRUE	.
rs8081717	45653921	*NPEPPS*	Intronic	T	A	0.5	0.6	0.66	0.54	0	0		TRUE	.
rs4092460	45667550	*NPEPPS*	Intronic	G	A	0.29	0.5	0.53	0.5	0	0		TRUE	.
rs4793842	45671506	*NPEPPS*	Intronic	G	A	0.29	0.51	0.54	0.53	0	0	H3K4me1	TRUE	.
rs12937560	45671869	*NPEPPS*	Intronic	C	T	0.29	0.51	0.51	0.51	0	0		TRUE	.
rs8072100	45674687	*NPEPPS*	Intronic	A	T	0.53	0.6	0.65	0.53	0	0	H3K27ac	TRUE	.
rs8078880	45676526	*NPEPPS*	Intronic	G	C	0.32	0.52	0.54	0.53	0	0	H3K27ac	TRUE	.
rs8075411	45676760	*NPEPPS*	Intronic	C	T	0.3	0.52	0.54	0.53	0	0	H3K27ac	TRUE	.
rs4793908	45678408	*NPEPPS*	Intronic	G	A	0.29	0.51	0.54	0.52	0	0	H3K9ac	TRUE	.

SNP	Position on Chr17	RefSeq Gene	Functional annotation	Reference allele	Alternative allele	AFR	AMR	ASN	EUR	SiPhy conservation score	GERP conservation score	Chromatin marks	eQTL	GWAS
rs4793913	45678579	*NPEPPS*	Intronic	C	T	0.3	0.52	0.54	0.52	0	0	H3K9ac	TRUE	.
rs10163469	45687767	*NPEPPS*	Intronic	G	A	0.29	0.51	0.54	0.51	0	0	H3K27ac	TRUE	.
rs12943464	45690351	*NPEPPS*	Intronic	A	T	0.32	0.53	0.54	0.53	0	0	H3K27ac	TRUE	.
rs3760370	45695832	*NPEPPS*	Intronic	C	T	0.3	0.52	0.54	0.53	1	0	H3K4me1	TRUE	.
rs7207542	45697549	*NPEPPS*	Intronic	C	G	0.3	0.52	0.54	0.52	0	0	H3K4me1	TRUE	.
rs4793978	45698175	*NPEPPS*	Intronic	G	A	0.32	0.53	0.54	0.53	0	0	H3K4me1	TRUE	.
rs7223029	45701731	*NPEPPS*	.	T	C	0.3	0.53	0.54	0.53	0	0	H3K4me1	TRUE	.
rs11079784	45702280	*NPEPPS*	.	T	C	0.32	0.52	0.54	0.53	0	0	H3K27ac	TRUE	.
rs7220935	45703433	*NPEPPS*	.	C	T	0.47	0.6	0.65	0.53	0	0	H3K27ac	TRUE	.
rs4289035	45707211	*NPEPPS*	.	C	T	0.64	0.48	0.46	0.47	0	0	H3K9ac	TRUE	.
rs4375701	45712218	*NPEPPS*	.	C	T	0.46	0.4	0.35	0.46	0	0	H3K27ac	TRUE	.
rs11656855	45714655	*KPNB1*	.	C	T	0.3	0.52	0.54	0.53	0	0	H3K4me1	TRUE	.
rs10853090	45716124	*KPNB1*	.	C	T	0.33	0.53	0.54	0.53	0	0	H3K4me1	TRUE	.
rs12150231	45717091	*KPNB1*	.	G	A	0.48	0.6	0.65	0.53	0	0	H3K9ac	TRUE	.
rs55633793	45719816	*KPNB1*	.	G	A	0.6	0.47	0.46	0.47	0	0	H3K27ac	TRUE	.
rs4794001	45720849	*KPNB1*	.	T	C	0.54	0.46	0.46	0.47	1	0	H3K27ac	TRUE	.
rs8065437	45721653	*KPNB1*	.	C	T	0.46	0.4	0.35	0.47	0	0	H3K27ac	TRUE	.
rs11650372	45721795	*KPNB1*	.	G	A	0.4	0.39	0.35	0.47	0	0	H3K27ac	TRUE	.
rs8065669	45724996	*KPNB1*	.	G	A	0.4	0.39	0.35	0.47	0	0	H3K4me1	TRUE	.
rs4794016	45725931	*KPNB1*	.	T	A	0.46	0.6	0.65	0.53	0	0	H3K4me3	TRUE	.
rs2317826	45730899	*KPNB1*	Intronic	C	T	0.48	0.4	0.35	0.47	0	0	H3K4me1	TRUE	.
rs11870935	45732605	*KPNB1*	Intronic	G	A	0.46	0.6	0.65	0.53	1	1	H3K4me1	TRUE	.
rs11871606	45732774	*KPNB1*	Intronic	C	A	0.37	0.53	0.54	0.53	0	0	H3K4me1	TRUE	.
rs8078686	45735706	*KPNB1*	Intronic	T	C	0.46	0.59	0.67	0.51	0	0	H3K9ac	TRUE	.
rs4264433	45737275	*KPNB1*	Intronic	A	T	0.49	0.57	0.64	0.53	0	0	H3K27ac	TRUE	.
rs7219303	45737701	*KPNB1*	Intronic	A	G	0.37	0.53	0.54	0.53	0	0	H3K9ac	TRUE	.
rs4580230	45738861	*KPNB1*	Intronic	G	A	0.48	0.4	0.35	0.46	0	0	H3K4me3	TRUE	.
rs9905922	45740805	*KPNB1*	Intronic	A	G	0.57	0.47	0.46	0.47	0	0	H3K4me1	TRUE	.
rs9905583	45741016	*KPNB1*	Intronic	C	T	0.47	0.4	0.35	0.47	0	0	H3K4me1	TRUE	.
rs56226803	45742233	*KPNB1*	Intronic	T	A	0.57	0.47	0.46	0.47	0	0	H3K4me1	TRUE	.
rs73315878	45744993	*KPNB1*	Intronic	G	A	0.37	0.52	0.54	0.52	0	0	H3K27ac	TRUE	.
rs4399567	45749172	*KPNB1*	Intronic	C	G	0.37	0.53	0.54	0.53	0	0	H3K27ac	TRUE	.
rs3809868	45750596	*KPNB1*	Intronic	G	A	0.46	0.6	0.65	0.53	0	0	H3K27ac	TRUE	.
rs4239162	45755810	*KPNB1*	Intronic	G	A	0.57	0.46	0.46	0.47	0	0	H3K27ac	TRUE	.
rs4239163	45755954	*KPNB1*	Intronic	C	T	0.57	0.46	0.46	0.47	0	0	H3K27ac	TRUE	.
rs7210738	45757598	*KPNB1*	Intronic	G	A	0.46	0.6	0.65	0.53	0	0	H3K27ac	TRUE	.
rs4794048	45763073	*KPNB1*	.	A	C	0.4	0.39	0.35	0.47	0	0	H3K4me1	TRUE	.
rs8070463	45768836	*TBKBP1*	.	T	C	0.63	0.47	0.45	0.5	0	0	H3K4me1	TRUE	Ankylosing spondylitis, Multiple sclerosis
rs67919208	45771873	*TBKBP1*	.	T	C	0.63	0.47	0.44	0.5	1	0	H3K4me3	TRUE	.
rs62074055	45771933	*TBKBP1*	.	G	C	0.54	0.4	0.33	0.49	0	0	H3K4me3	TRUE	.
rs4439799	45781599	*TBKBP1*	Intronic	C	T	0.44	0.4	0.34	0.49	0	0	H3K4me3	TRUE	.
rs4561509	45781799	*TBKBP1*	Intronic	G	A	0.42	0.4	0.34	0.49	1	1	H3K9ac	TRUE	.
rs4794057	45786452	*TBKBP1*	Synonymous	C	T	0.64	0.45	0.48	0.51	1	1	H3K4me3	TRUE	.

Supplementary Table 4.1: HaploReg annotation of all prioritised SNPs.

SNP	Position	Score	Rank	ChIP	DNase	Footprint	QTL	RegulomeDB Rank
rs11079784	45702280	0.9	1b	TRUE	TRUE	TRUE	TRUE	1
rs4794048	45763073	0.45933	1b	TRUE	TRUE	TRUE	TRUE	2
rs8070463	45768836	0.55436	1f	TRUE	TRUE	FALSE	TRUE	3
rs11871606	45732774	0.55324	1f	TRUE	FALSE	FALSE	TRUE	4
rs8072644	45653364	0.22388	1f	FALSE	TRUE	FALSE	TRUE	5
rs8074149	45583375	0.04	1f	FALSE	TRUE	TRUE	TRUE	6
rs10163469	45687767	0.65734	2b	TRUE	TRUE	TRUE	FALSE	7
rs11870935	45732605	0.59816	2b	TRUE	TRUE	TRUE	FALSE	8
rs7219303	45737701	0.56896	2b	TRUE	TRUE	TRUE	FALSE	9
rs4793978	45698175	0.4975	2b	TRUE	TRUE	TRUE	FALSE	10
rs4794057	45786452	0.49614	2b	TRUE	TRUE	TRUE	FALSE	11
rs7210738	45757598	0.90505	3a	TRUE	TRUE	FALSE	FALSE	12
rs4561509	45781799	0.85537	3a	TRUE	TRUE	FALSE	FALSE	13
rs3809868	45750596	0.57429	3a	TRUE	TRUE	FALSE	FALSE	14
rs11869940	45578435	0.57068	3a	TRUE	TRUE	FALSE	FALSE	15
rs7405889	45571676	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4794051	45577914	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs62076549	45579555	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4794007	45590722	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs8079821	45611704	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs12603290	45650196	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4092460	45667550	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4793842	45671506	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs3760370	45695832	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs7220935	45703433	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4794001	45720849	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs8065437	45721653	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs11650372	45721795	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs8065669	45724996	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4794016	45725931	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs56226803	45742233	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4399567	45749172	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4239162	45755810	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs67919208	45771873	0.70497	4	TRUE	TRUE	TRUE	FALSE	16
rs62074055	45771933	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4439799	45781599	0.60906	4	TRUE	TRUE	FALSE	FALSE	16
rs4338849	45512483	1	5	TRUE	FALSE	FALSE	FALSE	37
rs7222851	45592169	0.94917	5	TRUE	FALSE	FALSE	FALSE	38
rs55633793	45719816	0.94917	5	TRUE	FALSE	FALSE	FALSE	38
rs8077106	45653291	0.9223	5	FALSE	TRUE	FALSE	FALSE	40
rs7223029	45701731	0.89167	5	TRUE	FALSE	FALSE	FALSE	41
rs9905922	45740805	0.76455	5	TRUE	FALSE	FALSE	FALSE	42
rs4794372	45627760	0.71021	5	TRUE	FALSE	TRUE	FALSE	43
rs7350906	45588112	0.65852	5	FALSE	TRUE	FALSE	FALSE	44
rs4794371	45627693	0.63783	5	TRUE	FALSE	FALSE	FALSE	45
rs62076542	45558090	0.63123	5	FALSE	TRUE	FALSE	FALSE	46
rs9635763	45649694	0.612	5	TRUE	FALSE	FALSE	FALSE	47
rs9895274	45539117	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs79443387	45626961	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs8070759	45638760	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs4793908	45678408	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs2317826	45730899	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs8078686	45735706	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs4580230	45738861	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs9905583	45741016	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs4239163	45755954	0.58955	5	TRUE	FALSE	FALSE	FALSE	48
rs4264433	45737275	0.42044	5	TRUE	FALSE	FALSE	FALSE	57
rs8067286	45594590	0.41613	5	FALSE	TRUE	FALSE	FALSE	58
rs10432035	45624913	0.30338	5	FALSE	TRUE	FALSE	FALSE	59
rs11651766	45574931	0.13454	5	FALSE	TRUE	FALSE	FALSE	60
rs78551919	45626002	0.13454	5	FALSE	TRUE	FALSE	FALSE	60
rs7207542	45697549	0.13454	5	FALSE	TRUE	FALSE	FALSE	60
rs4289035	45707211	0.13454	5	FALSE	TRUE	FALSE	FALSE	60
rs8075411	45676760	0.03026	5	FALSE	TRUE	FALSE	FALSE	64
rs55687600	45630497	0.012	5	FALSE	TRUE	TRUE	FALSE	65
rs8065990	45591166	0.01	5	FALSE	TRUE	TRUE	FALSE	66
rs9901869	45575206	0	5	FALSE	TRUE	FALSE	FALSE	67
rs8081717	45653921	0	5	FALSE	TRUE	FALSE	FALSE	67
rs9635762	45648446	0.34319	6	FALSE	FALSE	FALSE	FALSE	69
rs8072100	45674687	0.33176	6	FALSE	FALSE	FALSE	FALSE	70
rs12051716	45582395	0.27598	6	FALSE	FALSE	FALSE	FALSE	71
rs4375701	45712218	0.18347	6	FALSE	FALSE	FALSE	FALSE	72
rs12602134	45582136	0.114	6	FALSE	FALSE	TRUE	FALSE	73
rs11079775	45516667	0.11318	6	FALSE	FALSE	FALSE	FALSE	74
rs12937560	45671869	0.11093	6	FALSE	FALSE	FALSE	FALSE	75
rs12943464	45690351	0.01306	6	FALSE	FALSE	TRUE	FALSE	76
rs77454872	45509496	0	6	FALSE	FALSE	FALSE	FALSE	77
rs12452315	45479446	0.51392	7	FALSE	FALSE	FALSE	TRUE	78
rs2085755	45545378	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs9903898	45575354	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs12103836	45577469	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs4794058	45597098	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs11869730	45614763	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs11079779	45640770	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs8082284	45647067	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs2125873	45660887	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs62074015	45668760	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs8078880	45676526	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs4793913	45678579	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs11656855	45714655	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs12150231	45717091	0.18412	7	FALSE	FALSE	FALSE	FALSE	79
rs73315878	45744993	0.18412	7	FALSE	FALSE	FALSE	FALSE	79

Supplementary Table 4.2: RegulomeDB classification and ranking of all prioritised SNPs.

SNP	Position	Raw Score	PHRED	Rank	Consequence	Gene Name	FeatureID
rs67919208	45771873	2.19	21.30	1	REGULATORY	NA	ENSR00000283592
rs4794057	45786452	1.58	16.16		SYNONYMOUS	TBKBP1	ENST00000361722
rs11870935	45732605	1.57	16.13	3	INTRONIC	KPNB1	ENST00000290158
rs4561509	45781799	1.36	15.20	4	REGULATORY	NA	ENSR00000095199
rs4794058	45597098	1.05	13.60	5	UPSTREAM	NPEPPS	ENST00000530173
rs4794371	45627693	0.65	10.68	6	INTRONIC	NPEPPS	ENST00000322157
rs62074055	45771933	0.57	10.04	7	REGULATORY	NA	ENSR00000283592
rs56226803	45742233	0.45	8.94	8	INTRONIC	KPNB1	ENST00000290158
rs62076549	45579555	0.45	8.86	9	REGULATORY	NA	ENSR00000095183
rs4338849	45512483	0.44	8.84	10	INTRONIC	EFCAB13	ENST00000331493
rs8065669	45724996	0.43	8.69	11	INTRONIC	RP11-580I16.2	ENST00000580045
rs9903898	45575354	0.40	8.41	12	INTERGENIC	NA	NA
rs4794001	45720849	0.36	7.94	13	INTRONIC	RP11-580I16.2	ENST00000580045
rs4092460	45667550	0.35	7.78	14	INTRONIC	NPEPPS	ENST00000322157
rs3760370	45695832	0.32	7.50	15	INTRONIC	RP11-580I16.2	ENST00000582389
rs4060785	45576936	0.32	7.43	16	INTERGENIC	NA	NA
exm2264630	45732774	0.31	7.30	17	INTRONIC	KPNB1	ENST00000290158
rs8072644	45653364	0.29	7.03	18	INTRONIC	NPEPPS	ENST00000322157
rs8072100	45674687	0.27	6.84	19	INTRONIC	NPEPPS	ENST00000322157
rs62074015	45668760	0.26	6.66	20	INTRONIC	NPEPPS	ENST00000322157
rs9635763	45649694	0.25	6.55	21	INTRONIC	NPEPPS	ENST00000322157
rs8070759	45638760	0.24	6.46	22	INTRONIC	NPEPPS	ENST00000322157
rs7221548	45582952	0.21	5.93	23	INTERGENIC	NA	NA
rs7219303	45737701	0.15	5.10	24	REGULATORY	NA	ENSR00000095191
exm-rs8070463	45768836	0.15	5.03	25	UPSTREAM	TBKBP1	ENST00000361722
rs8075411	45676760	0.14	4.88	26	INTRONIC	NPEPPS	ENST00000322157
rs7405889	45571676	0.13	4.78	27	UPSTREAM	MRPL45P2	ENST00000425159
rs8082284	45647067	0.12	4.55	28	INTRONIC	NPEPPS	ENST00000322157
rs12452315	45479446	0.12	4.52	29	INTRONIC	EFCAB13	ENST00000331493
rs9635762	45648446	0.11	4.42	30	INTRONIC	NPEPPS	ENST00000322157
rs11869940	45578435	0.10	4.29	31	INTERGENIC	NA	NA
rs11079784	45702280	0.07	3.72	32	DOWNSTREAM	NPEPPS	ENST00000322157
rs10853090	45716124	0.06	3.68	33	INTRONIC	RP11-580I16.2	ENST00000580045
rs8067286	45594590	0.05	3.51	34	INTERGENIC	NA	NA
rs4794016	45725931	0.03	3.07	35	INTRONIC	RP11-580I16.2	ENST00000580045
rs12603290	45650196	0.02	2.99	36	INTRONIC	NPEPPS	ENST00000322157
rs4794051	45577914	0.00	2.66	37	REGULATORY	NA	ENSR00000283591
rs55687600	45630497	0.00	2.55	38	UPSTREAM	Y_RNA	ENST00000384007
rs4794048	45763073	-0.01	2.54	29	UPSTREAM	RP11-138C9.1	ENST00000578482
rs8077106	45653291	-0.01	2.45	40	INTRONIC	NPEPPS	ENST00000322157
rs4399567	45749172	-0.01	2.45	41	INTRONIC	KPNB1	ENST00000290158
rs12051716	45582395	-0.02	2.33	42	INTERGENIC	NA	NA
rs11650372	45721795	-0.05	1.93	43	REGULATORY	NA	ENSR00000376577
rs62076542	45558090	-0.06	1.87	44	INTRONIC	MRPL45P2	ENST00000425159
rs12937560	45671869	-0.06	1.79	45	INTRONIC	NPEPPS	ENST00000322157
rs11079779	45640770	-0.07	1.71	46	INTRONIC	NPEPPS	ENST00000322157
rs4793908	45678408	-0.07	1.71	47	DOWNSTREAM	RP11-580I16.2	ENST00000582389
rs8074149	45583375	-0.07	1.70	48	INTERGENIC	NA	NA
rs77454872	45509496	-0.07	1.66	49	INTRONIC	EFCAB13	ENST00000331493
rs9895274	45539117	-0.08	1.63	50	INTRONIC	MRPL45P2	ENST00000425159
rs8081717	45653921	-0.08	1.57	51	INTRONIC	NPEPPS	ENST00000322157
rs8078880	45676526	-0.10	1.43	52	DOWNSTREAM	RP11-580I16.2	ENST00000582389
rs4289035	45707211	-0.10	1.39	53	INTRONIC	RP11-580I16.2	ENST00000580045
rs4375701	45712218	-0.11	1.32	54	INTRONIC	RP11-580I16.2	ENST00000580045
rs4239162	45755810	-0.11	1.31	55	DOWNSTREAM	RP11-138C9.1	ENST00000578482
rs12150231	45717091	-0.11	1.31	56	INTRONIC	RP11-580I16.2	ENST00000580045
rs4580230	45738861	-0.14	1.05	57	INTRONIC	KPNB1	ENST00000290158
rs8079821	45611704	-0.15	1.00	58	INTRONIC	NPEPPS	ENST00000322157
rs2085755	45545378	-0.16	0.94	59	INTRONIC	MRPL45P2	ENST00000425159
rs79443387	45626961	-0.17	0.90	60	INTRONIC	NPEPPS	ENST00000322157
rs11656855	45714655	-0.17	0.90	61	INTRONIC	RP11-580I16.2	ENST00000580045
rs4439799	45781599	-0.17	0.85	62	REGULATORY	NA	ENSR00000095199
rs11651766	45574931	-0.18	0.83	63	INTERGENIC	NA	NA
rs2317826	45730899	-0.18	0.83	64	REGULATORY	NA	ENSR00000376578
rs12602134	45582136	-0.18	0.80	65	INTERGENIC	NA	NA
rs8065990	45591166	-0.18	0.79	66	INTERGENIC	NA	NA
rs7350906	45588112	-0.19	0.78	67	INTERGENIC	NA	NA
rs8078686	45735706	-0.20	0.72	68	INTRONIC	KPNB1	ENST00000290158
rs7223029	45701731	-0.22	0.64	69	DOWNSTREAM	NPEPPS	ENST00000322157
rs11079775	45516667	-0.22	0.63	70	INTRONIC	EFCAB13	ENST00000331493
rs7210738	45757598	-0.22	0.62	71	DOWNSTREAM	RP11-138C9.1	ENST00000578482
rs10432035	45624913	-0.22	0.62	72	INTRONIC	NPEPPS	ENST00000322157
rs7207542	45697549	-0.24	0.56	73	INTRONIC	RP11-580I16.2	ENST00000582389
rs8065437	45721653	-0.25	0.50	74	REGULATORY	NA	ENSR00000376577
rs7222851	45592169	-0.26	0.48	75	INTERGENIC	NA	NA
rs12943464	45690351	-0.26	0.47	76	INTRONIC	NPEPPS	ENST00000322157
rs4793978	45698175	-0.28	0.42	77	INTRONIC	NPEPPS	ENST00000322157
rs4264433	45737275	-0.28	0.41	78	REGULATORY	NA	ENSR00000095191
rs9901869	45575206	-0.30	0.37	79	INTERGENIC	NA	NA
rs3809868	45750596	-0.32	0.31	80	INTRONIC	KPNB1	ENST00000290158
rs11869730	45614763	-0.33	0.30	81	INTRONIC	NPEPPS	ENST00000322157
rs78551919	45626002	-0.34	0.27	82	INTRONIC	NPEPPS	ENST00000322157
rs2125873	45660887	-0.36	0.24	83	INTRONIC	NPEPPS	ENST00000322157
rs7220935	45703433	-0.36	0.23	84	DOWNSTREAM	NPEPPS	ENST00000322157
rs4794007	45590722	-0.38	0.21	85	INTERGENIC	NA	NA
rs12103836	45577469	-0.39	0.20	86	INTERGENIC	NA	NA
rs4793913	45678579	-0.40	0.18	87	INTRONIC	NPEPPS	ENST00000322157
rs9905583	45741016	-0.41	0.16	88	INTRONIC	KPNB1	ENST00000290158
rs4239163	45755954	-0.42	0.15	89	DOWNSTREAM	RP11-138C9.1	ENST00000578482
rs9905922	45740805	-0.46	0.12	90	INTRONIC	KPNB1	ENST00000290158
rs73315878	45744993	-0.50	0.09	91	INTRONIC	KPNB1	ENST00000290158
rs10163469	45687767	-0.53	0.07	92	INTRONIC	RP11-580I16.2	ENST00000582389
rs55633793	45719816	-0.54	0.06	93	INTRONIC	RP11-580I16.2	ENST00000580045
rs4793842	45671506	-0.67	0.03	94	INTRONIC	NPEPPS	ENST00000322157
rs4794372	45627760	-0.96	0.00	95	INTRONIC	NPEPPS	ENST00000322157

Supplementary Table 4.3: CADD classification and ranking of all prioritised SNPs.

8.2.4 Chapter 5 supplementary tables

	Primer efficiency (%)		
Primer	C1R-B27	U937	Average
GAPDH	91	89	90
HPRT1	105	91	98
RPL32	109	99	104
NPEPPS_P11	102	101	101.5
NPEPPS_P13	94	97	95.5
NPEPPS_P15	99	91	95
NPEPPS_P17	106	106	106

Supplementary Table 5.1: Primer efficiencies calculated for each primer for the two cell-lines (C1R-B27 and U937). Average primer efficiency for all primers in between 90% and 110% and are therefore classed as approximately equal for the assumption of the $2^{-\Delta\Delta Ct}$ method.

RT-qPCR primer	*rs11869940* genotype	All Samples			AS cases			Healthy controls		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
NPEPPS_P13	TT	14	1.40	1.59	11	1.69	1.69	3	0.33	0.33
	TA	50	1.43	2.11	40	1.97	1.85	10	-0.71	1.72
	AA	19	1.10	2.02	9	2.76	1.45	10	-0.40	1.01
NPEPPS_P17	TT	15	2.80	1.35	12	3.29	1.01	3	0.87	0.22
	TA	54	2.47	1.26	42	2.95	0.98	12	0.79	0.35
	AA	20	1.97	1.48	10	3.20	1.03	10	0.75	0.47
NPEPPS_P11	TT	15	1.95	1.18	12	2.13	1.26	3	1.24	0.45
	TA	54	1.58	1.24	42	1.87	1.19	12	0.58	0.85
	AA	20	1.64	1.44	10	2.63	0.79	10	0.66	1.26
NPEPPS_P15	TT	15	3.57	0.82	12	3.70	0.87	3	3.03	0.16
	TA	54	2.95	0.92	42	3.12	0.81	12	2.38	1.10
	AA	20	3.18	0.86	10	3.57	0.79	10	2.80	0.79

Supplementary Table 5.2: Summary eQTL analysis data of gene expression of *NPEPPS* for *rs11869940* at all four *NPEPPS* primer sets. RT-qPCR primer ascribes *NPEPPS* primer set. *rs11869940* genotype ascribes sample genotype (T = risk allele). N ascribes number of samples with genotype. Mean provides average of ΔC_t for samples. SD provides standard deviation.

RT-qPCR primer	*rs67919208* genotype	All Samples			AS cases			Healthy controls		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
TBKBP1	CC	15	8.67	2.48	12	9.42	2.00	3	5.66	1.99
	TC	50	8.41	3.61	34	10.29	2.73	16	4.41	1.02
	TT	15	7.53	3.02	9	9.81	0.97	6	4.12	0.86

Supplementary Table 5.3: Summary eQTL analysis data of gene expression of *TBKBP1* for *rs67919208*. RT-qPCR primer ascribes *TBKBP1* primer set. *rs67919208* genotype ascribes sample genotype (C = risk allele). N ascribes number of samples with genotype. Mean = average of ΔC_t for samples. SD = standard deviation.

8.2.5 Supplementary table of online tool websites

Online tool	Website
CADD	cadd.gs.washington.edu/score
HaploReg v4.1	pubs.broadinstitute.org/mammals/haploreg/haploreg.php
LDhap	ldlink.nci.nhi.gov/?tab=ldhap
NCBI Primer BLAST	www.ncbi.nim.nih.gov/tools/primer-blast/
NEBioCalculator v1.10.1	nebiocalculator.neb.com/
RegulomeDB v2.0	regulomedb.org/regulome-search/
Ensembl VEP	grch37.ensembl.org/Tools/VEP
WashU epigenome browser	epigenomegateway.wustl.edu/legacy/

Supplementary Table 8.2.5: Table of online tool websites used in project.

9 Bibliography

ADAMS, D., ALTUCCI, L., ANTONARAKIS, S. E., BALLESTEROS, J., BECK, S., BIRD, A., BOCK, C., BOEHM, B., CAMPO, E., CARICASOLE, A., DAHL, F., DERMITZAKIS, E. T., ENVER, T., ESTELLER, M., ESTIVILL, X., FERGUSON-SMITH, A., FITZGIBBON, J., FLICEK, P., GIEHL, C., GRAF, T., GROSVELD, F., GUIGO, R., GUT, I., HELIN, K., JARVIUS, J., KUPPERS, R., LEHRACH, H., LENGAUER, T., LERNMARK, A., LESLIE, D., LOEFFLER, M., MACINTYRE, E., MAI, A., MARTENS, J. H., MINUCCI, S., OUWEHAND, W. H., PELICCI, P. G., PENDEVILLE, H., PORSE, B., RAKYAN, V., REIK, W., SCHRAPPE, M., SCHUBELER, D., SEIFERT, M., SIEBERT, R., SIMMONS, D., SORANZO, N., SPICUGLIA, S., STRATTON, M., STUNNENBERG, H. G., TANAY, A., TORRENTS, D., VALENCIA, A., VELLENGA, E., VINGRON, M., WALTER, J. & WILLCOCKS, S. 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol,* 30**,** 224-6.

AL-MOSSAWI, H., YAGER, N., TAYLOR, C. A., LAU, E., DANIELLI, S., DE WIT, J., GILCHRIST, J., NASSIRI, I., MAHE, E. A., LEE, W., RIZVI, L., MAKINO, S., CHEESEMAN, J., NEVILLE, M., KNIGHT, J. C., BOWNESS, P. & FAIRFAX, B. P. 2019. Context-specific regulation of surface and soluble IL7R expression by an autoimmune risk allele. *Nat Commun,* 10**,** 4575.

ALLEN, R. L., O'CALLAGHAN, C. A., MCMICHAEL, A. J. & BOWNESS, P. 1999. Cutting Edge: HLA-B27 Can Form a Novel β2 -Microglobulin-Free Heavy Chain Homodimer Structure. *The Journal of Immunology,* 162**,** 5045-5048.

ANDERSSON, R., GEBHARD, C., MIGUEL-ESCALADA, I., HOOF, I., BORNHOLDT, J., BOYD, M., CHEN, Y., ZHAO, X., SCHMIDL, C., SUZUKI, T., NTINI, E., ARNER, E., VALEN, E., LI, K., SCHWARZFISCHER, L., GLATZ, D., RAITHEL, J., LILJE, B., RAPIN, N., BAGGER, F. O., JORGENSEN, M., ANDERSEN, P. R., BERTIN, N., RACKHAM, O., BURROUGHS, A. M., BAILLIE, J. K., ISHIZU, Y., SHIMIZU, Y., FURUHATA, E., MAEDA, S., NEGISHI, Y., MUNGALL, C. J., MEEHAN, T. F., LASSMANN, T., ITOH, M., KAWAJI, H., KONDO, N., KAWAI, J., LENNARTSSON, A., DAUB, C. O., HEUTINK, P., HUME, D. A., JENSEN, T. H., SUZUKI, H., HAYASHIZAKI, Y., MULLER, F., FORREST, A. R. R., CARNINCI, P., REHLI, M. & SANDELIN, A. 2014. An atlas of active enhancers across human cell types and tissues. *Nature,* 507**,** 455-461.

ANDRES, A. M., DENNIS, M. Y., KRETZSCHMAR, W. W., CANNONS, J. L., LEE-LIN, S., HURLE, B., PROGRAM, N. C. S., SCHWARTZBERG, P. L., WILLIAMSON, S. H., BUSTAMANTE, C. D., NIELSEN, R., CLARK, A. G. & GREEN, E. D. 2010. Balancing Selection Maintains a Form of ERAP2

that Undergoes Nonsense-Mediated Decay and Affects Antigen Presentation. *PLoS Genet,* 6.

APARICIO, O., GEISBERG, J. V., SEKINGER, E., YANG, A., MOQTADERI, Z. & STRUHL, K. 2005. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences In vivo. *Current Protocols in Molecular Biology*.

BANG, S. Y., KIM, T. H., LEE, B., KWON, E., CHOI, S. H., LEE, K. S., SHIM, S. C., POPE, A., RAHMAN, P., REVEILLE, J. D. & INMAN, R. D. 2011. Genetic studies of ankylosing spondylitis in Koreans confirm associations with ERAP1 and 2p15 reported in white patients. *J Rheumatol,* 38**,** 322-4.

BARRETT, J. C., FRY, B., MALLER, J. & DALY, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-5.

BENJAMIN, R. & PARHAM, P. 1992. HLA-B27 and disease: a consequence of inadvertent antigen presentation? *Rheum Dis Clin North Am,* 18**,** 11-21.

BERNSTEIN, B. E., STAMATOYANNOPOULOS, J. A., COSTELLO, J. F., REN, B., MILOSAVLJEVIC, A., MEISSNER, A., KELLIS, M., MARRA, M. A., BEAUDET, A. L., ECKER, J. R., FARNHAM, P. J., HIRST, M., LANDER, E. S., MIKKELSEN, T. S. & THOMSON, J. A. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol,* 28**,** 1045-8.

BOYLE, A. P., HONG, E. L., HARIHARAN, M., CHENG, Y., SCHAUB, M. A., KASOWSKI, M., KARCZEWSKI, K. J., PARK, J., HITZ, B. C., WENG, S., CHERRY, J. M. & SNYDER, M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res,* 22**,** 1790-7.

BRAUN, J., BOLLOW, M., NEURE, L., SEIPELT, E., SEYREKBASAN, F., HERBST, H., EGGENS, U., DISTLER, A. & SIEPER, J. 1995. Use of immunohistologic and in situ hybridization techniques in the examination of sacroiliac joint biopsy specimens from patients with ankylosing spondylitis. *Arthritis & Rheumatism,* 38**,** 499-505.

BRAUN, J., BOLLOW, M., REMLINGER, G., EGGENS, U., RUDWALEIT, M., DISTLER, A. & SIEPER, J. 1998. Prevalence of spondylarthropathies in HLA-B27 positive and negative blood donors. *Arthritis Rheum,* 41**,** 58-67.

BRAUN, J., BRANDT, J., LISTING, J., ZINK, A., ALTEN, R., GOLDER, W., GROMNICA-IHLE, E., KELLNER, H., KRAUSE, A., SCHNEIDER, M., SORENSEN, H., ZEIDLER, H., THRIENE, W. & SIEPER, J. 2002. Treatment of active ankylosing spondylitis with infliximab: a randomised controlled multicentre trial. *Lancet,* 359**,** 1187-93.

BREWERTON, D. A., HART, F. D., NICHOLLS, A., CAFFREY, M., JAMES, D. C. & STURROCK, R. D. 1973. Ankylosing spondylitis and HL-A 27. *Lancet,* 1**,** 904-7.

BROWN, M. A. 2018. Solving the pathogenesis of ankylosing spondylitis. *Clin Immunol,* 186**,** 46-50.

BROWN, M. A., KENNA, T. & WORDSWORTH, B. P. 2016. Genetics of ankylosing spondylitis--insights into pathogenesis. *Nat Rev Rheumatol,* 12**,** 81-91.

BROWN, M. A., KENNEDY, L. G., MACGREGOR, A. J., DARKE, C., DUNCAN, E., SHATFORD, J. L., TAYLOR, A., CALIN, A. & WORDSWORTH, P. 1997. Susceptibility to ankylosing spondylitis in twins: the role of genes, HLA, and the environment. *Arthritis Rheum,* 40**,** 1823-8.

BROWN, M. A., LAVAL, S. H., BROPHY, S. & CALIN, A. 2000. Recurrence risk modelling of the genetic susceptibility to ankylosing spondylitis. *Ann Rheum Dis,* 59**,** 883-6.

BUENROSTRO, J. D., GIRESI, P. G., ZABA, L. C., CHANG, H. Y. & GREENLEAF, W. J. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods,* 10**,** 1213-8.

BUSTIN, S. A., BENES, V., GARSON, J. A., HELLEMANS, J., HUGGETT, J., KUBISTA, M., MUELLER, R., NOLAN, T., PFAFFL, M. W., SHIPLEY, G. L., VANDESOMPELE, J. & WITTWER, C. T. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem,* 55**,** 611-22.

CAFFREY, M. F. & JAMES, D. C. 1973. Human lymphocyte antigen association in ankylosing spondylitis. *Nature,* 242**,** 121.

CHANG, C. C., CHOW, C. C., TELLIER, L. C., VATTIKUTI, S., PURCELL, S. M. & LEE, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience,* 4**,** 7.

CONG, L., RAN, F. A., COX, D., LIN, S., BARRETTO, R., HABIB, N., HSU, P. D., WU, X., JUIANG, W., MARRAFFINI, L. A. & ZHANG, F. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science,* 339**,** 819-823.

CORTES, A. & BROWN, M. A. 2011. Promise and pitfalls of the immunochip. *Arthritis Res Ther,* 13**,** 101.

CORTES, A., PULIT, S. L., LEO, P. J., POINTON, J. J., ROBINSON, P. C., WEISMAN, M. H., WARD, M., GENSLER, L. S., ZHOU, X., GARCHON, H. J., CHIOCCHIA, G., NOSSENT, J., LIE, A., FØRRE, Ø., TUOMILEHTO, J., LAIHO, K., BRADBURY, L. A., ELEWAUT, D., BURGOS-VARGAS, R., STEBBINGS, S., APPLETON, L., FARRAH, C., LAU, J., HAROON, N., MULERO, J., BLANCO, F. J., GONZALEZ-GAY, M. A., LOPEZ-LARREA, C., BOWNESS, P., GAFFNEY, K., GASTON, H., GLADMAN, D. D., RAHMAN, P., MAKSYMOWYCH, W. P., CRUSIUS, J. B., VAN DER HORST-BRUINSMA, I. E., VALLE-OÑATE, R., ROMERO-

SÁNCHEZ, C., HANSEN, I. M., PIMENTEL-SANTOS, F. M., INMAN, R. D., MARTIN, J., BREBAN, M., WORDSWORTH, B. P., REVEILLE, J. D., EVANS, D. M., DE BAKKER, P. I. & BROWN, M. A. 2015. Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1. *Nature Communications,* 6.

COSTANTINO, F., BREBAN, M. & GARCHON, H. J. 2018. Genetics and Functional Genomics of Spondyloarthritis. *Front Immunol,* 9**,** 2933.

CREYGHTON, M. P., CHENG, A. W., WELSTEAD, G. G., KOOISTRA, T., CAREY, B. W., STEINE, E. J., HANNA, J., LODATO, M. A., FRAMPTON, G. M., SHARP, P. A., BOYER, L. A., YOUNG, R. A. & JAENISCH, R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A,* 107**,** 21931-6.

DANOY, P., PRYCE, K., HADLER, J., BRADBURY, L. A., FARRAR, C., POINTON, J. J., AUSTRALO-ANGLO-AMERICAN SPONDYLOARTHRITIS CONSORTIUM, WARD, M., WEISMAN, M., REVEILLE, J. D., WORDSWORTH, B. P., STONE, M. A., SPONDYLOARTHRITIS RESEARCH CONSORTIUM OF CANADA, MAKSYMOWYCH, W. P., RAHMAN, P., GLADMAN, D., INMAN, R. D. & BROWN, M. A. 2010. Association of Variants at 1q32 and STAT3 with Ankylosing Spondylitis Suggests Genetic Overlap with Crohn's Disease. *PLoS Genet,* 6.

DAVIDSON, S. I., WU, X., LIU, Y., WEI, M., DANOY, P. A., THOMAS, G., CAI, Q., SUN, L., DUNCAN, E., WANG, N., YU, Q., XU, A., FU, Y., BROWN, M. A. & XU, H. 2009. Association of ERAP1, but not IL23R, with ankylosing spondylitis in a Han Chinese population. *Arthritis & Rheumatism,* 60**,** 3263-3268.

DE BLECOURT, J. J., POLMAN, A. & DE BLECOURT-MEINDERSMA, T. 1961. Hereditary factors in rheumatoid arthritis and ankylosing spondylitis. *Ann Rheum Dis,* 20**,** 215-20.

DIMAS, A. S., DEUTSCH, S., STRANGER, B. E., MONTGOMERY, S. B., BOREL, C., ATTAR-COHEN, H., INGLE, C., BEAZLEY, C., GUTIERREZ ARCELUS, M., SEKOWSKA, M., GAGNEBIN, M., NISBETT, J., DELOUKAS, P., DERMITZAKIS, E. T. & ANTONARAKIS, S. E. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science,* 325**,** 1246-50.

ELLINGHAUS, D., JOSTINS, L., SPAIN, S. L., CORTES, A., BETHUNE, J., HAN, B., PARK, Y. R., RAYCHAUDHURI, S., POUGET, J. G., HUBENTHAL, M., FOLSERAAS, T., WANG, Y., ESKO, T., METSPALU, A., WESTRA, H. J., FRANKE, L., PERS, T. H., WEERSMA, R. K., COLLIJ, V., D'AMATO, M., HALFVARSON, J., JENSEN, A. B., LIEB, W., DEGENHARDT, F., FORSTNER, A. J., HOFMANN, A., INTERNATIONAL I. B. D. GENETICS CONSORTIUM, INTERNATIONAL GENETICS OF

ANKYLOSING SPONDYLITIS CONSORTIUM, INTERNATIONAL P. S. C. STUDY GROUP, GENETIC ANALYSIS OF PSORIASIS CONSORTIUM, PSORIASIS ASSOCIATION GENETICS EXTENSION, SCHREIBER, S., MROWIETZ, U., JURAN, B. D., LAZARIDIS, K. N., BRUNAK, S., DALE, A. M., TREMBATH, R. C., WEIDINGER, S., WEICHENTHAL, M., ELLINGHAUS, E., ELDER, J. T., BARKER, J. N., ANDREASSEN, O. A., MCGOVERN, D. P., KARLSEN, T. H., BARRETT, J. C., PARKES, M., BROWN, M. A. & FRANKE, A. 2016. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet,* 48**,** 510-8.

ERNST, J. & KELLIS, M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods,* 9**,** 215-6.

EVANS, D. M., SPENCER, C. C., POINTON, J. J., SU, Z., HARVEY, D., KOCHAN, G., OPPERMANN, U., DILTHEY, A., PIRINEN, M., STONE, M. A., APPLETON, L., MOUTSIANAS, L., LESLIE, S., WORDSWORTH, T., KENNA, T. J., KARADERI, T., THOMAS, G. P., WARD, M. M., WEISMAN, M. H., FARRAR, C., BRADBURY, L. A., DANOY, P., INMAN, R. D., MAKSYMOWYCH, W., GLADMAN, D., RAHMAN, P., SPONDYLOARTHRITIS RESEARCH CONSORTIUM OF CANADA, MORGAN, A., MARZO-ORTEGA, H., BOWNESS, P., GAFFNEY, K., GASTON, J. S., SMITH, M., BRUGES-ARMAS, J., COUTO, A. R., SORRENTINO, R., PALADINI, F., FERREIRA, M. A., XU, H., LIU, Y., JIANG, L., LOPEZ-LARREA, C., DIAZ-PENA, R., LOPEZ-VAZQUEZ, A., ZAYATS, T., BAND, G., BELLENGUEZ, C., BLACKBURN, H., BLACKWELL, J. M., BRAMON, E., BUMPSTEAD, S. J., CASAS, J. P., CORVIN, A., CRADDOCK, N., DELOUKAS, P., DRONOV, S., DUNCANSON, A., EDKINS, S., FREEMAN, C., GILLMAN, M., GRAY, E., GWILLIAM, R., HAMMOND, N., HUNT, S. E., JANKOWSKI, J., JAYAKUMAR, A., LANGFORD, C., LIDDLE, J., MARKUS, H. S., MATHEW, C. G., MCCANN, O. T., MCCARTHY, M. I., PALMER, C. N., PELTONEN, L., PLOMIN, R., POTTER, S. C., RAUTANEN, A., RAVINDRARAJAH, R., RICKETTS, M., SAMANI, N., SAWCER, S. J., STRANGE, A., TREMBATH, R. C., VISWANATHAN, A. C., WALLER, M., WESTON, P., WHITTAKER, P., WIDAA, S., WOOD, N. W., MCVEAN, G., REVEILLE, J. D., WORDSWORTH, B. P., BROWN, M. A., DONNELLY, P., AUSTRALO-ANGLO-AMERICAN SPONDYLOARTHRITIS CONSORTIUM & WELLCOME TRUST CASE CONTROL CONSORTIUM 2011. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet,* 43**,** 761-7.

GARRETT, S., JENKINSON, T., KENNEDY, L. G., WHITELOCK, H., GAISFORD, P. & CALIN, A. 1994. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J Rheumatol,* 21**,** 2286-91.

GEIRSSON, A. J., KRISTJANSSON, K. & GUDBJORNSSON, B. 2010. A strong familiality of ankylosing spondylitis through several generations. *Ann Rheum Dis,* 69**,** 1346-8.

GIRESI, P. G., KIM, J., MCDANIELL, R. M., IYER, V. R. & LIEB, J. D. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res,* 17**,** 877-85.

GROSSCHEDL, R. 2001. Enhancers. *In:* BRENNER, S. & MILLER, J. M. (eds.) *Encyclopedia of Genetics.* Academic Press.

HANSON, A. L., CUDDIHY, T., HAYNES, K., LOO, D., MORTON, C. J., OPPERMANN, U., LEO, P., THOMAS, G. P., LE CAO, K. A., KENNA, T. J. & BROWN, M. A. 2018. Genetic Variants in ERAP1 and ERAP2 Associated With Immune-Mediated Diseases Influence Protein Expression and the Isoform Profile. *Arthritis Rheumatol,* 70**,** 255-265.

HEINTZMAN, N. D., HON, G. C., HAWKINS, R. D., KHERADPOUR, P., STARK, A., HARP, L. F., YE, Z., LEE, L. K., STUART, R. K., CHING, C. W., CHING, K. A., ANTOSIEWICZ-BOURGET, J. E., LIU, H., ZHANG, X., GREEN, R. D., LOBANENKOV, V. V., STEWART, R., THOMAS, J. A., CRAWFORD, G. E., KELLIS, M. & REN, B. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature,* 459**,** 108-12.

HIMES, S. R. & SHANNON, M. F. 2000. Assays for Transcriptional Activity Based on the Luciferase Reporter Gene. *Transcription Factor Protocols. Methods in Molecular Biology.* Humana Press.

HUANG, Q. Q., RITCHIE, S. C., BROZYNSKA, M. & INOUYE, M. 2018. Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Res,* 46**,** e133.

INTERNATIONAL GENETICS OF ANKYLOSING SPONDYLITIS CONSORTIUM, CORTES, A., HADLER, J., POINTON, J. P., ROBINSON, P. C., KARADERI, T., LEO, P., CREMIN, K., PRYCE, K., HARRIS, J., LEE, S., JOO, K. B., SHIM, S. C., WEISMAN, M., WARD, M., ZHOU, X., GARCHON, H. J., CHIOCCHIA, G., NOSSENT, J., LIE, B. A., FORRE, O., TUOMILEHTO, J., LAIHO, K., JIANG, L., LIU, Y., WU, X., BRADBURY, L. A., ELEWAUT, D., BURGOS-VARGAS, R., STEBBINGS, S., APPLETON, L., FARRAH, C., LAU, J., KENNA, T. J., HAROON, N., FERREIRA, M. A., YANG, J., MULERO, J., FERNANDEZ-SUEIRO, J. L., GONZALEZ-GAY, M. A., LOPEZ-LARREA, C., DELOUKAS, P., DONNELLY, P., AUSTRALO-ANGLO-AMERICAN SPONDYLOARTHRITIS CONSORTIUM, GROUPE FRANCAISE D'ETUDE GENETIQUE DES SPONDYLARTHRITES, NORD-TRONDELAG HEALTH STUDY, SPONDYLOARTHRITIS RESEARCH CONSORTIUM OF CANADA, WELLCOME TRUST CASE CONTROL CONSORTIUM, BOWNESS, P., GAFNEY, K., GASTON, H., GLADMAN, D. D., RAHMAN, P.,

MAKSYMOWYCH, W. P., XU, H., CRUSIUS, J. B., VAN DER HORST-BRUINSMA, I. E., CHOU, C. T., VALLE-ONATE, R., ROMERO-SANCHEZ, C., HANSEN, I. M., PIMENTEL-SANTOS, F. M., INMAN, R. D., VIDEM, V., MARTIN, J., BREBAN, M., REVEILLE, J. D., EVANS, D. M., KIM, T. H., WORDSWORTH, B. P. & BROWN, M. A. 2013. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat Genet,* 45**,** 730-8.

INTERNATIONAL MULTIPLE SCLEROSIS GENETICS CONSORTIUM, 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., WELLCOME TRUST CASE CONTROL, C., INTERNATIONAL, I. B. D. G. C., 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 SAPIO, 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., et al. 2013. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet,* 45**,** 1353-60.

KENNA, T. & BROWN, M. A. 2013. Immunopathogenesis of ankylosing spondylitis. *International Journal of Clinical Rheumatology,* 8**,** 265-274.

KENNA, T. J., HANSON, A., COSTELLO, M. E. & BROWN, M. A. 2016. Functional Genomics and Its Bench-to-Bedside Translation Pertaining to the Identified Susceptibility Alleles and Loci in Ankylosing Spondylitis. *Curr Rheumatol Rep,* 18**,** 63.

KLEMM, S. L., SHIPONY, Z. & GREENLEAF, W. J. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet,* 20**,** 207-220.

LAU, M. C., KEITH, P., COSTELLO, M. E., BRADBURY, L. A., HOLLIS, K. A., THOMAS, R., THOMAS, G. P., BROWN, M. A. & KENNA, T. J. 2017. Genetic association of ankylosing spondylitis with TBX21 influences T-bet

and pro-inflammatory cytokine expression in humans and SKG mice as a model of spondyloarthritis. *Ann Rheum Dis,* 76**,** 261-269.

LEE, W., REVEILLE, J. D., DAVIS, J. C., JR., LEARCH, T. J., WARD, M. M. & WEISMAN, M. H. 2007. Are there gender differences in severity of ankylosing spondylitis? Results from the PSOAS cohort. *Ann Rheum Dis,* 66**,** 633-8.

LI, Y. R. & KEATING, B. J. 2014. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med,* 6**,** 91.

LI, Z. & BROWN, M. A. 2017. Progress of genome-wide association studies of ankylosing spondylitis. *Clin Transl Immunology,* 6**,** e163.

LI, Z., HAYNES, K., PENNISI, D. J., ANDERSON, L. K., SONG, X., THOMAS, G. P., KENNA, T., LEO, P. & BROWN, M. A. 2017. Epigenetic and gene expression analysis of ankylosing spondylitis-associated loci implicate immune cells and the gut in the disease pathogenesis. *Genes Immun,* 18**,** 135-143.

LIVAK, K. J. & SCHMITTGEN, T. D. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2 -ddCt Method. *Methods,* 25**,** 402-408.

MACHIELA, M. J. & CHANOCK, S. J. 2015. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics,* 31**,** 3555-7.

MAHMOUDI, M., JAMSHIDI, A. R., AMIRZARGAR, A. A., FARHADI, E., NOURIJELYANI, K., FALLAHI, S., ORAEI, M., NOORI, S. & NICKNAM, M. H. 2012. Association between endoplasmic reticulum aminopeptidase-1 (ERAP-1) and susceptibility to ankylosing spondylitis in Iran. *Iran Journal Aleergy Asthma Immunology,* 11**,** 294-300.

MALI, P., YANG, L., ESVELT, K. M., AACH, J., GUELL, M., DICARLO, J. E., NORVILLE, J. E. & CHURCH, G. M. 2013. RNA-guided human genome engineering via Cas9. *Science,* 339**,** 823-6.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. & DONNELLY, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet,* 39**,** 906-13.

MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet,* 7**,** 29-59.

MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H., BRODY, J., SHAFER, A., NERI, F., LEE, K., KUTYAVIN, T., STEHLING-SUN, S., JOHNSON, A. K., CANFIELD, T. K., GISTE, E., DIEGEL, M., BATES, D.,

HANSEN, R. S., NEPH, S., SABO, P. J., HEIMFELD, S., RAUBITSCHEK, A., ZIEGLER, S., COTSAPAS, C., SOTOODEHNIA, N., GLASS, I., SUNYAEV, S. R., KAUL, R. & STAMATOYANNOPOULOS, J. A. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science,* 337**,** 1190-5.

MCCARTHY, S., DAS, S., KRETZSCHMAR, W., DELANEAU, O., WOOD, A. R., TEUMER, A., KANG, H. M., FUCHSBERGER, C., DANECEK, P., SHARP, K., LUO, Y., SIDORE, C., KWONG, A., TIMPSON, N., KOSKINEN, S., VRIEZE, S., SCOTT, L. J., ZHANG, H., MAHAJAN, A., VELDINK, J., PETERS, U., PATO, C., VAN DUIJN, C. M., GILLIES, C. E., GANDIN, I., MEZZAVILLA, M., GILLY, A., COCCA, M., TRAGLIA, M., ANGIUS, A., BARRETT, J. C., BOOMSMA, D., BRANHAM, K., BREEN, G., BRUMMETT, C. M., BUSONERO, F., CAMPBELL, H., CHAN, A., CHEN, S., CHEW, E., COLLINS, F. S., CORBIN, L. J., SMITH, G. D., DEDOUSSIS, G., DORR, M., FARMAKI, A. E., FERRUCCI, L., FORER, L., FRASER, R. M., GABRIEL, S., LEVY, S., GROOP, L., HARRISON, T., HATTERSLEY, A., HOLMEN, O. L., HVEEM, K., KRETZLER, M., LEE, J. C., MCGUE, M., MEITINGER, T., MELZER, D., MIN, J. L., MOHLKE, K. L., VINCENT, J. B., NAUCK, M., NICKERSON, D., PALOTIE, A., PATO, M., PIRASTU, N., MCINNIS, M., RICHARDS, J. B., SALA, C., SALOMAA, V., SCHLESSINGER, D., SCHOENHERR, S., SLAGBOOM, P. E., SMALL, K., SPECTOR, T., STAMBOLIAN, D., TUKE, M., TUOMILEHTO, J., VAN DEN BERG, L. H., VAN RHEENEN, W., VOLKER, U., WIJMENGA, C., TONIOLO, D., ZEGGINI, E., GASPARINI, P., SAMPSON, M. G., WILSON, J. F., FRAYLING, T., DE BAKKER, P. I., SWERTZ, M. A., MCCARROLL, S., KOOPERBERG, C., DEKKER, A., ALTSHULER, D., WILLER, C., IACONO, W., RIPATTI, S., et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet,* 48**,** 1279-83.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol,* 17**,** 122.

MEAR, J. P., SCHREIBER, K. L., MUNZ, C., ZHU, X., STEVANOVIC, S., RAMMENSEE, H., ROWLAND-JONES, S. L. & COLBERT, R. A. 1999. Misfolding of HLA-B27 as a Result of Its B Pocket Suggests a Novel Mechanism for Its Role in Susceptibility to Spondyloarthropathies. *The Journal of Immunology,* 163**,** 6665-6670.

MIELANTS, H., VEYS, E. M., GOEMAERE, S., GOETHALS, K., CUVELIER, C. & DE VOS, M. 1991. Gut inflammation in the spondyloarthropathies: clinical, radiologic, biologic and genetic features in relation to the type of histology. A prospective study. *J Rheumatol,* 18**,** 1542-51.

NGUYEN, T. T., CHANG, S. C., EVNOUCHIDOU, I., YORK, I. A., ZIKOS, C., ROCK, K. L., GOLDBERG, A. L., STRATIKOS, E. & STERN, L. J. 2011. Structural basis for antigenic peptide precursor processing by the

endoplasmic reticulum aminopeptidase ERAP1. *Nat Struct Mol Biol,* 18**,** 604-13.

ONG, C. T. & CORCES, V. G. 2014. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet,* 15**,** 234-46.

PENNACCHIO, L. A., BICKMORE, W., DEAN, A., NOBREGA, M. A. & BEJERANO, G. 2013. Enhancers: five essential questions. *Nature Reviews Genetics,* 14**,** 288-295.

POINTON, J. J., HARVEY, D., KARADERI, T., APPLETON, L. H., FARRAR, C., STONE, M. A., STURROCK, R. D., REVEILLE, J. D., WEISMAN, M. H., WARD, M. M., BROWN, M. A. & WORDSWORTH, B. P. 2010. The chromosome 16q region associated with ankylosing spondylitis includes the candidate gene tumour necrosis factor receptor type 1-associated death domain (TRADD). *Ann Rheum Dis,* 69**,** 1243-6.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

RENTZSCH, P., WITTEN, D., COOPER, G. M., SHENDURE, J. & KIRCHER, M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res,* 47**,** D886-D894.

ROBERTS, A. R., VECELLIO, M., CHEN, L., RIDLEY, A., CORTES, A., KNIGHT, J. C., BOWNESS, P., COHEN, C. J. & WORDSWORTH, B. P. 2015. An ankylosing spondylitis-associated genetic variant in the IL23R-IL12RB2 intergenic region modulates enhancer activity and is associated with increased Th1-cell differentiation. *Ann Rheum Dis,* 75**,** 2150-2156.

ROBINSON, J., BARKER, D. J., GEORGIOU, X., COOPER, M. A., FLICEK, P. & MARSH, S. G. E. 2020. IPD-IMGT/HLA Database. *Nucleic Acids Res,* 48**,** D948-D955.

ROCKMAN, M. V. & KRUGLYAK, L. 2006. Genetics of global gene expression. *Nat Rev Genet,* 7**,** 862-72.

SALOMON, M. P., LI, W. L., EDLUND, C. K., MORRISON, J., FORTINI, B. K., WIN, A. K., CONTI, D. V., THOMAS, D. C., DUGGAN, D., BUCHANAN, D. D., JENKINS, M. A., HOPPER, J. L., GALLINGER, S., LE MARCHAND, L., NEWCOMB, P. A., CASEY, G. & MARJORAM, P. 2016. GWASeq: targeted re-sequencing follow up to GWAS. *BMC Genomics,* 17**,** 176.

SCHAID, D. J., CHEN, W. & B., L. N. 2018. From genome-wide association to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics,* 19**,** 491-504.

SCHLOSSTEIN, L., TERASAKI, P. I., BLUESTONE, R. & PEARSON, C. M. 1973. High association of an HL-A antigen, W27, with ankylosing spondylitis. *N Engl J Med,* 288**,** 704-6.

SHEN, H., GOODALL, J. C. & HILL GASTON, J. S. 2009. Frequency and phenotype of peripheral blood Th17 cells in ankylosing spondylitis and rheumatoid arthritis. *Arthritis Rheum,* 60**,** 1647-56.

SIEPER, J., APPEL, H., BRAUN, J. & RUDWALEIT, M. 2008. Critical appraisal of assessment of structural damage in ankylosing spondylitis: implications for treatment outcomes. *Arthritis Rheum,* 58**,** 649-56.

SIEPER, J., RUDWALEIT, M., BARALIAKOS, X., BRANDT, J., BRAUN, J., BURGOS-VARGAS, R., DOUGADOS, M., HERMANN, K.-G., LANDEWE, R., MAKSYMOWYCH, W. & VAN DER HEIJDE, D. 2009. The Assessment of SpondyloArthritis international Society (ASAS) handbook: a guide to assess spondyloarthritis. *Ann Rheum Dis,* 68.

SPAIN, S. L. & BARRETT, J. C. 2015. Strategies for fine-mapping complex traits. *Hum Mol Genet,* 24**,** R111-9.

SPIELMANN, M. & MUNDLOS, S. 2016. Looking beyond the genes: the role of non-coding variants in human disease. *Hum Mol Genet,* 25**,** R157-R165.

STEPHENS, M. & BALDING, D. J. 2009. Bayesian statistical methods for genetic association studies. *Nat Rev Genet,* 10**,** 681-90.

STOLTZE, L., SCHIRLE, M., SCHWARZ, G., SCHROTER, C., THOMPSON, M. W., HERSH, L. B., KALBACHER, H., STEVANOVIC, S., RAMMENSEE, H. G. & SCHILD, H. 2000. Two new proteases in the MHC class I processing pathway. *Nat Immunol,* 1**,** 413-8.

STOLWIJK, C., VAN ONNA, M., BOONEN, A. & VAN TUBERGEN, A. 2016. Global Prevalence of Spondyloarthritis: A Systematic Review and Meta-Regression Analysis. *Arthritis Care Res (Hoboken),* 68**,** 1320-31.

SYKES, M. P., DOLL, H., SENGUPTA, R. & GAFFNEY, K. 2015. Delay to diagnosis in axial spondyloarthritis: are we improving in the UK? *Rheumatology (Oxford),* 54**,** 2283-4.

SZALAY, B., MESZAROS, G., CSEH, A., ACS, L., DEAK, M., KOVACS, L., VASARHELYI, B. & BALOG, A. 2012. Adaptive Immunity in Ankylosing Spondylitis: Phenotype and Functional Alterations of T-Cells before and during Infliximab Therapy. *Clinical and Developmental Immunology*.

TAK, Y. G. & FARNHAM, P. J. 2015. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin,* 8.

TEAM, R. C. 2014. R: A language and environment for statistical computing.

THE AUSTRALO-ANGLO-AMERICAN SPONDYLOARTHRITIS CONSORTIUM 2010. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nature Genetics,* 42**,** 123.

THE ENCODE PROJECT CONSORTIUM 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

THJODLEIFSSON, B., GEIRSSON, A. J., BJORNSSON, S. & BJARNASON, I. 2007. A common genetic background for inflammatory bowel disease and ankylosing spondylitis: a genealogic study in Iceland. *Arthritis Rheum,* 56**,** 2633-9.

THOMAS, G. P. & BROWN, M. A. 2010. Genetics and genomics of ankylosing spondylitis. *Immunol Rev,* 233**,** 162-80.

THURMAN, R. E., RYNES, E., HUMBERT, R., VIERSTRA, J., MAURANO, M. T., HAUGEN, E., SHEFFIELD, N. C., STERGACHIS, A. B., WANG, H., VERNOT, B., GARG, K., JOHN, S., SANDSTROM, R., BATES, D., BOATMAN, L., CANFIELD, T. K., DIEGEL, M., DUNN, D., EBERSOL, A. K., FRUM, T., GISTE, E., JOHNSON, A. K., JOHNSON, E. M., KUTYAVIN, T., LAJOIE, B., LEE, B. K., LEE, K., LONDON, D., LOTAKIS, D., NEPH, S., NERI, F., NGUYEN, E. D., QU, H., REYNOLDS, A. P., ROACH, V., SAFI, A., SANCHEZ, M. E., SANYAL, A., SHAFER, A., SIMON, J. M., SONG, L., VONG, S., WEAVER, M., YAN, Y., ZHANG, Z., ZHANG, Z., LENHARD, B., TEWARI, M., DORSCHNER, M. O., HANSEN, R. S., NAVAS, P. A., STAMATOYANNOPOULOS, G., IYER, V. R., LIEB, J. D., SUNYAEV, S. R., AKEY, J. M., SABO, P. J., KAUL, R., FUREY, T. S., DEKKER, J., CRAWFORD, G. E. & STAMATOYANNOPOULOS, J. A. 2012. The accessible chromatin landscape of the human genome. *Nature,* 489**,** 75-82.

TRYNKA, G. & RAYCHAUDHURI, S. 2013. Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. *Curr Opin Genet Dev,* 23**,** 635-41.

TSUI, F. W., HAROON, N., REVEILLE, J. D., RAHMAN, P., CHIU, B., TSUI, H. W. & INMAN, R. D. 2010. Association of an ERAP1 ERAP2 haplotype with familial ankylosing spondylitis. *Ann Rheum Dis,* 69**,** 733-6.

VAN DE BUNT, M., CORTES, A., IGAS CONSORTIUM, BROWN, M. A., MORRIS, A. P. & MCCARTHY, M. I. 2015. Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. *PLoS Genet,* 11**,** e1005535.

VAN DE LINDEN, S., VALKENBURG, H. A. & CATS, A. 1984. Evaluation of diagnostic criteria for ankylosing spondylitis. A Proposal for Modification of the New York Criteria. *Arthritis and Rheumatism,* 27**,** 361-368.

VECELLIO, M., CORTES, A., ROBERTS, A. R., ELLIS, J., COHEN, C. J., KNIGHT, J. C., BROWN, M. A., BOWNESS, P. & WORDSWORTH, B. P. 2018. Evidence for a second ankylosing spondylitis-associated RUNX3 regulatory polymorphism. *RMD Open,* 4**,** e000628.

VECELLIO, M., ROBERTS, A. R., COHEN, C. J., CORTES, A., KNIGHT, J. C., BOWNESS, P. & WORDSWORTH, B. P. 2016. The genetic association of RUNX3 with ankylosing spondylitis can be explained by allele-specific effects on IRF4 recruitment that alter gene expression. *Ann Rheum Dis,* 75**,** 1534-1540.

VISEL, A., RUBIN, E. M. & PENNACCHIO, L. A. 2009. Genomic view of distant-acting enhancers. *Nature,* 461**,** 199-205.

WARD, L. D. & KELLIS, M. 2012. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res,* 40**,** D930-4.

WELLCOME TRUST CASE CONTROL CONSORTIUM, MALLER, J. B., MCVEAN, G., BYRNES, J., VUKCEVIC, D., PALIN, K., SU, Z., HOWSON, J. M., AUTON, A., MYERS, S., MORRIS, A., PIRINEN, M., BROWN, M. A., BURTON, P. R., CAULFIELD, M. J., COMPSTON, A., FARRALL, M., HALL, A. S., HATTERSLEY, A. T., HILL, A. V., MATHEW, C. G., PEMBREY, M., SATSANGI, J., STRATTON, M. R., WORTHINGTON, J., CRADDOCK, N., HURLES, M., OUWEHAND, W., PARKES, M., RAHMAN, N., DUNCANSON, A., TODD, J. A., KWIATKOWSKI, D. P., SAMANI, N. J., GOUGH, S. C., MCCARTHY, M. I., DELOUKAS, P. & DONNELLY, P. 2012. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet,* 44**,** 1294-301.

WELLCOME TRUST CASE CONTROL CONSORTIUM & THE AUSTRALO-ANGLO-AMERICAN SPONDYLITIS CONSORTIUM 2007. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genetics,* 39**,** 1329.

WRIGHT, C., EDELMANN, M., DIGLERIA, K., KOLLNBERGER, S., KRAMER, H., MCGOWAN, S., MCHUGH, K., TAYLOR, S., KESSLER, B. & BOWNESS, P. 2009. Ankylosing spondylitis monocytes show upregulation of proteins involved in inflammation and the ubiquitin proteasome pathway. *Ann Rheum Dis,* 68**,** 1626-1632.

YE, J., COULOURIS, G., ZARETSKAYA, I., CUTCUTACHE, I., ROZEN, S. & MADDEN, T. L. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics,* 13**,** 134.

YIN, W., XIANG, P. & LI, Q. 2005. Investigations of the effect of DNA size in transient transfection assay using dual luciferase system. *Anal Biochem,* 346**,** 289-94.

ZHOU, V. W., GOREN, A. & BERNSTEIN, B. E. 2011a. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet,* 12**,** 7-18.

ZHOU, X., MARICQUE, B., XIE, M., LI, D., SUNDARAM, V., MARTIN, E. A., KOEBBE, B. C., NIELSEN, C., HIRST, M., FARNHAM, P., KUHN, R. M.,

ZHU, J., SMIRNOV, I., KENT, W. J., HAUSSLER, D., MADDEN, P. A., COSTELLO, J. F. & WANG, T. 2011b. The Human Epigenome Browser at Washington University. *Nature Methods,* 8**,** 989-90.

ZHU, L., LI, Y., XIE, X., ZHOU, X., GU, M., JIE, Z., KO, C. J., GAO, T., HERNANDEZ, B. E., CHENG, X. & SUN, S. C. 2019. TBKBP1 and TBK1 form a growth factor signalling axis mediating immunosuppression and tumourigenesis. *Nat Cell Biol,* 21**,** 1604-1614.

ZINOVIEVA, E., BOURGAIN, C., KADI, A., LETOURNEUR, F., IZAC, B., SAID-NAHAL, R., LEBRUN, N., CAGNARD, N., VIGIER, A., JACQUES, S., MICELI-RICHARD, C., GARCHON, H., S., H., CHARON, C., BACQ, D., BOLAND, A., ZELENIKA, D., CHIOCCHIA, G. & BREBAN, M. 2009. Comprehensive linkage and association analyses identify haplotype, near to the TNFSF15 gene, significantly associated with spondyloarthritis. *PLoS Genet,* 5.