

Genetics of Ankylosing Spondylitis



Tugce Karaderi
University College

D.Phil Thesis
Trinity Term 2012

ABSTRACT

Ankylosing spondylitis (AS) is a common inflammatory arthritis of the spine and other affected joints, which is highly heritable, being strongly influenced by the HLA-B27 status, as well as hundreds of mostly unknown genetic variants of smaller effect. The aim of my research was to confirm some of the previously observed genetic associations and to identify new associations, many of which are in biological pathways relevant to AS pathogenesis, most notably the IL-23/T_H17 axis (*IL23R*) and antigen presentation (*ERAP1* and *ERAP2*). Studies presented in this thesis include replication and refinement of several potential associations initially identified by earlier GWAS (WTCCC-TASC, 2007 and TASC, 2010). I conducted an extended study of *IL23R* association with AS and undertook a meta-analysis, confirming the association between AS and *IL23R* (non-synonymous SNP rs11209026, $p=1.5 \times 10^{-9}$, OR=0.61). An extensive re-sequencing and fine mapping project, including a meta-analysis, to replicate and refine the association of *TNFRSF1A* with AS was also undertaken; a novel variant in intron 6 was identified and a weak association with a low frequency variant, rs4149584 ($p=0.01$, OR=1.58), was detected. Somewhat stronger associations were seen with rs4149577 ($p=0.002$, OR=0.91) and rs4149578 ($p=0.015$, OR=1.14) in the meta-analysis. Associations at several additional loci had been identified by a more recent GWAS (WTCCC2-TASC, 2011). I used *in silico* techniques, including imputation using a denser panel of variants from the 1000 Genomes Project, conditional analysis and rare/low frequency variant analysis, to refine these associations. Imputation analysis (1782 cases/5167 controls) revealed novel associations with *ERAP2* (rs4869313, $p=7.3 \times 10^{-8}$, OR=0.79) and several additional candidate loci including *IL6R*, *UBE2L3* and 2p16.3. Ten SNPs were then directly typed in an independent sample (1804 cases/1848 controls) to replicate selected associations and to determine the imputation accuracy. I established that imputation using the 1000 Genomes Project pilot data was largely reliable, specifically for common variants (genotype concordance~97%). However, more accurate imputation of low frequency variants may require larger reference populations, like the most recent 1000 Genomes reference panels. The results of my research provide a better understanding of the complex genetics of AS, and help identify future targets for genetic and functional studies.

ATTRIBUTIONS

The work undertaken in this thesis is my own original work. Parts of the thesis contain work done in collaboration with others. The contribution of others is listed below and in some cases within the text.

I would like to acknowledge the contribution of the following people:

Patient Recruitment and Examination - Professor B. Paul Wordsworth, Professor Matthew Brown, Claire Farrar, Linda Bradbury, Jonathan Lau (Institute of Musculoskeletal Sciences, The Botnar Research Centre, Oxford, UK).

National Ankylosing Spondylitis Society (Unit 0.2, One Victoria Villas, Richmond, Surrey, UK) for their invaluable help in the recruitment of patients.

DNA preparation - Louise H. Appleton, Tom Wordsworth, David Harvey and the other members of the Wordsworth Group.

Genotyping Outsourced to Wellcome Trust & KBiosciences - Lyn-Louise Johnson and Chris Allan, (Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, UK).

KBiosciences (Hoddesdon, Herts, UK).

Bioinformatics Support - Dr. John Broxholme and Dr. Tim Bardsley (Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, UK).

ACKNOWLEDGEMENTS

I would like to thank the patients and their families who participated in this research. Particular thanks go to Claire Farrar for the recruitment of patients and the collection of all the associated clinical information. I would also like to thank the Osteoarthritis Group in Oxford, Millicent Stone, Bath and Roger Sturrock, Glasgow for the use of their samples. I am indebted to the University College, University of Oxford for the Henni Mester Scholarship, the National Ankylosing Spondylitis Society (NASS) and the Arthritis Research Campaign, UK that funded the research carried out in this thesis.

I cannot thank enough to my supervisors Professor B. Paul Wordsworth, Dr. Jennifer J. Pointon and Dr. David M. Evans for their invaluable advice and assistance. I especially would like to thank Professor Wordsworth and Dr. Jenny Pointon who showed great belief in me and were always supportive. This work would not have been possible without them. I am particularly thankful to Dr. Jenny Pointon who was extremely supportive and helpful during my lab work. I also would like to thank Dr. David Evans for his assistance with the *in silico* techniques including the WTCCC2 imputation and statistical analysis. I am grateful to Carla Cohen for the time she spent on reading and commenting on this thesis. The last but not least, I am thankful to the remaining members of the Wordsworth group; Louise Appleton, David Harvey, Peter Cain, Yun Zhang, Pille Harrison and Perdy Dobson.

I am grateful to my family, particularly my parents, for their endless support. I would also like to thank my friends, who were always there for me.

ABBREVIATIONS

1958 BBC – 1958 British Birth Cohort

APC – Antigen-Presenting Cell

AS – Ankylosing Spondylitis

ASAS - Assessment in SpondyloArthritis International Society

B27₂ – HLA-B27 homodimer

BASDAI – Bath Ankylosing Spondylitis Disease Activity Index

BASFI – Bath Ankylosing Spondylitis Functional Index

BASMI – Bath Ankylosing Spondylitis Metrology Index

CARD9 – Caspase Recruitment Domain-containing Protein 9

CCR6 – Chemokine (C-C Motif) Receptor 6

CCL20 – Chemokine (C-C Motif) Ligand 20

CD – Crohn's Disease

CI – Confidence interval

CCRaVAT – Case-Control Rare Variant Analysis Tool

DC – Dendritic Cell

ddNTPs - Dideoxynucleotide Triphosphate

dNTP – Deoxynucleotide Triphosphate

DZ – Dizygotic

EDTA - Ethylenediaminetetraacetic Acid

eQTL – Expression Quantitative Trait Locus

ER – Endoplasmic Reticulum

ERAD - Endoplasmic Reticulum-Associated Protein Degradation

ERAP1 - Endoplasmic Reticulum Aminopeptidase 1

ERAP2 – Endoplasmic Reticulum Aminopeptidase 2

GRANVIL – Gene- or Region-based Analysis of Variants of Intermediate and Low Frequency

GWAS – Genomewide Association Study(s)

HHAT – Hedgehog Acyltransferase

HLA- Human Leukocyte Antigen

I² – Inconsistency (%)

IBD – Inflammatory Bowel Disease

IFN γ – Interferon gamma

IKK - The Inhibitor of NF- κ B (I κ B)-Kinase

IL1 – Interleukin 1

IL12B – Interleukin 12B

IL17 – Interleukin 17

IL21 – Interleukin 21

IL22 – Interleukin 22

IL23 – Interleukin 23

IL23p19 – p19 subunit of interleukin 23

IL6R – Interleukin 6 Receptor

IL23R – Interleukin 23 Receptor

IL6 – Interleukin 6

IRAK4 - IL-1R-Associated Kinase 4

IRF4 – Interferon Regulatory Factor 4

JAK2 – Janus Kinase 2

KIR – Killer Immunoglobulin-Like Receptors

LD – Linkage Disequilibrium

MACH – Markov Chain-based Haplotyper

MAF – Minor Allele Frequency

MALDI-TOF – Matrix-Associated Laser Desorption Ionization Time of Flight

Mass Spectrometry

MgCl₂ – Magnesium Chloride

MHC – Major Histocompatibility Complex

MRI – Magnetic Resonance Imaging

microRNA – Micro Ribonucleic Acid

mRNA – Messenger Ribonucleic Acid

MS – Multiple Sclerosis

MZ - Monozygotic

NCBI – National Centre for Biotechnology Information

NF- κ B – Nuclear Factor κ B

NK – Natural Killer

NOD2 - Nucleotide-Binding Oligomerization Domain containing protein 2

NOD3 - Nucleotide-Binding Oligomerization Domain containing protein 3

NSAID – Non-Steroidal Anti-Inflammatory Drug

NS SNP - Non-Synonymous Single Nucleotide Polymorphism

OR – Odds Ratio

PBMC – Peripheral Blood Mononuclear Cells

PCR – Polymerase Chain Reaction

PLC – Peptide Loading Complex

PolyPhen – Polymorphism Phenotyping

PRR – Pattern Recognition Receptor

RA – Rheumatoid Arthritis

rcf – Relative Centrifugal Force

ReA – Reactive Arthritis

RFLP – Restriction Fragment Length Polymorphism

RORC - RAR-related orphan receptor C

RORyt - RAR-related orphan receptor gamma t

RPM – Revolutions Per Minute

RUNX1 – Runt-related Transcription Factor 1

RUNX3 – Runt-related Transcription Factor 3

SIFT – Sorting Intolerant From Tolerant

SNP – Single Nucleotide Polymorphism

SpA – Spondyloarthropathy

STAT3 - Signal Transducer and Activator of Transcription 3

T2D – Type 2 Diabetes

TAP – Transporter Associated with Antigen Processing

TASC – The Australo-Anglo-American Spondyloarthritis Consortium

TBE - Tris-Borate-EDTA

TCR – T Cell Receptor

TE - Tris-EDTA

TGF- β – Transforming Growth Factor Beta

TLR4 – Toll-Like Receptor 4

TNF – Tumour Necrosis Factor

TNFR1 - Tumour Necrosis Factor Receptor 1

TNFRSF1A – Tumour Necrosis Factor Receptor Superfamily member 1A

TNF α – Tumour Necrosis Factor Alpha

TRADD - TNFR1-Associated Death Domain

TYK2 – Tyrosine Kinase 2

UBE2L3 – Ubiquitin-conjugating Enzyme E2L 3

UC – Ulcerative Colitis

UPR – Unfolded Protein Response

UTR – Untranslated Region

WTCCC – The Wellcome Trust Case-Control Consortium

β_2m – Beta-2 Microglobulin

χ^2 – Chi-Square

TABLE OF CONTENTS

ABSTRACT	2
ATTRIBUTIONS.....	3
ACKNOWLEDGEMENTS.....	4
ABBREVIATIONS	5
CHAPTER 1.....	18
1. INTRODUCTION TO ANKYLOSING SPONDYLITIS	18
1.1. INTRODUCTION.....	18
1.1.1. <i>Clinical Features & Pathology of AS</i>	20
1.1.2. <i>Definition of the Phenotype: Diagnostic Criteria</i>	24
1.1.2.1. The Modified New York Criteria.....	24
1.1.3. <i>Management & Treatment</i>	25
1.1.3.1. Drugs.....	26
1.1.3.1.1. Non-Steroidal Anti-Inflammatory Drugs (NSAIDs)	26
1.1.3.1.2. Anti-TNF Treatment.....	26
1.1.3.1.3. Potential New Treatments.....	27
1.1.3.2. Physical Therapy	27
1.2. EPIDEMIOLOGY	28
1.2.1. <i>Prevalence of AS</i>	28
1.2.2. <i>Male Predominance & Age at Onset</i>	28
1.3. IMMUNOLOGY OF ANKYLOSING SPONDYLITIS	29
1.3.1. <i>Components of the Immune System</i>	31
1.3.1.1. Innate Immunity	31
1.3.1.1.1. Pattern Recognition Receptors.....	31
1.3.1.1.2. Killer Immunoglobulin-like Receptors.....	32
1.3.1.1.3. Leukocytes: Cell-mediated Innate Immune Response	33
1.3.1.2. Adaptive Immunity.....	34
1.3.1.2.1. T Lymphocytes & Cell-mediated Immunity	35
1.3.1.3. Major Histocompatibility Complex.....	39
1.3.1.3.1. The Evolution of Major Histocompatibility Complex.....	40
1.3.1.4. Antigen Processing & Presentation	41

1.3.1.5. Innate Immune Recognition & Control of Adaptive Immune Response.....	45
1.3.2. Immune/Inflammatory Responses.....	47
1.3.2.1. HLA-B27 & Pathogenesis of AS	50
1.3.2.1.1. Unfolded Protein Response.....	50
1.3.2.1.2. Other Proposed Mechanisms.....	50
1.4. GENETIC EPIDEMIOLOGY OF ANKYLOSING SPONDYLITIS.....	51
1.4.1. Genetic Predisposition & Heritability	52
1.4.2. Major Histocompatibility Complex: Human Leukocyte Antigen.....	53
1.4.2.1. HLA-B27 in Ankylosing Spondylitis	53
1.4.2.2. Prevalence of AS & HLA-B27 Subtypes.....	53
1.4.3. Genetic Linkage Studies.....	56
1.4.3.1. Genetic Linkage Studies in AS.....	57
1.5. CANDIDATE GENE ASSOCIATION STUDIES	58
1.5.1. The Basics of Genetic Association Studies.....	58
1.5.1.1. Single Nucleotide Polymorphisms.....	58
1.5.1.2. Linkage Disequilibrium	59
1.5.2. Candidate Gene Association Studies of Complex Diseases.....	62
1.5.3. MHC Region: HLA-B Associations.....	64
1.5.3.1. HLA-B27 & AS.....	64
1.5.3.2. HLA-B60 & AS	64
1.5.4. Other Genes with a Role in Related Inflammatory Diseases that are Candidates for AS.....	65
1.6. THE ERA OF GENOMEWIDE ASSOCIATION STUDIES.....	66
1.6.1. Genomewide Association Studies.....	67
1.6.2. The International HapMap Project.....	68
1.6.3. Limitations & Issues Related to GWAS	69
1.6.4. Main Findings of GWAS in AS.....	71
1.6.4.1. WTCCC1 2007	71
1.6.4.2. WTCCC-TASC 2007.....	72
1.6.4.3. TASC 2010	73
1.6.4.4. WTCCC2 & TASC Meta-analysis 2011	73
1.6.5. Associated Genes Common to AS & Other Related Inflammatory Diseases.....	74

1.7. THE USE OF IMPUTATION & LARGE-SCALE SEQUENCING	76
1.7.1. <i>A Brief Explanation of Imputation: In silico Genotyping</i>	76
1.7.1.1. Imputation in the Sequencing Era	78
1.7.1.1.1. The 1000 Genomes Project	78
1.8. THE AIMS OF THE THESIS	79
1.9. REFERENCES	80
CHAPTER 2.....	90
2. GENERAL METHODS.....	90
2.1. STUDY POPULATIONS	90
2.2. DNA EXTRACTION FROM BLOOD & SALIVA	91
2.2.1. <i>DNA Extraction from Blood</i>	91
2.2.2. <i>DNA Extraction from Saliva</i>	92
2.3. POLYMERASE CHAIN REACTION (PCR) & DNA ELECTROPHORESIS	92
2.3.1. <i>Primer Design, Preparation & PCR</i>	92
2.3.2. <i>Agarose Gel Electrophoresis</i>	95
2.4. SNP GENOTYPING.....	98
2.4.1. <i>Restriction Fragment Length Polymorphism</i>	98
2.4.2. <i>Tag SNP Selection & High Throughput Genotyping</i>	99
2.4.2.1. KASPar Genotyping	100
2.4.2.2. iPLEX Genotyping.....	102
2.4.2.3. Quality Control	105
2.5. DNA SEQUENCING & DETECTION OF NOVEL SNPs.....	107
2.5.1. <i>PCR Amplification</i>	108
2.5.2. <i>Purification of the PCR Products</i>	108
2.5.3. <i>Sequencing PCR</i>	108
2.5.4. <i>Purification of the Sequencing PCR Products</i>	110
2.5.5. <i>Analysis of DNA Sequencing Reaction Products</i>	110
2.5.6. <i>Analysis of DNA Sequences with Polyphred for Detection of SNPs</i>	111
2.6. STATISTICAL ANALYSIS OF DIRECTLY GENOTYPED VARIANTS	111
2.6.1. <i>Statistical Power Calculations</i>	111
2.6.1.1. Estimation of the Prevalence for Power Calculations	112
2.6.2. <i>Genetic Association Analysis</i>	112
2.6.3. <i>Meta-analysis</i>	113

2.6.4. Case-Control Rare Variant Analysis Tool (CCRaVAT)	115
2.7. GENOTYPE IMPUTATION	116
2.7.1. Input Data: Preliminary Checks & Data Filtering	116
2.7.2. MACH: Estimating Model Parameters.....	117
2.7.3. MACH: Carrying Out Genotype Imputation.....	117
2.8. STATISTICAL ANALYSIS OF IMPUTED GENOTYPES	118
2.8.1. Mach2Dat: Logistic Regression Analysis of Binary Data on Imputed Genotypes.....	118
2.8.1.1. Logistic Regression Analysis	118
2.8.1.2. Inflation Factor, Q-Q and Manhattan Plots.....	119
2.8.1.3. Regional Association Plots	120
2.8.1.4. Conditional Logistic Regression Analysis.....	120
2.8.1.5. HLA-B27-Stratified Logistic Regression Analysis	120
2.8.2. GRANVIL: Gene- or Region-Based Analysis of Variants of Intermediate and Low Frequency.....	120
2.8.2.1. Data Re-formatting & GRANVIL Analysis	121
2.8.3. Imputation Accuracy: Concordance between Imputed & Directly Genotyped SNPs.....	123
2.9. REAGENTS	124
2.10. ONLINE RESOURCES.....	125
2.11. REFERENCES	127
CHAPTER 3.....	129
3. INTERLEUKIN 23 RECEPTOR & ANKYLOSING SPONDYLITIS	129
3.1. INTRODUCTION.....	129
3.2. PATIENTS, MATERIALS & METHODS	130
3.2.1. UK Case-Control Study & Extended Study of Association between IL23R & AS 130	
3.2.2. Stratified Analysis.....	134
3.2.3. Meta-analysis	134
3.3. RESULTS	136
3.3.1. UK Case-Control Study & Stratified Analysis.....	136
3.3.2. Meta-analysis	139
3.4. DISCUSSION	145

3.4.1. UK Case-Control Study & Meta-analysis.....	145
3.4.2. Recent Studies on IL23R & AS	147
3.4.3. Statistical Considerations	148
3.4.4. T_H17 T Cell Subset.....	150
3.4.5. Related Inflammatory Diseases associated with the T_H17 Pathway: Psoriasis & Inflammatory Bowel Disease.....	154
3.4.6. T_H17 Pathway as a Therapeutic Target	156
3.5. REFERENCES	158
CHAPTER 4.....	161
4. INVESTIGATION OF <i>TNFRSF1A</i> AS A CANDIDATE SUSCEPTIBILITY LOCUS IN ANKYLOSING SPONDYLITIS	161
4.1. INTRODUCTION.....	161
4.2. PATIENTS, MATERIALS & METHODS	163
4.2.1. Re-sequencing of <i>TNFRSF1A</i>	163
4.2.2. Detection of SNPs.....	166
4.2.3. Genotyping of Possible Novel SNPs.....	168
4.2.4. Case-Control Study.....	170
4.2.4.1. Study A.....	170
4.2.4.2. Study B	175
4.2.4.3. Stratified Analyses	175
4.2.5. Meta-analysis	176
4.2.6. Haplotype Analysis.....	176
4.3. RESULTS	178
4.3.1. SNP Detection.....	178
4.3.2. Case-Control Study.....	181
4.3.2.1. Study A.....	181
4.3.2.2. Study B	181
4.3.2.3. Stratified Analyses	181
4.3.3. Meta-analysis	183
4.3.4. Haplotype Analysis.....	185
4.4. DISCUSSION	186
4.4.1. Case-Control Study, Meta-analysis & Haplotype Analysis	186
4.4.1.1. Case-Control Study & Extended Study	186

4.4.1.2. Stratified Analyses.....	188
4.4.1.3. Meta-analysis.....	189
4.4.1.4. Haplotype Analysis.....	191
4.4.2. <i>Problems with Re-sequencing of TNFRSF1A</i>	192
4.4.2.1. A Brief Summary of the Sequencing Process.....	192
4.4.2.2. Re-sequencing Problems due to the Initial Amplification Step.	192
4.4.2.3. Re-sequencing Problems due to Purification of the Sequencing Reaction	194
4.4.3. <i>Implications for the Variants of TNFRSF1A & their Associations with AS</i> 194	
4.4.3.1. <i>TNFRSF1A Variants & Inflammatory Diseases</i>	197
4.5. REFERENCES.....	201
CHAPTER 5.....	204
5. USING IMPUTATION TO FINE MAP KNOWN REGIONS & IDENTIFY LOW FREQUENCY VARIANTS ASSOCIATED WITH AS.....	204
5.1. INTRODUCTION.....	204
5.2. PATIENTS, MATERIALS & METHODS.....	207
5.2.1. <i>Patients & Quality Control</i>	207
5.2.2. <i>Imputation & Statistical Analysis</i>	209
5.3. RESULTS.....	211
5.3.1. <i>Logistic Regression Analysis</i>	211
5.3.1.1. Q-Q Plots & Inflation Factors.....	211
5.3.1.2. Genomewide Associations.....	214
5.3.2. <i>Conditional Analysis</i>	232
5.3.3. <i>Further Analyses of the ERAP1-ERAP2 Region</i>	239
5.3.4. <i>HLA-B27 Stratified Analysis</i>	243
5.3.4.1. HLA-B27 Positive Associations with AS.....	243
5.3.4.2. HLA-B27 Negative Associations with AS.....	248
5.3.5. <i>Rare/Low-Frequency Variant Analysis</i>	256
5.3.5.1. CCRaVAT Analysis.....	256
5.3.5.2. GRANVIL Analysis.....	258
5.4. DISCUSSION.....	259
5.4.1. <i>Logistic Regression Analysis</i>	259

5.4.2. Conditional Analysis for Independent Association Signals.....	265
5.4.2.1. IL23R Region	265
5.4.2.2. HLA-B & the MHC Region	266
5.4.2.3. ERAP1-ERAP2 Region	267
5.4.3. HLA-B27-Stratified Analysis	269
5.4.4. Rare/Low-frequency Variant Analysis.....	273
5.4.5. Technical & Statistical Considerations.....	276
5.5. REFERENCES	278
CHAPTER 6.....	283
6. CONFIRMATION OF IMPUTATION RESULTS BY DIRECT GENOTYPING....	283
6.1. INTRODUCTION.....	283
6.2. PATIENTS, MATERIALS & METHODS	285
6.2.1. Selection of Genes/Regions of Interest.....	285
6.2.2. Selection of SNPs.....	285
6.2.3. Sample Population, Genotyping & Statistical Analysis	292
6.2.4. Imputation Accuracy: Concordance between Imputed & Directly Typed Genotypes	293
6.3. RESULTS	294
6.3.1. Concordance between Imputation & Direct Genotyping.....	294
6.3.2. Case-Control Study & HLA-B27-Stratified Analysis.....	298
6.4. DISCUSSION	302
6.4.1. Imputed Genotypes vs. Directly Typed Genotypes	302
6.4.2. Chromosome 1q32 – HHAT	308
6.4.3. Chromosome 5q31 – IL12B.....	312
6.4.4. Chromosome 6 – rs4349859 (HLA-B27).....	313
6.4.5. Other Genes/Regions of Interest.....	313
6.4.5.1. Chromosome 19p13 – JMJD2B.....	314
6.4.6. Control Group & Selection Bias: Healthy Controls & Osteoarthritis Cases	314
6.4.7. Further Considerations & Future Work.....	315
6.5. REFERENCES	317
CHAPTER 7.....	319

7. CONCLUSIONS & FUTURE DIRECTIONS	319
7.1. GENERAL CONCLUSIONS.....	319
7.2. GENERAL DISCUSSION & FUTURE DIRECTIONS.....	323
7.2.1. <i>Gene-Set Analysis</i>	325
7.2.2. <i>Functional Studies</i>	327
7.2.3. <i>DNA Sequencing</i>	329
7.3. REFERENCES	333
APPENDICES	336
APPENDIX 1.....	336
APPENDIX 2.....	340
APPENDIX 3.....	342
PUBLICATIONS.....	355

CHAPTER 1

1. INTRODUCTION TO ANKYLOSING SPONDYLITIS

In this thesis I have used various types of genetic association studies to identify genes contributing susceptibility to the condition with a view to identifying potential new therapeutic targets.

1.1. Introduction

Ankylosing spondylitis (AS) is a common form of inflammatory arthritis with a complex aetiology. It is a polygenic disease with a major contribution from one gene, HLA-B27, which contributes about 30% of the heritability of the disease (Brown *et al.* 1997). Although treatments for the condition have improved considerably in the last 10 years, there is still no cure. Little is known about the potential environmental trigger factors for the condition and research has concentrated on exploring the full genetic component in the expectation that this will lead to the identification of potential new drug targets.

In this introductory chapter, I describe:

- 1- Clinical and pathological aspects of the disease.

- 2- Certain aspects of the immune system that are thought to be relevant to AS.
- 3- The use of genetic epidemiology to identify important biological pathways associated with AS.

AS is a highly heritable disease with a polygenic contribution. Identifying individual genetic variants that influence the risk of AS provides clues to the pathogenesis of the condition and biological pathways that are potential targets for developing new treatments. The major genetic risk variant for the condition, HLA-B27, was one of the first genetic associations to be discovered for any complex disease (Brewerton *et al.* 1973; Schlosstein *et al.* 1973). The strong familial aggregation of cases initially led Brewerton *et al.* to try to identify what genetic factors might be involved. They identified an extremely strong genetic association with the transplant antigen known then as HL-A27, now known as HLA-B27 (Brewerton *et al.* 1973). Parallel work in the USA by Schlosstein *et al.* replicated these findings (Schlosstein *et al.* 1973). The odds ratio of ~120 for developing AS associated with HLA-B27 remains among the strongest for any complex disease trait.

Subsequently, it also became apparent that other disorders, including psoriasis, inflammatory bowel disease and iritis were more prevalent in the families of index cases with AS. However, HLA-B27 is not related to psoriasis or IBD, and so does not explain the co-segregation of psoriasis or IBD in these families, or in individuals with AS themselves, which suggests that there are other genetic factors contributing susceptibility to all these conditions. In recent years,

following the application of genomewide association techniques, fourteen further genes/loci have been implicated in the condition (Burton *et al.* 2007; Reveille *et al.* 2010; Evans *et al.* 2011). In the Wellcome Trust Case Control Consortium 2 (WTCCC2) genomewide association study (GWAS), associations with *ERAP1*, *IL23R*, *KIF21B*, *HLA-B*, *2p15*, *21q22*, *RUNX3*, *IL12B*, *LTBR-TNFRSF1A*, *ANTXR2*, *PTGER4*, *CARD9* and *TBKBP1* were observed along with the *HLA-B27-ERAP1* interaction in AS (Evans *et al.* 2011). However, like most complex diseases, the majority of genetic variation underlying the condition remains unexplained- a phenomenon called the “missing heritability” problem (Maher 2008; Manolio *et al.* 2009).

The most important goal for understanding the genetic basis of AS is to facilitate a clearer understanding of the pathological processes involved and thereby, to identify potential therapeutic targets for the condition. Even weak genetic associations could reveal potential novel forms of treatment as exemplified by the association between CTLA4 (modulates T cell activation through the CD28/B7 signalling pathway) and rheumatoid arthritis (RA). Despite the weak effect size (Odds ratio ~ 1.05), there is a highly beneficial therapeutic effect of human recombinant CTLA4/immunoglobulin fusion protein (abatacept) in RA (Plenge *et al.* 2005; Furst *et al.* 2012).

1.1.1. Clinical Features & Pathology of AS

Ankylosing spondylitis (AS) is a common chronic inflammatory arthritis, predominantly affecting the spine and sacroiliac joints. Typically it causes inflammation and pain in the axial skeleton followed by progressive spinal fusion

and stiffness. Early in the disease inflammatory changes predominate, particularly around the sacroiliac joints. Subsequent new bone formation, arising at the sites of spinal inflammation, causes progressive spinal fusion and the development of a progressively flexed posture and the classic radiographic appearance of a “bamboo spine” (Figure 1.1). The inflammatory process particularly affects sites of high mechanical stress, such as entheses, including the sacroiliac joint, the Achilles tendon and plantar fascia.



Figure 1.1 (A)

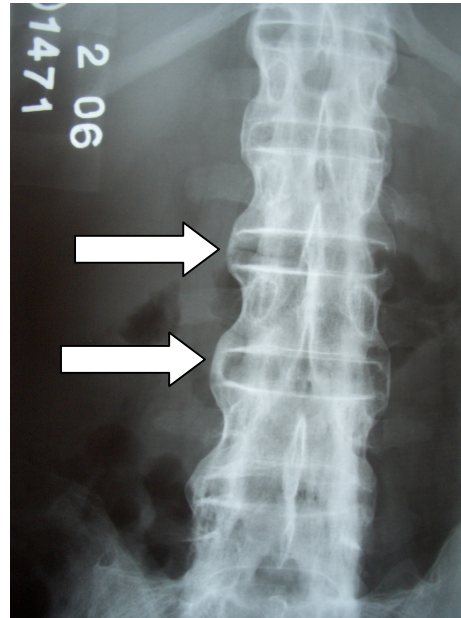


Figure 1.1 (B)

Figure 1.1 (A) Classic “question mark” posture of advanced ankylosing spondylitis due to progressive fusion of the spine leading to flexed posture. **(B)**, Anteroposterior radiograph of the lumbar spine showing multiple bridging syndesmophytes (arrows) causing the appearance of a classic bamboo spine.

Extra-spinal features are also common in AS and include anterior uveitis (iritis), peripheral arthritis and low-grade inflammation of the small intestine. Anterior uveitis has previously been reported in 5-35% of AS cases (Careless & Inman 1995) and was 40% in our cohort. Arthritis occurs more frequently in the lower limb joints; hip involvement is particularly observed constituting an adverse prognostic indicator, and osteoporosis is common as well (Vander Cruyssen *et al.* 2010; Sambrook & Geusens 2012). Approximately 15 per cent of individuals with AS may also develop psoriasis and/or inflammatory bowel disease (IBD), conditions that can in part be regarded as part of the same overarching disease spectrum as AS (Feldtkeller *et al.* 2003; El Maghraoui 2011).

The absolute risk of developing AS is about 2-8% in HLA-B27 positive individuals and 90% of the AS cases are HLA-B27 positive in Caucasian populations. The disease risk is approximately 13% in HLA-B27 positive first-degree relatives of HLA-B27 positive AS patients (Brewerton *et al.* 1973; van der Linden *et al.* 1984; van der Linden & van der Heijde 1998; Sieper *et al.* 2006; Khan 2010). The onset of AS is typically between the ages of 15 and 35 (80% of individuals with AS develop the disease before the age of 30 years) (Feldtkeller *et al.* 2003) and it occurs more than twice as commonly in men than women (Calin 1985; Calin *et al.* 1999). Typically axial skeletal inflammation begins in the second decade with onset of low back pain and alternating buttock pain. Sacroiliitis due to inflammation of the sacroiliac joints is usually the earliest feature of AS, and is a prerequisite for its diagnosis. Skeletal radiographs were formerly used to identify sacroiliitis, but it is now more appropriate to use magnetic resonance imaging

(MRI), which can identify inflammation in the sacroiliac joints much earlier in the disease (Figure 1.2).

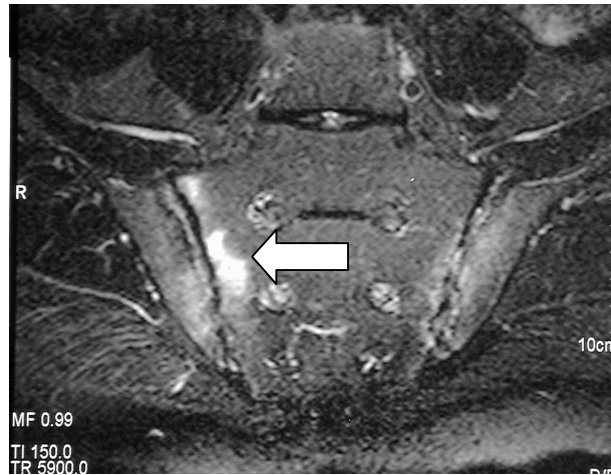


Figure 1.2. MRI of the sacroiliac joints showing high signal predominantly on the sacral side of the right SI joint (arrow).

AS causes significant physical and psychological pain, decreased quality of life and reduced lifespan. The annual cost of AS was estimated as £6,765 per patient in 2002 (Kobelt *et al.* 2004), and the direct costs alone (pre-anti-TNF agents) have been estimated in European countries at €1,800-2,800 per patient annually (Boonen & Severens 2002) and \$1,750 in the United States (Ward 2002).

Although anti-TNF drugs (e.g. infliximab, etanercept) produce improvements in acute inflammation in AS, there are no treatments which have to date induced remission of AS, or been proven to retard the progressive joint fusion that typifies the disease. There is thus an urgent need for the development of effective new therapies and a better understanding of the pathogenesis of the condition.

1.1.2. Definition of the Phenotype: Diagnostic Criteria

1.1.2.1. The Modified New York Criteria

For epidemiological research, it is essential to have an accurate definition of the disease being studied. Nowhere is this more evident than in the study of arthritis. The original 1963 Rome criteria for AS highlighted the importance of spinal symptoms and signs, but used rather vague radiological signs of sacroiliac joint involvement. The 1968 New York revision of these criteria initiated a grading system for radiographic changes in the sacroiliac joints and excluded iritis as a clinical criterion (Moll & Wright 1973). In 1984, the modified New York criteria were proposed, which emphasized the inflammatory quality of the back pain and the need for its chronicity (≥ 3 months). The 1984 modified New York criteria were widely adopted for disease classification and for use in the clinic (van der Linden *et al.* 1984). Several other diagnostic classification criteria were subsequently proposed, but they generally lacked specificity (Goie The *et al.* 1985). MRI of the axial skeleton has now often replaced radiography for the early diagnosis of inflammatory changes in AS, but it is not universally available. The “Assessment in Ankylosing Spondylitis” (ASAS) society have developed MRI based criteria with increased sensitivity for detecting early disease, although they probably lack specificity for AS (Cheung *et al.* 2012). The patients described in my studies all fulfill the 1984 modified New York criteria, which provide the high degree of specificity for the diagnosis of AS essential for studies of the genetic epidemiology of disease.

The Modified New York Criteria 1984

Clinical criteria:

- Low back pain with duration ≥ 3 months that is improved by exercise and not relieved by rest.
- Limited sagittal and coronal back movement.
- Reduced chest expansion (compared to age and sex-matched values).

Radiographic criteria:

- Bilateral sacroiliitis (grade ≥ 2).
- Unilateral sacroiliitis (grade ≥ 3).

Grade 0 is normal; Grade 1 is suspicious; Grade 2 indicates sclerosis around the joint; Grade 3 is widening, erosion and sclerosis on both sides of the SI joint; Grade 4 is partial or complete ankylosis.

AS is diagnosed if significant radiographic evidence of sacroiliitis is present along with one or more clinical criteria. MRI allows earlier detection of sacroiliitis and enthesitis. Thus, it is now preferred over X-ray imaging in the early diagnosis of AS.

1.1.3. Management & Treatment

The management of AS has traditionally relied on physical therapy to maintain mobility and analgesics to relieve pain. Only in recent years have drugs that could truly be regarded as modifying the course of the disease been introduced.

1.1.3.1. Drugs

1.1.3.1.1. Non-Steroidal Anti-Inflammatory Drugs (NSAIDs)

Simple analgesics can be used to relieve pain but, more commonly, NSAIDs are used. These reduce pain and stiffness while facilitating mobility, and possibly have a weak effect on retarding the progression of spinal ankylosis (Wanders *et al.* 2005; El Maghraoui 2011).

1.1.3.1.2. Anti-TNF Treatment

The treatment of refractory forms of rheumatoid arthritis with anti-TNF biologics started in the 1990s. They have proved remarkably successful and safe, and have subsequently been introduced for the management of many other inflammatory diseases, including psoriasis, psoriatic arthritis and inflammatory bowel disease (Burmester *et al.* 2012; van Dartel *et al.* 2012). They have proved to be at least as effective in the management of AS (Davis *et al.* 2003; van der Heijde *et al.* 2009; Braun *et al.* 2012). Currently in the UK three anti-TNF biologic agents are recommended by the National Institute for Health and Clinical Excellence (NICE) for use in refractory cases of AS; adalimumab, golimumab and etanercept. These treatments are highly effective in symptom relief and clearly have a major effect in reducing local bone inflammation as judged by MRI. Their effects on quality of life and benefits in health economics terms have been clearly demonstrated (Braun *et al.* 2003; Boonen *et al.* 2006). However, there are some doubts about whether these expensive agents actually suppress new bone formation and retard progression of the disease (van der Heijde *et al.* 2008).

1.1.3.1.3. Potential New Treatments

None of the current AS treatments are truly curative nor are there any proven preventative strategies. It is clearly desirable to develop such agents/strategies and/or less expensive alternatives to the biologic therapies currently available. Some of the newer agents for AS currently being assessed in late-stage human clinical trials include monoclonal antibodies (e.g. secukinumab, Novartis) targeting IL17, a logical drug target given the association of the T_H17 pathway in AS. Other monoclonal antibodies targeting components of this pathway include ustekinumab, which targets the common IL12 β 1 chain of interleukins-12 and -23. This has been shown to be effective in psoriatic arthritis and skin psoriasis, but has not yet been used in AS (see Chapter 3 for more details) (Gottlieb *et al.* 2009; Maddur *et al.* 2012; Zhu & Qian 2012).

1.1.3.2. Physical Therapy

Physical therapy is an effective treatment to preserve mobility for those with AS, and has been extensively used not least because there were, until recently, few alternatives. The natural history of AS leads to progressive ankylosis of the spine and an increasingly flexed posture. At its worst, this can leave the patient unable to raise the head to see directly ahead and to engage in a social conversation. Exercise regimes emphasizing the importance of spinal extension and the maintenance of mobility have been shown to improve not only posture and flexibility, but also to reduce pain, at least in the medium term (Wordsworth *et al.* 1984; Dagfinrud *et al.* 2008).

1.2. Epidemiology

1.2.1. Prevalence of AS

The prevalence of AS is highly variable between different ethnic groups. It is most prevalent in the circumpolar regions of the northern hemisphere, particularly among the Inuit and certain Amerindian groups. In contrast, it is rare in indigenous sub-Saharan Africans and entirely absent from Australian aborigines. In the developed world, its prevalence appears to be approximately 1-5 per thousand. There are no definitive studies from the UK, but a large radiograph-based study in Holland, estimated the prevalence to be 0.1-0.2% (van der Linden *et al.* 1984). A higher prevalence of 0.55% has been reported in Germany (Braun *et al.* 1998; Braun & Sieper 2007), and a prevalence of 0.24% was observed in Greece (Andrianakos *et al.* 2003). In two non-European studies, the prevalence of AS was estimated as 0.4% in Alaskan Eskimos (Boyer *et al.* 1994) and 0.1% in Chinese (Dai *et al.* 2003). These estimates are dependent on the methods of ascertainment and the diagnostic criteria used to define the disease. The prevalence of AS tends to reflect the local population prevalence of the major genetic risk factor HLA-B27 (See Section 1.4.2.2).

1.2.2. Male Predominance & Age at Onset

AS more commonly affects males; the prevalence of AS in males is estimated as 2.5 times of that in women (Calin 1985; Calin *et al.* 1999). However, the male predominance becomes less obvious with increasing age. Furthermore, HLA-B27 negative patients tend to have a later age of onset (~27.7) compared to the HLA-B27 positive AS patients (~24.8). AS usually develops in young adulthood with

only 5% of individuals developing the disease after 45 years old. However, the diagnosis is often delayed by 6 years or more. This delay may in part be due to the late emergence of sacroiliitis on radiographs – a key historical requirement for the diagnosis of AS, or it may reflect the often indolent onset of the disease (Feldtkeller *et al.* 2003).

1.3. Immunology of Ankylosing Spondylitis

The strongest early evidence for the involvement of the immune system in AS came from the discovery of its association with *HLA-B27*, one of the immune response genes (Brewerton *et al.* 1973; Schlosstein *et al.* 1973). It is also important to recognise that there is a strong relationship between AS and another form of spondyloarthropathy known as reactive arthritis, which clearly represents an abnormal host response to a variety of bacterial triggers. Longitudinal studies suggest that up to one-quarter of patients with reactive arthritis may eventually develop AS (Thomson *et al.* 1995). There has also been substantial interest in the possibility that flares of AS may be triggered by increased faecal carriage of bacteria, such as *Klebsiella aerogenes* (Ebringer 1989). The complexity of immune/inflammatory responses is such that it is critical to have an understanding of the workings of the immune system as a foundation for interpreting the relevance of genetics associations with AS. In this and other related inflammatory diseases, it is now recognized that there may be multiple genetic associations with individual components of important immune/inflammatory pathways. For example, in the T_H17 pathway, the interleukin 23 receptor (*IL23R*), interleukin 12 (*IL12B*) and signal transducer and

activator of transcription 3 (*STAT3*) are associated with AS. In the MHC class I antigen presentation pathway, *HLA-B27* and endoplasmic reticulum aminopeptidase 1 (*ERAP1*) are associated with AS (see Chapters 3, 4 and 5 for details of these associations) (Burton *et al.* 2007; Harvey *et al.* 2009; Karaderi *et al.* 2009; Danoy *et al.* 2010; Reveille *et al.* 2010; Evans *et al.* 2011; Karaderi *et al.* 2012).

The immune system plays a crucial role in maintaining anti-microbial defence mechanisms, but it can also cause deleterious inflammation directed at the host tissues. Activation of the immune system typically provokes an inflammatory response of the sort that is commonly observed in many inflammatory disorders, including AS and other inflammatory arthritides. In most of these idiopathic disorders the precise trigger for the inflammatory/immunological response is unknown. However, the genetic associations of many of these conditions with components of the immune system strongly suggest that some form of dysregulated immunological reactivity is involved. Intense infiltration of the affected tissues by activated effector cells has also been well described in AS (Appel *et al.* 2006a; Appel *et al.* 2006b; Appel *et al.* 2006c). However, despite the fact that association of AS with HLA-B27 has been known for nearly 40 years the underlying causal mechanisms are still unclear.

1.3.1. Components of the Immune System

1.3.1.1. Innate Immunity

Innate immune responses are relatively non-specific and are found in almost all forms of life. In humans, the innate immune system has cell-mediated and humoral components. There is no innate immunological memory. Innate immunity comprises components as simple as physical barriers, such as skin and mucous membranes, as well as more intricate mechanisms including the complement system, cytokines (interleukins, chemokines, interferons), pattern recognition receptors (e.g. toll-like receptors), cellular barriers such as leukocytes (macrophages, neutrophils, dendritic cells, mast cells, eosinophils, basophils, natural killer cells). Innate immunity is genetically programmed for the detection of common or invariant features of various microbes. Pattern recognition to detect these microbes is the main response mechanism in innate immunity. Pattern recognition receptors (PRRs) recognise the pathogen-associated molecular patterns (PAMPs) and upregulate the immune response, including specific immune responses (Medzhitov & Janeway 2000; Iwasaki & Medzhitov 2010).

1.3.1.1.1. Pattern Recognition Receptors

One particular family of PRRs that has been associated with AS in Finnish and Canadian patients is the transmembrane toll-like receptors (TLRs), specifically variants in *TLR4* (Snelgrove *et al.* 2007; Pointon *et al.* 2008). TLR4 on the cell surface recognises lipopolysaccharide (LPS) derived from gram negative bacteria. Transmembrane PRRs are expressed in specific cell types such as dendritic cells

(DCs) and macrophages, where they lead to cell-extrinsic innate immune recognition by mostly non-infected cells (Stetson 2009). TLRs also recognize endogenous signals, such as chromatin fragments or ribonucleoprotein complexes, originating from injured tissues and dead cells. If these potential signals are not cleared sufficiently, they can lead to TLR activation on DCs and B cells. Such activated TLRs may lead to systemic autoimmune diseases (Marshak-Rothstein & Rifkin 2007). This may be particularly relevant to forms of reactive arthritis triggered by bacterial infection.

Another type of PRRs important for AS is cytosolic receptors such as nucleotide-binding oligomerisation domain (NOD)-like receptors (NLRs). CARD9 is an NLR protein that has an N-terminal caspase-recruitment domain. It was found to be associated with IBD and subsequently, with AS (Zhernakova *et al.* 2008; Pointon *et al.* 2010a). NLRs are members of a large family of intracellular receptors that act as sensors for intracellular pathogens and stress signals. NLRs recognise degradation products of peptidoglycans from bacterial cell walls, stress signals such as UV radiation, and other microbial products, allowing cell-intrinsic innate immune responses by infected/damaged cells (Martinon *et al.* 2009). Thus, these receptors are expressed by most cells that are likely to be infected by pathogens, such as viruses (Stetson 2009).

1.3.1.1.2. Killer Immunoglobulin-like Receptors

In a meta-analysis of genomewide linkage studies in AS, the chromosome 19q13 region adjacent to the killer immunoglobulin-like receptor (KIR) genes showed a linkage signal (Carter *et al.* 2007). These genes are expressed on some T cell

subsets and natural killer (NK) cells, which are important for innate immune response against tumour cells and virally infected cells. These responses involve interactions between specific *KIR* and *HLA* alleles. Association between AS and the *KIR* gene cluster has been variable. Some studies have reported associations between *KIR3DL1* and *KIR3DS1* and AS (Lopez-Larrea *et al.* 2006; Diaz-Pena *et al.* 2010). In contrast, studies from the UK showed no obvious associations with *KIR* haplotypes or specific 12 *KIR3DL2* alleles (Harvey *et al.* 2009). Moreover, an association between alleles of *KIR3DL1* and *KIR3DS1* and AS was not found, supporting the previous findings (McCappin *et al.* 2010). However, the range of allelic variation at the *KIR* loci and the complexity of their interactions with *HLA* alleles renders any possible association with diseases, such as AS, extremely difficult to evaluate. Further studies of large AS populations compared with carefully *HLA* and *KIR* matched controls are needed to assess any potential *KIR* associations with AS systematically (See Chapter 7, Section 7.2.2).

1.3.1.1.3. Leukocytes: Cell-mediated Innate Immune Response

Various cell types of the leukocyte family are responsible for the cell-mediated immune component of the innate immune response, but also for activation and implementation of the adaptive immune response. Leukocytes include neutrophils, eosinophils, basophils, monocytes, macrophages and dendritic cells (DCs). Antigen-presenting cells (APCs), such as macrophages and DCs, express MHC class I and II receptors. Typically endogenous antigens derived from viruses and tumours are presented bound to the MHC class I receptors expressed by all nucleated cells, whereas exogenous antigens (e.g. from bacterial proteins) are

presented by dedicated APCs bound to MHC class II molecules (See Section 1.3.1.4).

1.3.1.2. Adaptive Immunity

There has been substantial interest in the role of the adaptive immune system in AS particularly because of the strong association with *HLA-B27*, an MHC class I gene involved in specific immune responses. Adaptive immunity is found in jawed vertebrates. Adaptive immune responses are pathogen- and antigen-specific, and there is a lag time between exposure and response reflecting the activation of antigen-specific memory cells. Furthermore, there are cell-mediated and humoral components to adaptive immunity. Immunological memory is a crucial feature of the adaptive immune response. Since it is antigen-specific, adaptive immunity requires the ability to distinguish between self and non-self in order to regulate the immune response. Adaptive immunity makes use of antigen receptors that are only partially encoded in the germ-line DNA. Further diversity is generated by the splicing together of various germ-line DNA segments with intervening diversity fragments leading to a specific and strong immune response (Iwasaki & Medzhitov 2010). The main cellular components of adaptive immunity are lymphocytes of two main types; B cells and T cells.

Activation of pathogen-specific receptors (e.g. PRRs) of the innate immune system upregulates antigen-specific adaptive immune responses. PRRs activating NF- κ B, IRF and NFAT, the main transcription factors, lead to the induction of both adaptive T and B cell responses (Palm & Medzhitov 2009). TLRs can direct components of the adaptive immune response, such as 1) immunoglobulin M (IgM), IgG and IgA antibody responses, 2) helper T cell - T_H1, T_H2 and T_H17 CD4+

T cell response, 3) cytotoxic CD8⁺ T cell response (Iwasaki & Medzhitov 2004). Specifically, signalling through TLR4 can lead to T_H2 and IgE antibody responses, whereas signaling through other PRRs, dectin-1 and -2, can lead to T_H17 responses against fungal infections (Robinson *et al.* 2009).

1.3.1.2.1. T Lymphocytes & Cell-mediated Immunity

T cells are responsible for the cell-mediated immune response of the adaptive immunity. T cells are produced in bone marrow and develop in the thymus where positive and negative selection takes place in order to generate a population of mature T cells that is tuned to recognize foreign antigens efficiently, but to be tolerant of self proteins. During thymic development, T cells with antigen receptors that recognize self-proteins bound to MHC molecules on the surface of thymic epithelial cells, undergo deletion from the repertoire by apoptosis. In contrast, those T cells with receptors that recognize MHC but not MHC/self protein undergo further development to become mature T cells. It is proposed that auto-reactive T cells escaping this selection lead to organ-specific autoimmune diseases upon activation by APCs (Gery & Egwuagu 2002). Foreign antigens, in the form of small peptide fragments, bound to the appropriate MHC receptors (class I or II) are typically presented by APCs to the T cells. Those T cells with receptors recognizing the MHC/antigen become activated and undergo a proliferative response (Yamane & Paul 2012).

- a. Cytotoxic T cells: These T cells also express CD8 co-receptors at the cell surfaces and are typically activated through abnormal endogenously derived antigens presented bound to MHC class I receptors found on all nucleated cells. These CD8⁺ cells exert

cytotoxic effects on cells that express abnormal antigens following virus infection or tumour antigens following malignant transformation. AS-associated *HLA-B* encodes an MHC class I receptor which is involved in such cytotoxic T cell responses (Brewerton *et al.* 1973; Schlosstein *et al.* 1973). To date, there is no evidence that viral infection plays any role in AS.

- b. Helper T cells: These T cells express CD4 co-receptors on their surfaces in addition to the T cell receptor (TCR). They are activated by antigen presentation in the context of MHC class II receptors expressed mainly on immune cells, such as DCs. Helper T cells can differentiate into one of a number of subsets depending on the type of antigen and cytokine exposure. These subsets include T_H1 , T_H2 , T_H17 and T_{REG} (Figure 1.3). Activated helper T cells eventually may become memory T cells, which can respond rapidly to antigenic re-stimulation. Cytokines produced by helper T cells facilitate maintenance of that T cell subset itself, as well as potentially stimulating macrophages, cytotoxic T cells and B cells. Helper T cells are important for regulation of both innate and adaptive immune responses.

- i. T_H1 T cell subset: This type of response is typically directed against microbes that activate macrophages and natural killer (NK) cells, as well as against viruses. T_H1 T cells produce interferon-gamma ($IFN\gamma$) that activate B cells and macrophages, and inhibits T_H2 responses (Figure 1.3) (Romagnani 1996).

- ii. T_H2 T cell subset: This type of response occurs against helminths, extraneous allergens, extracellular microbes and toxins. T_H2 T cells produce interleukins 4 (IL4), 5 (IL5) and 13 (IL13). IL4 inhibits T_H1 responses, and together with IL13 stimulates B-cells to produce IgE and IgG, whereas IL5 stimulates eosinophils (Figure 1.3) (Kidd 2003).
- iii. T_H17 T cell subset: These T cells are formed in the presence of the cytokines IL6 and TGF- β (Figure 1.3). They produce IL17, stimulating stromal cells to make pro-inflammatory cytokines, such as IL6 and IL8. The T_H17 pathway is strongly associated with certain inflammatory diseases such as AS, IBD and psoriasis (See Chapters 3 for a detailed discussion of the T_H17 pathway) (Boniface *et al.* 2008).
- iv. Regulatory T (T_{REG}) cell subset: Regulatory T cells help to control the immune responses. They require TGF β for differentiation and are characterised by the master transcription factor FOXP3. IL6 is the key balancing cytokine, which suppresses T_{REG} differentiation while promoting T_H17 differentiation (Wahl *et al.* 2004; Oukka 2007; Kimura & Kishimoto 2010; Leung *et al.* 2010).

DIFFERENTIATION OF CD4⁺ T CELL LINEAGE

(MHC Class II antigen presentation)

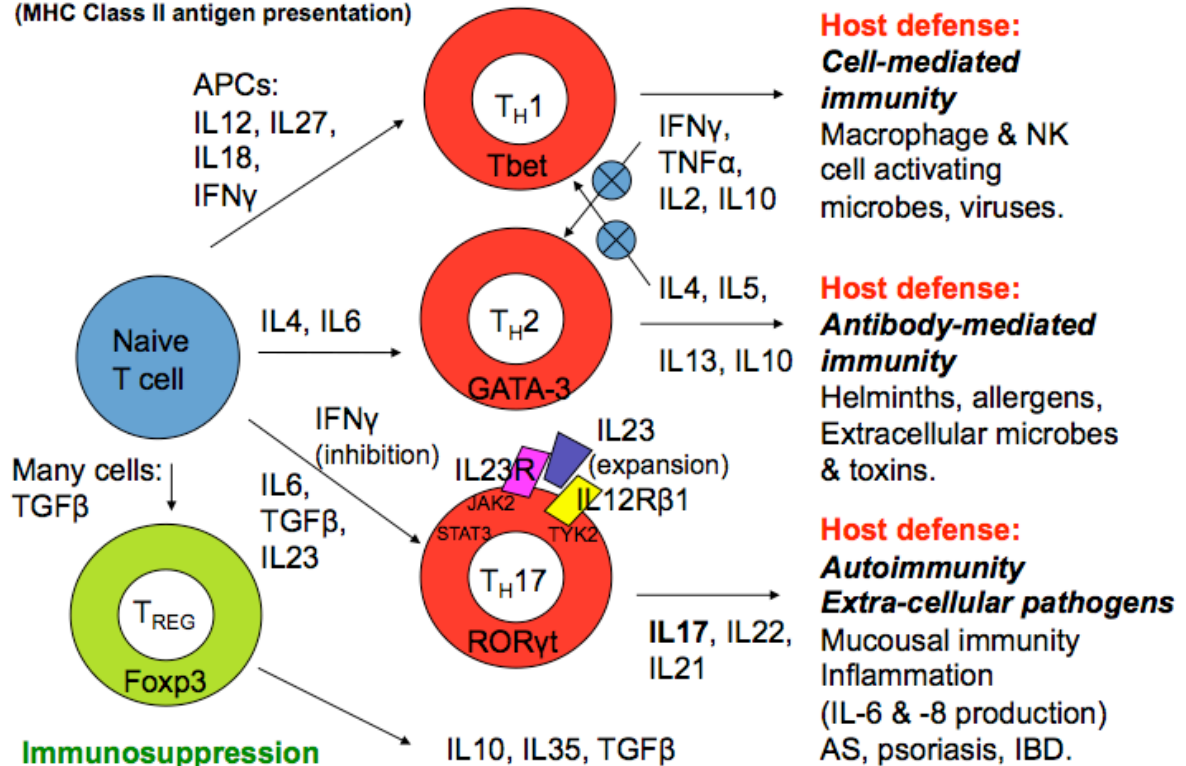


Figure 1.3. Differentiation of the CD4⁺ T cells into various T cell subsets; T_H1, T_H2, T_H17 and T_{REG}. The lineage specific cytokines drivers and transcription factors are indicated together with the major effects of terminally differentiated cells. Note that the IL12 β 1 chain is common to both the IL23 receptor and the IL12 receptor (Mosmann & Coffman 1989; Romagnani 1994; Kimura & Kishimoto 2010; Pillai *et al.* 2011; Annunziato *et al.* 2012; Wang *et al.* 2012; Zhang *et al.* 2012).

1.3.1.3. Major Histocompatibility Complex

The major histocompatibility complex (MHC) is located on the short arm of chromosome 6. All vertebrates have an MHC but the gene composition within the region varies. There are about 200 MHC genes in humans, many of which have immunological functions, including the human leukocyte antigens (HLA) that function as antigen-presenting proteins (Thorsby 2009). The MHC is divided into three regions:

- 1) MHC Class I: This includes the HLA-A, HLA-B and HLA-C loci, which encode the HLA class I molecules involved in cytotoxic responses to virus infections mediated by CD8⁺ T cells. HLA class I molecules are expressed on all nucleated cells and form heterotrimeric complexes (HLA/ β_2 -microglobulin/antigenic peptide) on the cell surface where they can be recognised by T cells with the appropriate antigen receptor. AS is strongly associated with the MHC class I molecules HLA-B27 (Brewerton *et al.* 1973; Schlosstein *et al.* 1973) and HLA-B60 (Robinson *et al.* 1989; Brown *et al.* 1996; Brown *et al.* 1997).
- 2) MHC Class II: HLA-DPA1, HLA-DPB1, HLA-DQ1, HLA-DQB1, HLA-DRA and HLA-DRB1 are encoded by this class of MHC genes which lie centromeric to MHC class I. These molecules also form heterotrimeric complexes including peptide antigens, but in this case the antigen binds to heterodimers of α and β chains of HLA-DR, DQ or DP. They are particularly expressed by immune cells such as APCs (macrophages and dendritic cells), activated T cells and B cells. A member of this class, HLA-DRB1, is associated with AS (Brown *et al.* 1997; Brown *et al.* 1998; Sims *et al.* 2007).

- 3) MHC Class III: this region contains many diverse genes, some of which have immunological functions. These include the C3 and C4 components of the complement cascade, TNF, lymphotoxin and heat shock proteins.

In addition to the highly polymorphic nature of the MHC, there is a high level of linkage disequilibrium between its constituent genes making genetic studies more complicated. This intricate linkage causes high correlation between the genetic variants at different loci and makes it difficult to distinguish between a primarily causal variant and those that are correlated with it. This is particularly well demonstrated by the existence of numerous extended MHC haplotypes, perhaps the best known being the HLA-A1/B8/DR3/DQ2 haplotype.

1.3.1.3.1. The Evolution of Major Histocompatibility Complex

The MHC is the most diverse region in the human genome. Much of the allelic diversity is due to evolutionary pressure to maintain an effective immune response to a wide range of infectious agents (Hill *et al.* 1994). Balancing selection, acting through mechanisms such as frequency-dependent selection and heterozygote advantage, is responsible for the conservation of a range of different alleles for each HLA gene since no single allele is the most fit providing the best survival advantage (Potts & Slev 1995; Apanius *et al.* 1997). A significant example of the evolution of the MHC and maintenance of its polymorphism primarily through natural selection by pathogens has been demonstrated by Hill *et al.* in West Africans (Hill *et al.* 1991). In about 2000 West African children, an MHC class I antigen (HLA-Bw53) and an MHC class II haplotype (HLA-DRB1*1302-

DQB1*0501) were associated with protection against severe malaria. These variants were common in West Africans, but rare in other populations, showing the importance of pathogens in shaping the MHC diversity.

1.3.1.4. Antigen Processing & Presentation

Antigen processing and presentation appear to play a significant role in susceptibility to AS as variants in *HLA-B* (MHC class I) and *ERAP1* are strongly associated with the disease (Burton *et al.* 2007; Harvey *et al.* 2009; Reveille *et al.* 2010; Evans *et al.* 2011). Furthermore, an interaction between *ERAP1* and *HLA-B27* has also been identified in a genome-wide study, where the *ERAP1* association was only observed in *HLA-B27* positive cases (Evans *et al.* 2011). Moreover, the potential importance of autophagy, which also plays a role in antigen processing, in inflammatory diseases related to AS was highlighted by the identification of an association of Crohn's disease with autophagy genes including *ATG16L1*, *IRGM* and *NOD2* (*CARD15*). *ATG16L1* is a protein required for elongation of the autophagosomal membrane. Immunity-related GTPase family M (*IRGM*) encodes a protein necessary for degradation of intracellular microorganisms, and nucleotide-binding oligomerisation domain-containing protein 2 (*NOD2*) encodes a protein important for the initiation of autophagy (Hampe *et al.* 2007; Rioux *et al.* 2007; McCarroll *et al.* 2008; Brain *et al.* 2012; Lapaquette *et al.* 2012).

Presentation of antigens by MHC class I molecules involves a number of steps (Figure 1.4). Recognition of antigens presented by MHC class I molecules stimulates CD8⁺ T cell response against virally infected cells and malignant

transformations. These antigenic peptides presented by the MHC class I molecules are endogenous, 8-9 amino acids in length, and may originate either from the cytosol or the nucleus (Neefjes *et al.* 2011).

Complex intracellular antigens are degraded in the cytoplasm by proteasomes to smaller peptides, which are translocated via the transporter associated with antigen presentation (TAP) into the endoplasmic reticulum (ER) lumen, where they are loaded on to nascent MHC class I molecules (Procko & Gaudet 2009). TAP typically translocates antigenic peptides 8 to 16 amino acids in length into the ER lumen (Parcej & Tampe 2010). Peptides longer than 8-9 amino acids (optimal length for MHC class I loading) require further trimming in the ER, a function carried out by ER aminopeptidases 1 (ERAP1) and 2 (ERAP2) (Saveanu *et al.* 2005).

MHC class I molecules are heterodimers made of a polymorphic heavy chain and an invariant light chain called β_2 -microglobulin. This heterodimer is stabilised by either binding to the antigenic peptide (Elvin *et al.* 1991), or ER chaperone proteins such as calreticulin, ERp57 and tapasin. TAP, tapasin, MHC class I molecule, calreticulin and ERp57 form the peptide-loading complex (PLC). Tapasin promotes peptide translocation into ER by interacting with TAP and thus, coupling peptide translocation to peptide loading onto the MHC class I molecules. Upon antigenic peptide binding, ER chaperones and tapasin are freed back to the ER lumen. Peptide–MHC class I complexes are then released from the ER and transported via the Golgi to the plasma membrane for antigen presentation to the cytotoxic CD8⁺ T cells. Peptides that are not bound to MHC class I molecules are

transported back to the cytosol through ER-associated degradation (ERAD) pathway for either destruction or re-translocation to the ER via TAP (Vyas *et al.* 2008; Neefjes *et al.* 2011).

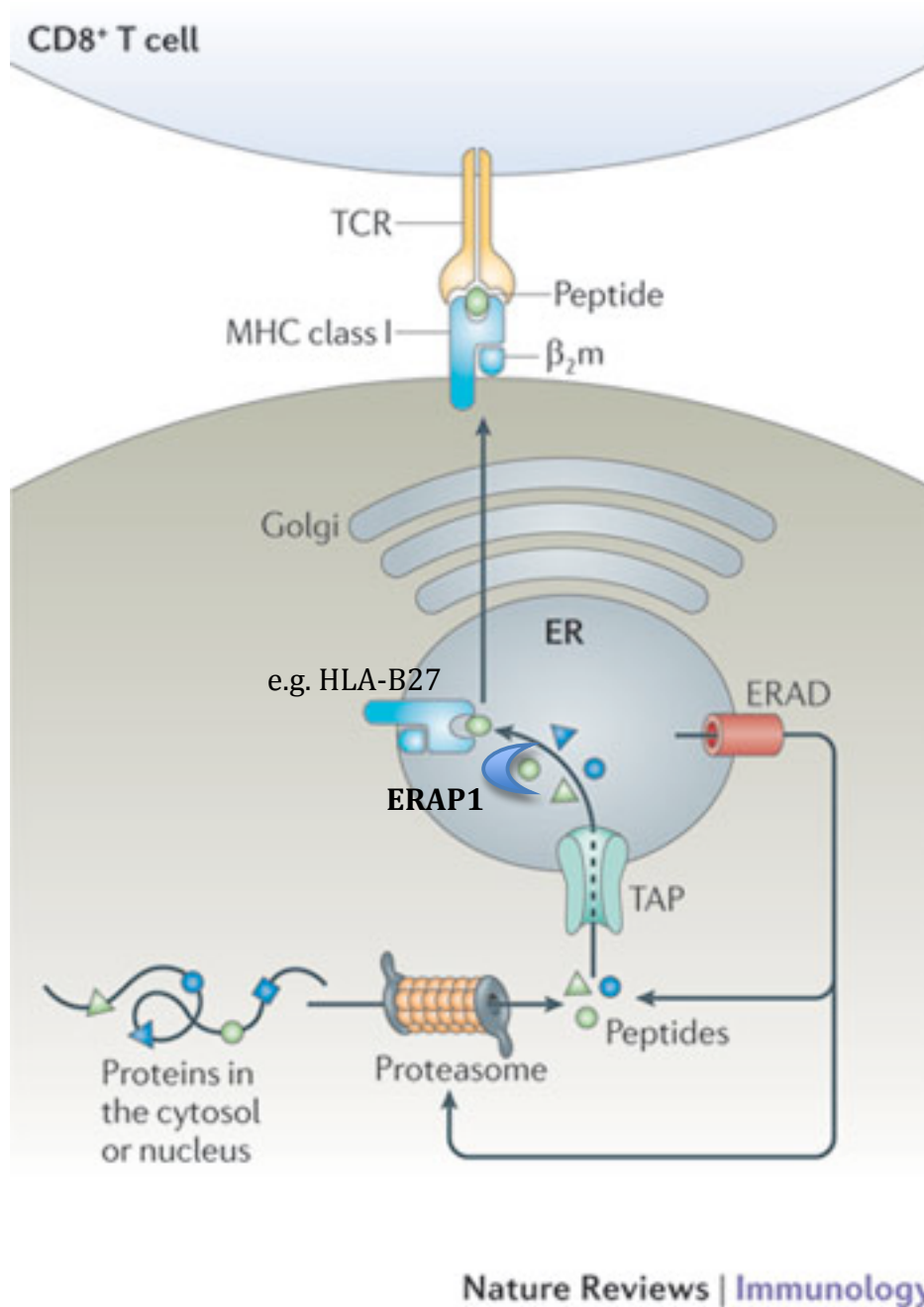


Figure 1.4. The presentation of intracellular antigenic peptides by MHC class I molecules (e.g. HLA-B27) as a result of a series of reactions. (β_2m , β_2 -microglobulin; ERAD, ER-associated protein degradation; TCR, T cell receptor.) Adapted and reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Immunology (Neeffjes *et al.* 2011), copyright (2011).

MHC class I molecules may also present exogenous peptides originating from outside the cell (Kurts *et al.* 2010). Exogenous peptides are generally degraded by lysosome-associated proteases and presented by the MHC class II molecules on macrophages and dendritic cells (DCs) thereby inducing a CD4⁺ T cell response. However, it is proposed that autophagy, which involves lysosomal degradation of damaged cellular organelle and macromolecules sequestered in vacuoles called autophagosomes, may play a role in both MHC class I and class II antigen presentation (Levine & Kroemer 2008; Vyas *et al.* 2008; English *et al.* 2009). Autophagy possibly leads to exogenous antigen presentation (i.e. cross-presentation) by the MHC class I molecules on DCs that engulf autophagic dying cells by phagocytosis (English *et al.* 2009; Uhl *et al.* 2009).

1.3.1.5. Innate Immune Recognition & Control of Adaptive Immune Response

Recognition of the pathogen-associated molecular pattern (PAMP) by pattern-recognition receptors, such as the TLRs, generates signals that activate the adaptive immune system (Figure 1.5). Endocytic pattern-recognition receptors, such as the macrophage mannose receptor, bind to components of pathogenic cell walls and mediate phagocytosis of pathogens by antigen-presenting cells such as DCs and macrophages. Pathogenic proteins are degraded in the lysosomes to generate antigenic peptides, which form a complex with MHC class II molecules on the surface of the antigen presenting cells. These peptides are recognized by CD4⁺ T cell receptors leading to activation of the T cells (Medzhitov & Janeway 2000; Vyas *et al.* 2008; Neefjes *et al.* 2011).

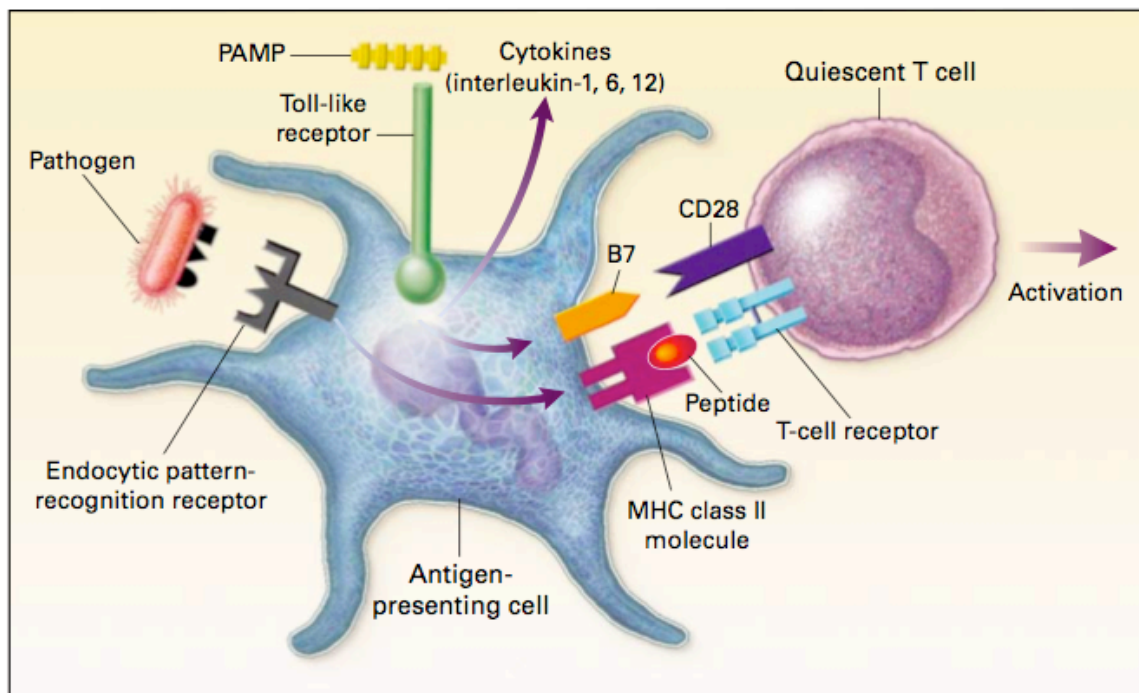


Figure 1.5. The recognition of pathogen-associated molecular patterns by toll-like receptors (TLRs) leads to the activation of signaling pathways that induce the expression of cytokines, chemokines, and costimulatory molecules. Thus, pattern-recognition receptors of the innate immune system have a role in the generation of both the peptide–MHC-molecule complex and the costimulation required for the activation of T cells. Reproduced with permission from (Medzhitov & Janeway 2000), Copyright Massachusetts Medical Society.

Receptors of the adaptive immune system are produced randomly and bind to antigens of all origins i.e. bacterial, environmental and self-antigens. However, receptors of the innate immune system bind specifically to PAMPs, and signal in the presence of infection. The adaptive immune system responds to a pathogen only after its initial recognition by the innate immune system. Thus, T cells of the adaptive immune system distinguish self from non-self via the innate immune

system. The recognition of the peptide–MHC ligand by antigen receptors is not sufficient to activate T cells as they cannot distinguish self from non-self; T cells require at least two signals to become activated: 1) the complex of a peptide and an MHC molecule; 2) a costimulatory signal mediated by, for example, the CD80 and CD86 molecules on the surface of the antigen-presenting cell (Figure 1.6). Recognition of an antigen in the absence of CD80 or CD86 molecules leads to permanent inactivation or apoptosis of the T cells. However, autoreactive T cells, which escape removal via thymic selection, may be activated by APCs leading to organ-specific autoimmune diseases, such as inflammation in the spine in AS (Gery & Egwuagu 2002).

The expression of CD80 and CD86 molecules on the surface of the antigen-presenting cell is controlled by the innate immune system via receptors, such as TLRs. Since PAMPs occur only on pathogens, TLRs induce CD80 and CD86 molecules only in the presence of infection. T cells receive both of the signals required for activation via receptors binding to the peptide that was derived from the pathogen as well as the CD80 or CD86 molecules induced by TLRs of the innate immunity (Medzhitov *et al.* 1997).

1.3.2. Immune/Inflammatory Responses

Inflammation is an adaptive response evolved for restoring homeostasis. It is clearly a key factor in the pathogenesis of AS. This is shown by histological evidence of inflammation at the involved sites in the spine and osteitis on MRI (Figure 1.2). In general terms, inflammation can be triggered by:

- 1) Infection leading to host defence against infection.

- 2) Tissue injury resulting in tissue repair response.
- 3) Tissue stress/malfunction leading to adaptation to stress and restoration of a homeostatic state.

A generic inflammatory pathway includes inducers, sensors, mediators and effectors (Figure 1.6(a)), where inducers of inflammation can be endogenous or exogenous (Figure 1.6(b)) (Medzhitov 2008).

In the case of bacterial infections, pathogenic PAMPs are detected by the pattern recognition receptors of the innate immune system such as toll-like receptors (TLRs) and NOD-like receptors (NLRs) on DCs. Macrophages and mast cells residing in the tissue mediate this innate recognition of the infection leading to production of inflammatory factors such as cytokines (e.g. IL18 and IL1 β) and chemokines, which in turn recruit neutrophils to the site of inflammation (Strowig *et al.* 2012).

(A) Inducers → Sensors → Mediators → Effectors

(B) Classification of inducers of inflammation

Inducers					
Endogenous				Exogenous	
ECM derived	Plasma derived	Tissue derived	Cell derived	Non-microbial	Microbial
- Signals released from stressed, malfunctioning or dead cells and damaged tissues - Endogenous crystals - Products of ECM breakdown				- Allergens - Irritants - Foreign bodies - Toxic compounds	- PAMPs - Virulence factors

Figure 1.6. (A) A generic inflammatory pathway. (B) Inducers of inflammation can be divided into two groups as exogenous and endogenous. Further classification of inducers is shown in the figure above. (ECM: Extra-cellular matrix; PAMPs: Pathogen-associated molecular patterns.) Adapted by permission from Macmillan Publishers Ltd: Nature (Medzhitov 2008), copyright (2008).

Infection-induced inflammation is relatively well understood compared to the systemic chronic inflammatory states such as AS. Inflammation due to infection and tissue stress/malfunction may have pathological consequences resulting in autoimmunity and inflammatory diseases (Medzhitov 2008). Amongst proposed mechanisms involved in the pathogenesis of AS is the unfolded protein response (UPR) causing ER stress (Colbert *et al.* 2010).

1.3.2.1. HLA-B27 & Pathogenesis of AS

1.3.2.1.1. Unfolded Protein Response

According to the proposed UPR mechanism, HLA-B27, which is strongly associated with AS, can misfold and form heavy chain homodimers (B27₂) due to formation of disulfide bonds by unpaired cysteine residues at position 67 (Bowness 2002). Cytokines upregulating the MHC class I expression and assembly pathway, such as interferons and TNF α , may trigger an unfolded protein response (UPR) in APCs of HLA-B27-expressing individuals when a critical threshold of misfolding is reached. *ERAP1* polymorphisms, which are also strongly associated with AS, may affect HLA-B27 folding/misfolding and its peptide repertoire via its peptide trimming function in the ER (Evans *et al.* 2011). The UPR may cause cells responding to innate pattern recognition receptor signals to produce more IL23 providing a stimulus for T_H17 survival and activation in individuals with permissive *IL23R* polymorphisms that are associated with AS (Colbert *et al.* 2010; Evans *et al.* 2011)

1.3.2.1.2. Other Proposed Mechanisms

Other proposed mechanisms of AS pathogenesis include (I) the arthritogenic peptide theory and (II) the HLA-B27-free heavy chain and homodimer hypothesis. According to the arthritogenic peptide theory, HLA-B27 presents an arthritogenic peptide, shared with disease-causing pathogens, leading to cross-recognition of a self antigen by cytotoxic T cells (Benjamin & Parham 1990). However, a self antigen, which could bind to HLA-B27 causing an auto-reactive cytotoxic T cell response, has not yet been identified in the joints. In the HLA-B27-free heavy chain and homodimer hypothesis, it has been proposed that cell surface

expression of abnormally folded HLA-B27 molecules leads to disease. The cell surface expression of the abnormal HLA-B27 heavy chains (monomeric or dimeric (B27₂)) may be triggered by conditions including infection or tissue stress. These HLA-B27-free heavy chains are then recognised by the innate immune receptors (e.g. KIRs) on cells such as NK cells and APCs, leading to a pro-inflammatory response (Allen *et al.* 1999). Cell surface homodimers, which may be due to dissociation of unstable cell surface trimeric (β_2m , HLA-B27 heavy chain and antigen) complexes, have been identified in HLA-B27-transfected cell lines and in peripheral blood mononuclear cells of AS patients providing evidence for the proposed disease mechanism (Kollnberger *et al.* 2002; Bird *et al.* 2003; Lan *et al.* 2004; Raine *et al.* 2006).

1.4. Genetic Epidemiology of Ankylosing Spondylitis

Studies of the genetic epidemiology of a disease range from demonstrating that the disease of interest has a heritable component, to determining the relative size of the genetic effect in relation to other sources of variation (e.g. environmental), to identifying the individual gene(s) responsible for disease risk. These aims can be achieved through:

- 1) Twin studies and studies of familial aggregation (to establish familiarity and determine heritability),
- 2) Genetic linkage studies in families (to identify chromosomal regions containing genetic variation implicated in disease),
- 3) Genetic association analysis and functional studies (to identify the genes and the functional variants that influence disease risk).

A major polygenic component to AS has been identified through a combination of twin studies, genetic linkage studies in affected sib pairs, and genetic association studies. These studies explained the increased familial recurrence of the disease as explained in the following sections.

1.4.1. Genetic Predisposition & Heritability

In the UK, 9.5% of the population is HLA-B27 positive, and > 85% of AS patients are HLA-B27 positive (Khan *et al.* 1977; Brown *et al.* 1996; Brown *et al.* 1997). There is a strong genetic predisposition to AS with clear familial aggregation of cases. The polygenic nature of the disease and the importance of other genetic factors besides HLA-B27 are highlighted by twin studies. The disparity between monozygotic twin concordance (63%) and dizygotic twins (13%) clearly indicates the importance of genetic factors in AS. This is only partly accounted for by the effect of HLA-B27, since dizygotic twins concordant for HLA-B27 carriage still exhibit a concordance rate of 23% for AS. Moreover, the broad sense heritability of AS has been estimated to be in excess of 97% and the HLA-B27 association is estimated to explain approximately 30% of the heritability of the disease (Brown *et al.* 1997). Furthermore, while more than 85% of AS cases carry the HLA-B27 allele, only 2-8% of HLA-B27 positive individuals in the general population develop AS. These observations show that the presence of HLA-B27 is almost a prerequisite for development of AS, but there are other genetic and environmental factors determining AS susceptibility (van der Linden & van der Heijde 1998; Sieper *et al.* 2006).

1.4.2. Major Histocompatibility Complex: Human Leukocyte Antigen

1.4.2.1. HLA-B27 in Ankylosing Spondylitis

HLA-B is an MHC class I gene expressed on almost all cell types. HLA-B27 is just one of the serotypes at this locus. There are more than 2,000 *HLA-B* alleles and in excess of 60 allelic variants of HLA-B27 alone at the DNA level (Khan 2010). In UK Caucasians, the major HLA-B27 subtypes associated with AS are *HLA-B*2705* and *HLA-B*2702* (Table 1.2) (Brown *et al.* 1996).

1.4.2.2. Prevalence of AS & HLA-B27 Subtypes

There is a strong correlation between the frequency of HLA-B27 in populations and the prevalence of AS. For instance, in Berlin the relatively high frequency of HLA-B27 (9.3%) is mirrored by a high prevalence of AS (0.55%) (Braun *et al.* 1998). Other examples of a high prevalence of AS reflecting relatively high frequencies of HLA-B27 in different populations include North American Indians (HLA-B27 18-50%), Norwegian Lapps (HLA-B27 25-30%), Alaskan Eskimos (HLA-B27 25-40%). In contrast, there is a very low prevalence of AS in most of sub-Saharan Africa (<1%), Australian aboriginals (absent) and South American Indians (absent), populations where the prevalence of HLA-B27 is low or even absent (Sieper *et al.* 2006).

Other genetic and/or environmental factors may also complicate the relationship between the frequency of HLA-B27 and the prevalence of AS. This is exemplified in West Africa where the prevalence of HLA-B27 is relatively high in the Gambia (~6%) but AS is extremely uncommon. In the Gambia, the main subtype of HLA-B27 was initially thought to be *HLA-B*2703*, not the European AS associated *HLA-*

*B*2705* subtype, and this was thought to account for the low prevalence of AS in this population (Hill *et al.* 1991). However, subsequent more detailed studies of this population in areas of the country known to be enriched for HLA-B27 revealed the *HLA-B*2705* subtype to be as common as *HLA-B*2703*. Despite this, AS is still extremely rare. Consequently, the difference in the prevalence of AS between the UK and the Gambia must reflect other differences, either genetic or environmental (Brown *et al.* 1997; van der Linden & van der Heijde 1998).

There are more than 60 subtypes of HLA-B27 (*HLA-B*2701-HLA-B*2763*), many of which are exceptionally rare, and not all are disease-associated. The HLA-B27 subtypes are generated by genetic events such as point mutations, gene conversion events and reciprocal recombination leading to amino acid changes and different peptide binding properties. Different HLA-B27 subtypes are associated with AS in different regions of the world. *HLA-B*2705* is the ancestral haplotype and the most widespread disease-associated subtype; *HLA-B*2702* and *HLA-B*2704* are disease-associated in Jews and Asians, respectively, but *HLA-B*2706* (in south East Asia) and *HLA-B*2709* (in Sardinia) are not disease-associated (Table 1.2) (Reveille & Maganti 2009; Khan 2010).

Population	*2701	*2702	*2703	*2704	*2705	*2706	*2707	*2708	*2709	*2711	*2713	*2714
Northern Europe		10			90							
Denmark		10			90							
UK		2			98							
Southern Europe		20			80							
Spain (Galicia)		18			80			3				
Northern Spain		7			91		1				1	
Italy		30			65		2		3			
Sardinia		3			77				20			
Greece		34	8		50		8					
Cyprus (Greeks)		52			32		17					
Turkey	7	30			43		14	5				
Lebanon		24	12		35		30					
Jewish		48		3	38		13					
Siberia		14			84							2
Northern India				33	61		6					
Western India				34	34		18	12				2
Japan				82	18							
Chinese		2		66	31		2					
Taiwan	0.05	0.5	3	87	4	7	2	0.02		0.02		
Taiwan (Han-Chinese)				94	6							
Taiwan (Aborigines)				100								
Chinese Indonesians				38		62						
Native Indonesians				6	6	89						
Thailand				42	5	53						
Brazil		10	6		80			3				<1
North Africa		50			50							
West Africans			32		68							
AS Associated Subtype	YES?	YES	YES?	YES	YES	NO?	YES	YES?	NO	?	YES?	YES?

Table 1.2. Examples of HLA-B27 subtypes (%), their geographic distributions and associations with AS (Brown *et al.* 1996; Sieper *et al.* 2006; Reveille & Maganti 2009).

1.4.3. Genetic Linkage Studies

Genetic linkage studies, based on the co-transmission of marker and disease loci that are physically close on chromosomes, proved to be a powerful tool for the localisation of mutant loci in monogenic diseases in the later part of the 20th century (e.g. hemochromatosis (Feder *et al.* 1996)). This approach was subsequently modified and applied to the study of complex diseases to identify loci contributing susceptibility to these polygenic disorders. For example, linkage analysis was applied to the study of type 1 diabetes mellitus both in the NOD mouse model (Cornall *et al.* 1991; Todd *et al.* 1991) and in humans (Davies *et al.* 1994). Similar approaches were applied to the collagen-induced model of rheumatoid arthritis (Remmers *et al.* 1996) and rheumatoid arthritis itself (MacKay *et al.* 2002; Eyre *et al.* 2004). Subsequently, these techniques were also applied to AS (See Section 1.4.3.1).

Genetic linkage studies examine the co-segregation of genetic markers and disease typically in large pedigrees of multiple affected individuals. Strategies for linkage analysis can take the form of genomewide screens where a few hundred genetic markers are typed across the genome, to studies involving the co-segregation of a single genetic marker. It may also be useful to focus on isolated populations. These population isolates are more genetically homogenous than outbred populations (minimizing genetic heterogeneity) and contain large pedigrees of multiple affected individuals. Examples of such isolates include the Amish and Ashkenazi Jews as well as selected populations from Finland, Iceland, Newfoundland, Quebec and Sardinia (Wright *et al.* 1999; Peltonen *et al.* 2000; Arcos-Burgos & Muenke 2002).

1.4.3.1. Genetic Linkage Studies in AS

A number of linkage studies have been carried out to identify regions linked to AS. These studies confirmed linkage with the MHC, and identified several non-MHC regions possibly linked to AS (Brown *et al.* 1998; Brown *et al.* 2000; Laval *et al.* 2001; Carter *et al.* 2007; Sims *et al.* 2008). These included a region on chromosome 16q (LOD=4.7) and a region on chromosome 2q, encompassing the interleukin 1 (IL1) cluster (LOD=2.5). A later association study suggested that an association with single nucleotide polymorphism (SNPs) near *TRADD* may account for the chromosome 16 linkage, but the causative SNP/gene has still to be unequivocally defined (Pointon *et al.* 2010b). The IL1 cluster was subsequently investigated in detail (Sims *et al.* 2008), but not convincingly replicated in later more comprehensive genomewide association studies (Burton *et al.* 2007; Reveille *et al.* 2010; Evans *et al.* 2011).

A major drawback of linkage studies is that the linked region is usually large (in the tens of centiMorgan range) and can encompass hundreds or thousands of genes, many of which might be functional candidates. This makes identification of the precise functional variant through positional cloning extremely difficult (although notable exceptions exist like in the case of *NOD2* and Crohn's disease (Ogura *et al.* 2001)). Additionally, linkage analysis of complex traits and diseases typically has low power and is often far less powerful than similar sized genetic association studies (Risch & Merikangas, 1996). It is perhaps not surprising then that many of the regions initially identified by linkage studies have subsequently failed to replicate and that the strategy as a whole has yielded little in the way of

robustly replicated genetic variants underlying complex disease (Altmuller *et al.* 2001).

1.5. Candidate Gene Association Studies

Genetic association studies have been applied to analyse countless candidate genes for evidence of association with many complex diseases, including AS.

1.5.1. The Basics of Genetic Association Studies

1.5.1.1. Single Nucleotide Polymorphisms

Genetic association studies typically use genetic markers known as single nucleotide polymorphisms (SNPs). A SNP is a DNA sequence variation that occurs when a nucleotide (Adenine - A, Thymine - T, Guanine - G, Cytosine - C) differs between individuals, or homologous chromosomes (Figure 1.7). There are about 53 million SNPs in the SNP database (dbSNP 137, genome build 37.3), about 22 million of which are found within or close to genes. Whole genomes or regions of interest are scanned for association signals by comparing genotype frequencies of SNPs in a study population of cases and ethnically matched healthy controls (HapMap 2003; Balding 2006). The genetic distance over which association signals can be detected depends on several variables, including the size of the study population, SNP allele frequencies and the extent of linkage disequilibrium in the region of interest.

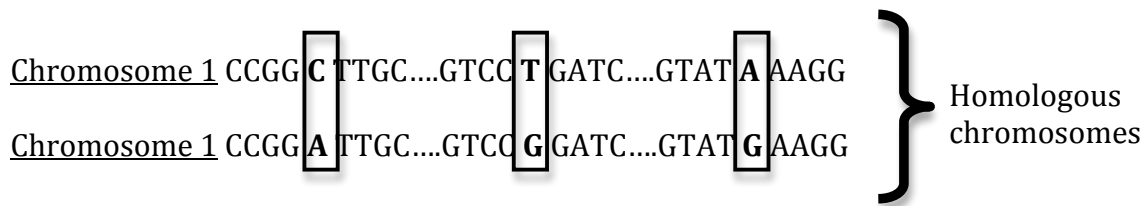


Figure 1.7. Three single nucleotide polymorphisms (SNPs; C/A, T/G, A/G in boxes) shown on a pairs of homologous chromosome (Chromosome 1). “....” represents flanking DNA sequences.

1.5.1.2. Linkage Disequilibrium

Linkage disequilibrium (LD) refers to non-random association (i.e. correlation) between alleles at different loci, in that they tend to be co-inherited together in a particular population. This measure is useful for selecting of tag SNPs (a strategy for determining the minimum number of SNPs required to scan a genetic region of interest for association signals by eliminating the highly correlated SNPs e.g. $r^2 > 0.8$) to be genotyped in an association study, as well as for comparing the power of different correlated variants to detect genetic association (Robbins 1918; Lewontin 1964; Balding 2006).

LD is influenced by a number of factors, such as physical distance, recombination frequency, mutation rate, natural selection, genetic drift, non-random mating, population structure, size and ancestry. The most important factor in determining LD between a disease and marker allele is the amount of historical recombination

that has occurred between them. Thus, loci that are physically close together typically will have had less ancestral recombination and therefore, exhibit stronger LD than those which are physically distant (Cardon & Abecasis 2003).

Two measures that are often used for quantifying LD are r^2 and D' . D is the deviation of the observed frequency of a haplotype (i.e combination of alleles) from the expected frequency under independent assortment (Figure 1.8). Since D depends on the allele frequencies, it is necessary to standardise it by calculating D' (If $D < 0$, $D' = D/D_{\min}$; if $D > 0$, $D' = D/D_{\max}$, where $D_{\min}$ and $D_{\max}$ are minimum and maximum values calculated for deviation, respectively). r^2 is the most common LD measure employed in genetic association studies and is mathematically equivalent to the correlation between two variants. r^2 is calculated using the formula:

$$r = D / (p_1 p_2 q_1 q_2)^{1/2}$$

where D = Deviation in the 2-SNP (AB) haplotype frequency

p_1 = SNP (A_1) major allele frequency

p_2 = SNP (B_1) major allele frequency

q_1 = SNP (A_2) minor allele frequency

q_2 = SNP (B_2) minor allele frequency

Two-locus model for calculation of deviation

Haplotype	Frequency		Allele	Frequency
A_1B_1	x_{11}		A_1	$p_1 = x_{11} + x_{12}$
A_1B_2	x_{12}		A_2	$p_2 = x_{21} + x_{22}$
A_2B_1	x_{21}		B_1	$q_1 = x_{11} + x_{21}$
A_2B_2	x_{22}		B_2	$q_2 = x_{12} + x_{22}$

Hence the observed frequency of haplotype $A_1B_1 \rightarrow x_{11}$

The expected frequency of $A_1B_1 \rightarrow x_{11} = p_1q_1$
(probability of two *independent* events happening together i.e. no LD)

Observed haplotype frequency - Expected haplotype frequency = Deviation (D)
 $D = x_{11} - p_1q_1$

Relationship between the haplotype frequencies, allele frequencies and D.

Formation of the 2x2 table:

$$D = x_{11} - p_1q_1 \rightarrow x_{11} = p_1q_1 + D$$

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
B_2	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	1

Number of times a given haplotype is observed and the allele frequencies are known.

If the number of times that a given haplotype is observed deviates from what is expected from the allele frequencies assuming two alleles are independent, then the alleles are in LD.

Figure 1.8. Calculation of D explained. D is the deviation of the observed frequency of a haplotype (i.e combination of alleles) from the expected frequency under independent assortment.

An understanding of the LD structure within the genetic region of interest is essential for the design and the interpretation of corresponding genetic association studies. This is because even though LD typically decays as a function of physical distance between two loci, the extent and distribution of LD is highly variable. In some genomic regions, two variants may be in LD even if they are 50 to 500 kb apart (Cardon & Abecasis 2003). This may be particularly extreme in the MHC.

1.5.2. Candidate Gene Association Studies of Complex Diseases

Most association studies compare unrelated cases (individuals with the disease of interest), and unrelated ethnically matched controls (healthy individuals with similar genetic ancestry as the case population) in a case-control study. The same SNPs are genotyped in the cases and controls, and allele, genotype or haplotype frequencies are compared. Statistically significant differences in frequencies between the two groups indicate an association with the disease which may be due to direct involvement of the marker studied, or linkage disequilibrium (LD) with the disease marker (Figure 1.9).

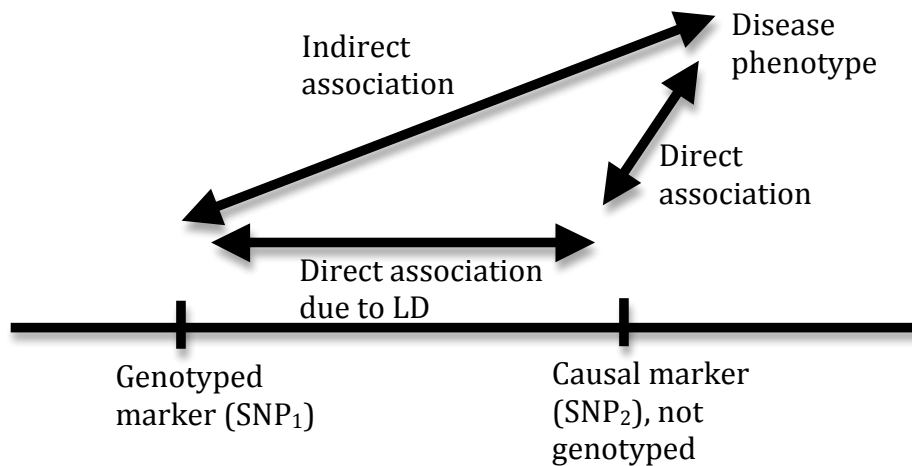


Figure 1.9. Strategy for genetic association studies involving the use of LD patterns to detect disease associated loci. The typed marker SNP₁ can be used to infer association at an untyped causal locus (SNP₂). Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Balding 2006), copyright (2006).

Candidate gene association studies are hypothesis-driven, looking at the relationship between genetic variants and disease risk in biologically and functionally plausible genes, as well as genes in regions of genetic linkage or genomewide association studies. Biologically plausible candidates may include genes associated with related diseases like IBD and psoriasis in the case of AS. Precisely, what constitutes a candidate gene is a subject of debate with particular implications for applying correction factors for levels of significance. Large study populations are required to reveal true associations and achieve the necessary statistical power to identify and confirm genetic effects that may be very small. An independent replication of observed associations is essential to eliminate type I errors (false positives), genotyping artefacts and confirm the initial results (Balding 2006).

There are several pitfalls associated with genetic case-control studies that can be reduced by careful experimental design. For example, spurious associations produced as a consequence of population substructure (leading to different allele frequencies and disease prevalence) can be minimized by ensuring that cases and controls are matched as closely as possible in terms of their ancestry (i.e. the controls should be comparable to the cases except for the presence of the disease). In addition, there now exists statistical methodology to detect and correct for the effect of population substructure in genetic studies of unrelated individuals including genomic control (Devlin & Roeder 1999), structured association (Pritchard *et al.* 2000), principal components analysis (Price *et al.* 2006), and linear mixed models (Zhou & Stephens 2012).

1.5.3. MHC Region: *HLA-B* Associations

1.5.3.1. HLA-B27 & AS

The association between HLA-B27 and AS was first described 39 years ago (Brewerton *et al.* 1973; Schlosstein *et al.* 1973). A number of genetic studies including twin studies and candidate gene association studies have confirmed this association, as discussed previously in this chapter (Brown *et al.* 1996; Brown *et al.* 1997). In the study by Brown *et al.*, 94% of AS cases and 9.5% of healthy controls were HLA-B27 positive giving an odds ratio of 171 (95% CI 135-218, $p < 10^{-99}$) (Brown *et al.* 1996).

1.5.3.2. HLA-B60 & AS

HLA-B60 has been identified as an independent risk factor for AS, initially in HLA-B27 positive individuals, but subsequently also in those patients who were HLA-B27 negative (Robinson *et al.* 1989; Brown *et al.* 1996; Brown *et al.* 1997; Brown *et*

al. 1998). In the study by Brown *et al.*, 12% of AS cases and 6.2% of healthy controls were HLA-B60 positive giving an odds ratio of 3.6 (95% CI 2.1 – 6.3, $p < 1 \times 10^{-5}$) in HLA-B27 positive cases and controls. The observed HLA-B60 association was independent of the HLA-B27 association as residual association was observed in that region after controlling for HLA-B27 (Brown *et al.* 1996). Moreover, an HLA-B60 association was also seen in Taiwanese HLA-B27 negative AS patients (Wei *et al.* 2004).

1.5.4. Other Genes with a Role in Related Inflammatory Diseases that are Candidates for AS

Some polygenic disorders tend to recur more commonly than expected by chance in individuals or their relatives. This is particularly true for individuals with AS, IBD and psoriasis, suggesting that these diseases may share common genetic susceptibility factors. Associations detected in any one of these diseases may therefore be plausible candidates for the others. For instance, *IL23R* and *TNFRSF1A* are both associated with IBD, PS and AS (see Chapters 3 and 4) (Karaderi *et al.* 2009; Karaderi *et al.* 2012). Other obvious candidates that have been studied include pattern recognition receptors (PRRs), essential components of the innate immune system, such as *NOD2* (*CARD15*), which is strongly associated with IBD (Hugot *et al.* 2001; Ogura *et al.* 2001). An association with *NOD2* was confirmed in cases with IBD, who also had AS, but no independent association was found in those with AS without IBD (Crane *et al.* 2002). An association between AS and a similar protein, *CARD9*, which also functions as a PRR and is associated with IBD, has been identified by a study in the UK (Pointon *et al.* 2010b). Other genes known to be associated with Crohn's disease (CD) were tested for association with AS in another study. Positive associations were identified with *STAT3* and 1q32 in this

study with odds ratios 0.84 (95% CI 0.77-0.91, $p < 2.6 \times 10^{-5}$) and 0.74 (95% CI 0.68-0.82, $p < 1.6 \times 10^{-10}$), respectively, confirming the genetic overlap between CD and AS (Danoy *et al.* 2010).

1.6. The Era of Genomewide Association Studies

Genomewide association studies (GWASs) are hypothesis-free genetic association studies where typically hundreds of thousands of markers spanning the genome are tested for association with the disease of interest. These studies have been facilitated by:

- (1) The publication of International Hapmap, which documented the extent and distribution of LD across the genome (HapMap 2003; HapMap 2005; Frazer *et al.* 2007).
- (2) The development of genotyping technologies, which made it not only technologically possible but economically attractive to genotype hundreds of thousands of genetic polymorphisms in thousands of individuals.
- (3) The concurrent availability of large patient and control collections ensuring adequate statistical power to detect common variants of small effect. Indeed the collaboration of research groups forming international consortia for sample collection, data generation and analysis has considerably facilitated the discovery of common genetic variants through the sharing of association summary statistics and subsequent meta-analysis.

1.6.1. Genomewide Association Studies

The “Common disease, common variant” hypothesis underlies the application of agnostic (hypothesis-free) GWAS, which aim to detect genetic associations between common variants and common diseases. Such common polygenic diseases are generally considered evolutionarily neutral, because it is expected that many common variants/genes are involved in influencing the disease risk, each of which is relatively unlikely to have a significant effect on fitness to reproduce overall. Thus, susceptibility variants may have frequencies > 5% in a given population with small additive or multiplicative effects on the disease risk. Therefore, at least some of the genetic influences on many common diseases are probably attributable to a limited number of common allelic variants that may be identified by GWAS (Carlson *et al.* 2004).

Since there is no *a priori* hypothesis and the whole genome is scanned for association signals, much more stringent controls for possible statistical errors are required. These standards were developed, and thresholds were set in the first large Wellcome Trust Case-Control Consortium study (WTCCC1 2007). The first GWAS in AS (and multiple sclerosis, auto-immune thyroid disease and breast cancer) genotyped only 14,500 non-synonymous coding SNPs and was itself, in part, an experiment to establish whether such a gene-targeted approach was more productive in identifying genetic associations than the truly genomewide approach used in a number of other common diseases (Burton *et al.* 2007; WTCCC1 2007). Subsequent studies have used different platforms capable of either genotyping larger numbers of SNPs, and/or targeting specific SNPs. For example, the latest WTCCC2-TASC GWAS genotyped ~2000 AS cases for >600,000 SNPs and controls

for 1.2 million SNPs (Evans *et al.* 2011). Subsequent replication studies have used a customised “ImmunoChip” for genotyping 200,000 SNPs specifically targeting genes potentially involved in a range of inflammatory/immunological diseases (Cortes & Brown 2011).

1.6.2. The International HapMap Project

The aim of the International HapMap Project was to catalog common genetic variants (MAF>5%) in humans. The project started in 2002 in order to develop a human haplotype map. Haplotypes are genomic regions of linked alleles inherited together that can be used for precise LD mapping, which facilitates genetic association studies. Such LD information is useful in the selection of SNPs for genotyping (See Chapter 2, Section 2.4.2 for tag SNP selection), as well as in the explanation of any observed associations. Researchers from Canada, China, Japan, Nigeria, the UK and the USA participated in this multi-national project. SNPs were detected by random shot-gun sequencing of whole-genomes and whole-chromosomes. The human haplotype map was produced by genotyping these SNPs in 270 samples from 4 populations with geographically diverse ancestry, including individuals from Yoruba in Nigeria (YRI), the Centre d'Etude du Polymorphisme Humain (CEPH) (Utah residents of Northern and Western European ancestry - CEU), Beijing in China (Han Chinese - CHB), and Tokyo in Japan (Japanese - JPT) (HapMap 2003; HapMap 2005).

The project had three separate phases:

- **Phase I:** Approximately 1 million SNPs with a minimum spacing of 5 kilobases were genotyped and their LD patterns were characterised (HapMap 2005).

- **Phase II:** An additional 2.1 million SNPs were genotyped, characterising 25-35% of the 9-10 million common SNP variation in humans (Frazer *et al.* 2007).
- **Phase III:** Additional samples from the same populations and additional populations were included in the project to allow better selection of tag SNPs for genetic association studies and more detailed analysis of diverse populations. 1.6 million common SNPs were genotyped in 1184 individuals from 11 populations and ten 100-kilobase regions were sequenced in 692 of these in order to have more data on lower frequency variants (frequency<5%) as well (Altshuler *et al.* 2010).

1.6.3. Limitations & Issues Related to GWAS

GWAS are designed to identify common disease-associated variants (minor allele frequency>5%) with modest effect sizes ($1.1 < \text{odds ratio} < 1.5$). This strategy has been spectacularly successful given that before 2007 only a handful of common genetic variants had been robustly associated with complex diseases, whereas there are now well over 1500 genetic variants reliably associated with these diseases directly as a result of the GWAS approach (Visscher *et al.* 2012). Although this represents an amazing advance in knowledge, for the majority of complex traits and diseases, the genetic variants that have been discovered explain a only a small proportion of the overall heritability (Maher 2008). For example, in Crohn's disease, 32 identified loci account for only 20% of the disease heritability (Manolio *et al.* 2009). It is still unclear precisely what the functional variants are that may explain this missing heritability. Some possibilities include (1) gene-gene interaction, (2) structural variation and (3) low frequency variation

(frequency<5%) not tagged by SNPs on GWAS genotyping platforms. New statistical methodology developed by Peter Visscher's group indicates that for many complex traits and diseases, around half of the missing heritability is present in the form of common variants of very small effect scattered across the genome (Yang *et al.* 2010). While the effect sizes of these SNPs are too small to be detected by standard GWAS, their combined variance can be estimated through the use of linear mixed models (Zhou & Stephens 2012).

Almost all of the disease-associated variants identified in GWAS are common SNPs (MAF>5%) that correlate with the disease and affect the risk of disease. However, this correlation does not always indicate a functional association. The functional variants that underlie susceptibility to common diseases may include (but are not limited to) (1) non-synonymous coding variants, (2) variants in transcription factor binding sites in promoter regions, or in other regulatory protein binding sites in 3' untranslated regions, (3) intronic variants affecting mRNA splicing, and (4) variants in inter-genic regions of the DNA and gene deserts encoding non-protein coding RNA species, such as microRNAs that control gene expression. In the recent findings of the Encyclopedia of DNA Elements (ENCODE) project, it was estimated that most of the disease-associated common SNPs (MAF>5%) in the GWAS are found within non-coding functional elements such as transcription binding sites (12%; 31% including the variants in strong LD) and DNA regulatory sites (e.g. enhancer binding sites; 34%; 71% including variants in LD) (Bernstein *et al.* 2012). Furthermore, the DNA regions containing functional elements are generally highly conserved as a result of selection; deleterious variants within them are therefore likely to be of lower frequency (Genomes 2010). Thus, in

addition to the common variants found in GWAS, it is important to analyse disease associations with low frequency variants.

Population stratification and multiple testing are the main source of false positive results (type I errors) in GWAS. Population stratification is controlled by careful choice of study populations, matching cases and controls for ethnicity. The inclusion of ancestry informative marker SNPs (AIMS) on genotyping platforms in large studies controls for population stratification at the analysis stage (Burton *et al.* 2007). Even though a stringent genomewide level of significance ($p\text{-value} < 5 \times 10^{-8}$) is applied to GWAS to account for multiple testing, false positive results still arise making robust replication studies essential (Balding 2006).

Replication of association in a different ethnicity can be particularly valuable since it can be informative about the causal SNP, as well as the disease aetiology. For instance, AS is associated with the *IL23R* non-synonymous coding variant rs11209026 in Caucasian populations, but this SNP is absent from Chinese populations, which correspondingly show no association with *IL23R*. This observation provides support for rs11209026 as a causal coding variant (see Chapter 3) (Davidson *et al.* 2009; Karaderi *et al.* 2009).

1.6.4. Main Findings of GWAS in AS

1.6.4.1. WTCCC1 2007

The Wellcome Trust Case-Control Consortium (WTCCC1 2007) study was the first large GWAS in common complex diseases; it was designed to assess 7 complex diseases. The study included 2000 cases from each disease and 3000 controls. This study set important standards and described methods for future GWA studies.

For instance, the genomewide p-value threshold of 5×10^{-7} was determined by the WTCCC 2007 study. This genomewide p-value threshold was later adjusted to 5×10^{-8} . Furthermore, the use of a common control group of healthy individuals to test for association with all 7 diseases, and the utility of combining unrelated disease cases with healthy controls to achieve a higher statistical power was also demonstrated by this study. This method was used in Chapter 6 of this thesis, in which osteoarthritis cases were included in the control group. The use of genomic control (λ) (Devlin & Roeder 1999) and principle components analysis (Price *et al.* 2006) for control of population stratification at the analysis stage was also illustrated (See Chapter 2, section 2.8.1.2 for the explanation of λ) (WTCCC1 2007).

1.6.4.2. WTCCC-TASC 2007

The WTCCC-TASC (Triple “A” (Anglo/Australian/American) Spondyloarthritis Consortium) GWAS was the first genomewide study in AS, and studied 1000 AS cases and 1500 healthy controls. The major findings from this study included the identification of associations with variants within *ERAP1* and *IL23R*, as well as confirming the *HLA-B* association. The *ERAP1* and *IL23R* associations were subsequently independently replicated (see Chapter 3 for the replication of the *IL23R* association) (Harvey *et al.* 2009; Karaderi *et al.* 2009). Only non-synonymous coding SNPs (which lead to changes in the amino acid sequence of proteins) were included in the study (14,436 NS SNPs and 897 MHC SNPs) because it was thought that these were more likely to be causally associated with disease. The study did not examine non-coding SNPs, which may include variants found in promoters, intronic regions and gene deserts (Burton *et al.* 2007).

1.6.4.3. TASC 2010

The TASC GWAS (examining 2053 cases and 5140 healthy controls in the discovery set) was the first comprehensive genome-wide association study to examine both coding and non-coding SNPs in AS (Reveille *et al.* 2010). Previously identified associations between variants within *IL23R*, *ERAP1*, and the MHC were replicated in this study, highlighting the importance of the T_H17 and IL1 cytokine pathways in AS susceptibility. Genetic variants in “gene deserts” on chromosomes 2p15 and 21q22 also showed genomewide significance ($p < 5 \times 10^{-8}$), whereas variants within *ANTXR2* and *IL1R2*, showed suggestive association ($p < 10^{-6}$).

1.6.4.4. WTCCC2 & TASC Meta-analysis 2011

The most recent GWAS of AS was a combined meta-analysis of the WTCCC2 and TASC (2010) studies, comprising a discovery set of 3,023 cases and 8,779 controls (Evans *et al.* 2011). Genome-wide significant association was demonstrated between variants within the genes *RUNX3*, *LTBR-TNFRSF1A*, and *IL12B*. Strong evidence of association was observed at loci containing the genes *PTGER4*, *TBKBP1*, *ANTXR2* and *CARD9*. An important finding of the GWAS was the genetic interaction between *ERAP1* and HLA-B27. This observation supported the role of ERAP1 in AS susceptibility as an aminopeptidase trimming peptides for HLA class I presentation (See Chapter 5). This study also employed a SNP, rs4349859, which could be used to tag HLA-B27 with high sensitivity and specificity, and has been used for this purpose in this thesis.

1.6.5. Associated Genes Common to AS & Other Related Inflammatory Diseases

GWAS and candidate gene association studies have been very effective in identifying potential causal variants in AS as well as other related inflammatory diseases such as IBD and psoriasis. For instance, identification of *IL23R* and *IL12B* provided strong evidence for a common aetiology between AS, IBD, psoriasis, and the T_H17 pathway (Burton *et al.* 2007) (Figure 1.10). Such discoveries highlighted the potential of new drug targets, such as IL12p40 and IL17 in these inflammatory diseases (Maddur *et al.* 2012; Zhu & Qian 2012). There are now several examples of the results from GWAS leading to "drug repositioning"- whereby a drug which was originally developed for another related disease, is tried in another disease on the basis of the results of a genome-wide association study (Sanseau *et al.* 2012). The anti-IL17 drug secukinumab (which was originally developed for the treatment of multiple sclerosis and rheumatoid arthritis), found promise in the treatment of psoriasis. Following the identification of the *IL23R* association with AS, which implicated the T_H17 pathway, secukinumab was repositioned for use in the treatment of AS, where it shows considerable promise for the treatment of this condition (Baraliakos & Braun 2012).

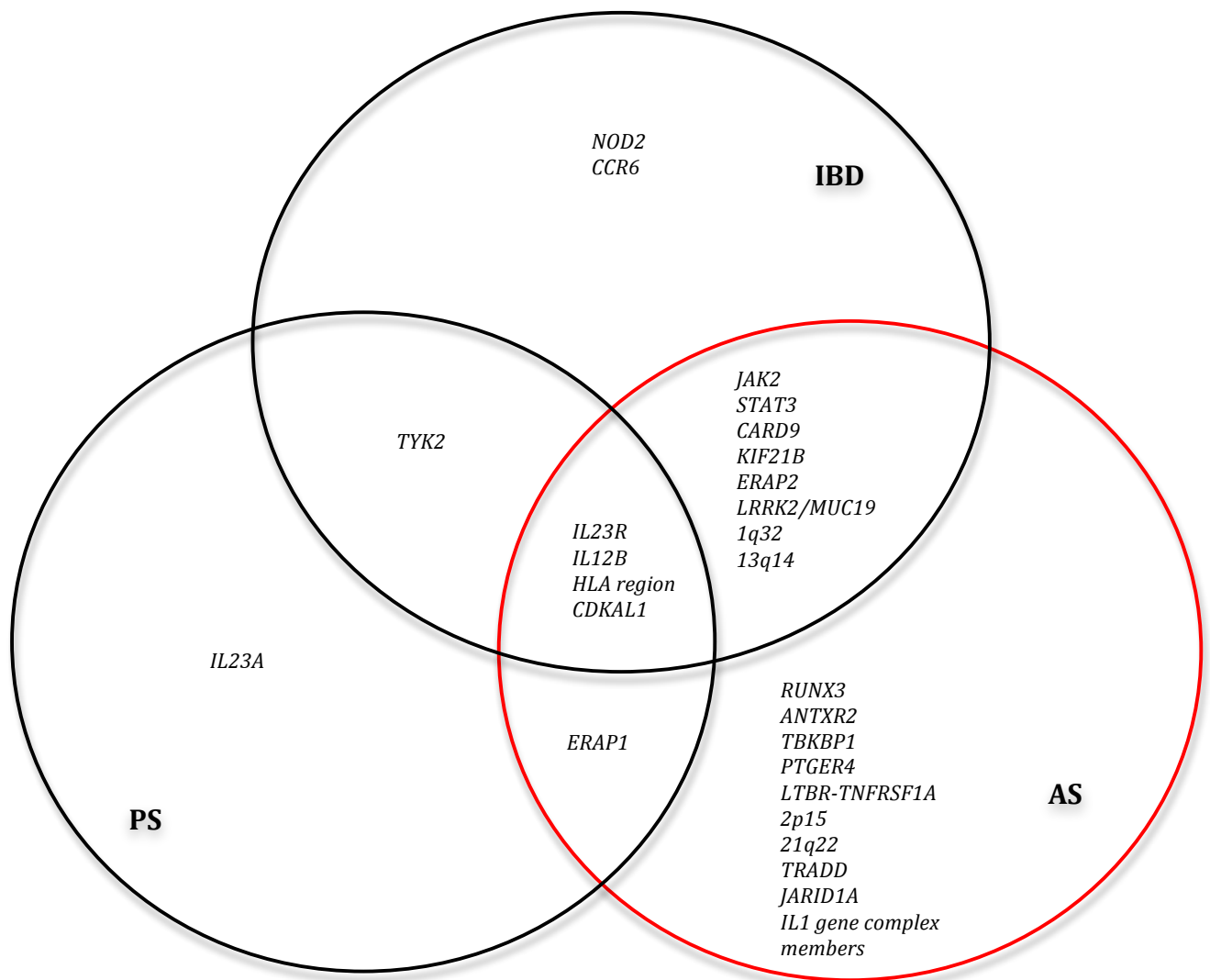


Figure 1.10. Major genetic association findings in AS, IBD and psoriasis showing the genetic overlap between these diseases (Barrett *et al.* 2009; Danoy *et al.* 2010; Franke *et al.* 2010; Strange *et al.* 2010; Anderson *et al.* 2011; Evans *et al.* 2011; Reveille 2011).

1.7. The Use of Imputation & Large-Scale Sequencing

1.7.1. A Brief Explanation of Imputation: *In silico* Genotyping

Imputation is a statistical method of inferring missing genotypes in a genetic association study (Figure 1.11). Reference haplotypes are constructed using a densely genotyped dataset often obtained from a large population based project, such as the International HapMap Project or the 1000 Genomes Consortium (HapMap 2005; Li *et al.* 2009; Genomes 2010). Advantages of imputation include fast genotyping with a much lower cost for fine mapping of genetic regions of interest. However, the use of appropriate reference haplotypes is crucial for precise inference of the genotypes. Factors determining the quality of these reference haplotypes include (1) the size of the reference population, (2) whether haplotypes are inferred from unrelated or family (trio) data, (3) the quality of the technology used for genotyping the reference population, and (4) the density at which the SNPs have been genotyped. As the size of the reference population increases, estimation of haplotypes is more likely to be accurate, and rarer haplotypes are more likely to be represented in the reference set. Increasing the density of SNPs in the reference population allows finer scale genetic variation to be imputed. Furthermore, the software used for imputation is another determining factor for accuracy of imputed genotypes. Widely used imputation software include MACH, IMPUTE, PLINK and BEAGLE (Biernacka *et al.* 2009; Huang *et al.* 2009; Li *et al.* 2009).

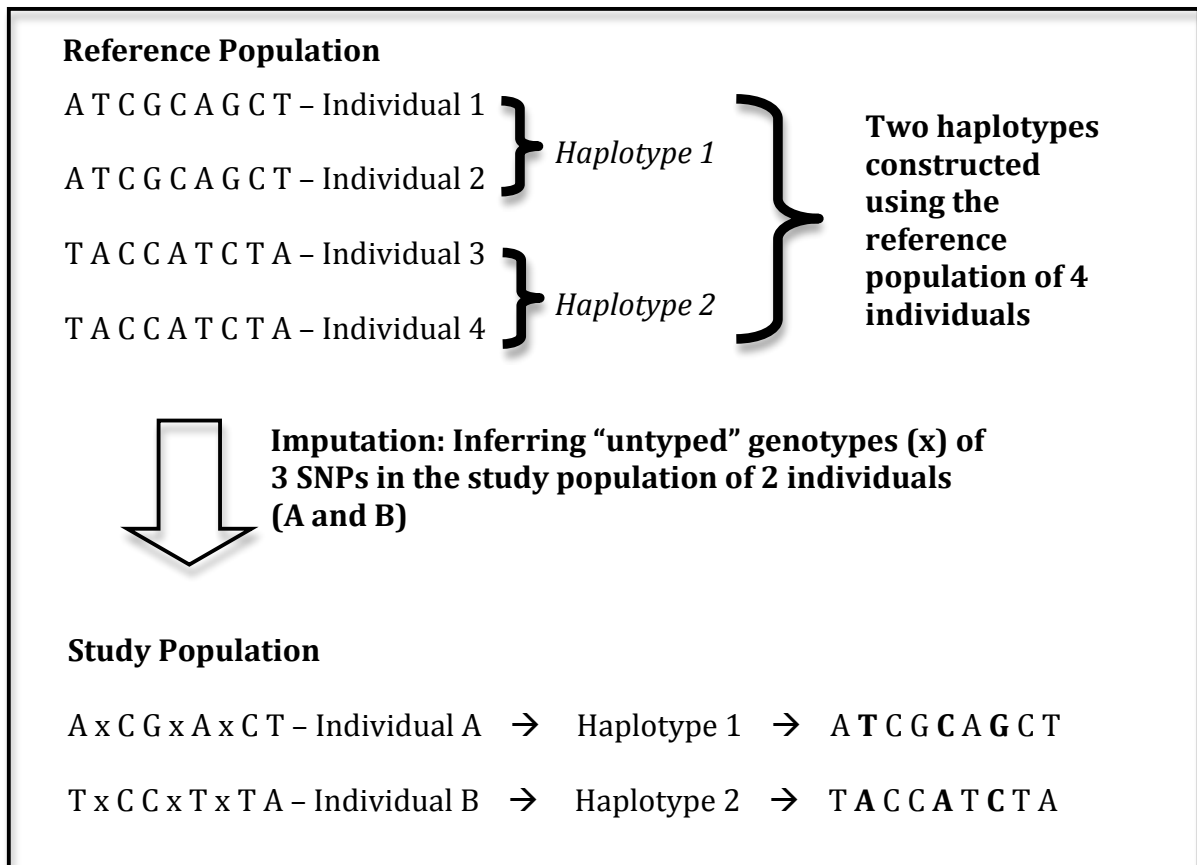


Figure 1.11. Imputation: Haplotypes from a reference population is used for statistically inferring missing genotypes in a study population. Two haplotypes are constructed using the data on genotypes of 9 SNPs from 4 individuals in the reference population. Later, these reference haplotypes are used to estimate the untyped 3 SNPs (x) in the study population of 2 individuals. The number of individuals, SNPs and haplotypes are decreased to simplify the figure.

1.7.1.1. Imputation in the Sequencing Era

1.7.1.1.1. The 1000 Genomes Project

The 1000 Genomes Project started in January 2008 with the collaboration of international research groups to sequence the genomes of a 1000 individuals from different ethnicities (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas) and provide a more complete catalogue of genetic variation within the human genome. The aim of the project was to identify >95% of variants in the accessible genome that have allele frequencies $\geq 1\%$. The first pilot stage of the project was completed in 2009. Data from the pilot study were made public for use in scientific studies, as well as providing a foundation for the future stages of the project. In the pilot study, low coverage whole genome sequencing data was produced on 61 CEU individuals (August 2009 pilot data: 122 haplotypes, number of SNPs: 9,232,128). This dataset was used as the reference set in the imputation performed in this thesis (See Chapter 5) (Meucci *et al.* 2005; Genomes 2010).

Further data were also released later in 2010 with more complete sequence information. The most recent release contains haplotype data from 1092 individuals with genotype data on about 38.3 million SNPs (data release 2010-11). The main advantage of using the 1000 Genomes Project data for imputation is that it is far denser than the HapMap, contains data on lower frequency variants, and thus, it complements GWA studies by making it possible to impute these variants to test for association with complex traits (See Chapter 2 and 5 for further details of application of this method).

1.8. The Aims of the Thesis

In this thesis, I systematically tested the reported association of two genetic loci, *IL23R* and *TNFRSF1A* that were identified in previous GWAS (Burton *et al.* 2007; Reveille *et al.* 2010). I have genotyped SNPs in these genes in a large sample of AS cases and controls in an attempt to validate these associations and to try to identify potential primary causal genetic variants. I have also investigated whether some of the missing heritability in AS exists as common variants that were not present in the International HapMap Project data, which were consequently not imputed in previous GWAS, and also whether some of the missing heritability might be explained by low frequency variants, which again was not imputed in previous GWAS. Specifically, I have imputed genomewide SNP data from the WTCCC2, the most recent GWAS study of AS, using the 1000 Genomes Project data. I then performed single locus tests of association and rare variant collapsing tests (Asimit & Zeggini 2009; Asimit & Zeggini 2010; Lawrence *et al.* 2010; Morris & Zeggini 2010; Magi *et al.* 2011; Neale *et al.* 2011) in an attempt to identify novel common and low frequency variants that may have been missed by previous GWAS. Ten selected SNPs were then genotyped in an independent cohort to replicate the observed associations. Reliability of the imputation was also analysed in a group of cases to estimate the accuracy of imputed genotypes for common and low frequency SNPs.

1.9. References

- Allen, R. L., O'Callaghan, C. A., *et al.* (1999). "Cutting edge: HLA-B27 can form a novel beta 2-microglobulin-free heavy chain homodimer structure." *J Immunol* **162**(9): 5045-5048.
- Altmuller, J., Palmer, L. J., *et al.* (2001). "Genomewide scans of complex human diseases: true linkage is hard to find." *Am J Hum Genet* **69**(5): 936-950.
- Altshuler, D. M., Gibbs, R. A., *et al.* (2010). "Integrating common and rare genetic variation in diverse human populations." *Nature* **467**(7311): 52-58.
- Anderson, C. A., Boucher, G., *et al.* (2011). "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47." *Nat Genet* **43**(3): 246-252.
- Andrianakos, A., Trontzas, P., *et al.* (2003). "Prevalence of rheumatic diseases in Greece: a cross-sectional population based epidemiological study. The ESORDIG Study." *J Rheumatol* **30**(7): 1589-1601.
- Annunziato, F., Cosmi, L., *et al.* (2012). "Defining the human T helper 17 cell phenotype." *Trends Immunol.*
- Apanius, V., Penn, D., *et al.* (1997). "The nature of selection on the major histocompatibility complex." *Crit Rev Immunol* **17**(2): 179-224.
- Appel, H., Kuhne, M., *et al.* (2006a). "Immunohistologic analysis of zygapophyseal joints in patients with ankylosing spondylitis." *Arthritis Rheum* **54**(9): 2845-2851.
- Appel, H., Kuhne, M., *et al.* (2006b). "Immunohistochemical analysis of hip arthritis in ankylosing spondylitis: evaluation of the bone-cartilage interface and subchondral bone marrow." *Arthritis Rheum* **54**(6): 1805-1813.
- Appel, H., Loddenkemper, C., *et al.* (2006c). "Correlation of histopathological findings and magnetic resonance imaging in the spine of patients with ankylosing spondylitis." *Arthritis Res Ther* **8**(5): R143.
- Arcos-Burgos, M. & Muenke, M. (2002). "Genetics of population isolates." *Clin Genet* **61**(4): 233-247.
- Asimit, J. & Zeggini, E. (2009). "Testing for rare variant associations in complex diseases." *Genome Med* **1**(11): 24.
- Asimit, J. & Zeggini, E. (2010). "Rare variant association analysis methods for complex traits." *Annu Rev Genet* **44**: 293-308.
- Balding, D. J. (2006). "A tutorial on statistical methods for population association studies." *Nat Rev Genet* **7**(10): 781-791.
- Baraliakos, X. & Braun, J. (2012). "Biologic therapies for spondyloarthritis: what is new?" *Curr Rheumatol Rep* **14**(5): 422-427.
- Barrett, J. C., Lee, J. C., *et al.* (2009). "Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region." *Nat Genet* **41**(12): 1330-1334.
- Benjamin, R. & Parham, P. (1990). "Guilt by association: HLA-B27 and ankylosing spondylitis." *Immunol Today* **11**(4): 137-142.
- Bernstein, B. E., Birney, E., *et al.* (2012). "An integrated encyclopedia of DNA elements in the human genome." *Nature* **489**(7414): 57-74.
- Biernacka, J. M., Tang, R., *et al.* (2009). "Assessment of genotype imputation methods." *BMC Proc* **3 Suppl 7**: S5.
- Bird, L. A., Peh, C. A., *et al.* (2003). "Lymphoblastoid cells express HLA-B27 homodimers both intracellularly and at the cell surface following endosomal recycling." *Eur J Immunol* **33**(3): 748-759.

- Boniface, K., Blom, B., *et al.* (2008). "From interleukin-23 to T-helper 17 cells: human T-helper cell differentiation revisited." Immunol Rev **226**: 132-146.
- Boonen, A. & Severens, J. L. (2002). "Ankylosing spondylitis: what is the cost to society, and can it be reduced?" Best Pract Res Clin Rheumatol **16**(4): 691-705.
- Boonen, A., van der Heijde, D., *et al.* (2006). "Markov model into the cost-utility over five years of etanercept and infliximab compared with usual care in patients with active ankylosing spondylitis." Ann Rheum Dis **65**(2): 201-208.
- Bowness, P. (2002). "HLA B27 in health and disease: a double-edged sword?" Rheumatology (Oxford) **41**(8): 857-868.
- Boyer, G. S., Templin, D. W., *et al.* (1994). "Prevalence of spondyloarthropathies in Alaskan Eskimos." J Rheumatol **21**(12): 2292-2297.
- Brain, O., Cooney, R., *et al.* (2012). "Functional consequences of mutations in the autophagy genes in the pathogenesis of Crohn's disease." Inflamm Bowel Dis **18**(4): 778-781.
- Braun, J., Baraliakos, X., *et al.* (2003). "Magnetic resonance imaging examinations of the spine in patients with ankylosing spondylitis, before and after successful therapy with infliximab: evaluation of a new scoring system." Arthritis Rheum **48**(4): 1126-1136.
- Braun, J., Baraliakos, X., *et al.* (2012). "Golimumab reduces spinal inflammation in ankylosing spondylitis: MRI results of the randomised, placebo- controlled GO-RAISE study." Ann Rheum Dis **71**(6): 878-884.
- Braun, J., Bollow, M., *et al.* (1998). "Prevalence of spondylarthropathies in HLA-B27 positive and negative blood donors." Arthritis Rheum **41**(1): 58-67.
- Braun, J. & Sieper, J. (2007). "Ankylosing spondylitis." Lancet **369**(9570): 1379-1390.
- Brewerton, D. A., Hart, F. D., *et al.* (1973). "Ankylosing spondylitis and HL-A 27." Lancet **1**(7809): 904-907.
- Brown, M. A., Edwards, S., *et al.* (2000). "Polymorphisms of the CYP2D6 gene increase susceptibility to ankylosing spondylitis." Hum Mol Genet **9**(11): 1563-1566.
- Brown, M. A., Jepson, A., *et al.* (1997). "Ankylosing spondylitis in West Africans--evidence for a non-HLA-B27 protective effect." Ann Rheum Dis **56**(1): 68-70.
- Brown, M. A., Kennedy, L. G., *et al.* (1998). "The effect of HLA-DR genes on susceptibility to and severity of ankylosing spondylitis." Arthritis Rheum **41**(3): 460-465.
- Brown, M. A., Kennedy, L. G., *et al.* (1997). "Susceptibility to ankylosing spondylitis in twins: the role of genes, HLA, and the environment." Arthritis Rheum **40**(10): 1823-1828.
- Brown, M. A., Pile, K. D., *et al.* (1996). "HLA class I associations of ankylosing spondylitis in the white population in the United Kingdom." Ann Rheum Dis **55**(4): 268-270.
- Brown, M. A., Pile, K. D., *et al.* (1998). "A genome-wide screen for susceptibility loci in ankylosing spondylitis." Arthritis Rheum **41**(4): 588-595.
- Burmester, G. R., Panaccione, R., *et al.* (2012). "Adalimumab: long-term safety in 23 458 patients from global clinical trials in rheumatoid arthritis, juvenile idiopathic arthritis, ankylosing spondylitis, psoriatic arthritis, psoriasis and Crohn's disease." Ann Rheum Dis.

- Burton, P. R., Clayton, D. G., *et al.* (2007). "Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants." Nat Genet **39**(11): 1329-1337.
- Calin, A. (1985). "Ankylosing spondylitis." Clin Rheum Dis **11**(1): 41-60.
- Calin, A., Brophy, S., *et al.* (1999). "Impact of sex on inheritance of ankylosing spondylitis: a cohort study." Lancet **354**(9191): 1687-1690.
- Cardon, L. R. & Abecasis, G. R. (2003). "Using haplotype blocks to map human complex trait loci." Trends Genet **19**(3): 135-140.
- Careless, D. J. & Inman, R. D. (1995). "Acute anterior uveitis: clinical and experimental aspects." Semin Arthritis Rheum **24**(6): 432-441.
- Carlson, C. S., Eberle, M. A., *et al.* (2004). "Mapping complex disease loci in whole-genome association studies." Nature **429**(6990): 446-452.
- Carter, K. W., Pluzhnikov, A., *et al.* (2007). "Combined analysis of three whole genome linkage scans for Ankylosing Spondylitis." Rheumatology (Oxford) **46**(5): 763-771.
- Cheung, P. P., Paternotte, S., *et al.* (2012). "Performance of the assessment in spondyloarthritis international society classification for axial and peripheral spondyloarthritis in an established clinical cohort: comparison with criteria sets of amor and the European spondylarthropathy study group." J Rheumatol **39**(4): 816-821.
- Colbert, R. A., DeLay, M. L., *et al.* (2010). "From HLA-B27 to spondyloarthritis: a journey through the ER." Immunol Rev **233**(1): 181-202.
- Cornall, R. J., Prins, J. B., *et al.* (1991). "Type 1 diabetes in mice is linked to the interleukin-1 receptor and Lsh/Ity/Bcg genes on chromosome 1." Nature **353**(6341): 262-265.
- Cortes, A. & Brown, M. A. (2011). "Promise and pitfalls of the Immunochip." Arthritis Res Ther **13**(1): 101.
- Crane, A. M., Bradbury, L., *et al.* (2002). "Role of NOD2 variants in spondylarthritis." Arthritis Rheum **46**(6): 1629-1633.
- Dagfinrud, H., Kvien, T. K., *et al.* (2008). "Physiotherapy interventions for ankylosing spondylitis." Cochrane Database Syst Rev(1): CD002822.
- Dai, S. M., Han, X. H., *et al.* (2003). "Prevalence of rheumatic symptoms, rheumatoid arthritis, ankylosing spondylitis, and gout in Shanghai, China: a COPCORD study." J Rheumatol **30**(10): 2245-2251.
- Danoy, P., Pryce, K., *et al.* (2010). "Association of variants at 1q32 and STAT3 with ankylosing spondylitis suggests genetic overlap with Crohn's disease." PLoS Genet **6**(12): e1001195.
- Davidson, S. I., Wu, X., *et al.* (2009). "Association of ERAP1, but not IL23R, with ankylosing spondylitis in a Han Chinese population." Arthritis Rheum **60**(11): 3263-3268.
- Davies, J. L., Kawaguchi, Y., *et al.* (1994). "A genome-wide search for human type 1 diabetes susceptibility genes." Nature **371**(6493): 130-136.
- Davis, J. C., Jr., Van Der Heijde, D., *et al.* (2003). "Recombinant human tumor necrosis factor receptor (etanercept) for treating ankylosing spondylitis: a randomized, controlled trial." Arthritis Rheum **48**(11): 3230-3236.
- Devlin, B. & Roeder, K. (1999). "Genomic control for association studies." Biometrics **55**(4): 997-1004.
- Diaz-Pena, R., Vidal-Castineira, J. R., *et al.* (2010). "Association of the KIR3DS1*013 and KIR3DL1*004 alleles with susceptibility to ankylosing spondylitis." Arthritis Rheum **62**(4): 1000-1006.

- Ebringer, A. (1989). "The relationship between Klebsiella infection and ankylosing spondylitis." *Baillieres Clin Rheumatol* **3**(2): 321-338.
- El Maghraoui, A. (2011). "Extra-articular manifestations of ankylosing spondylitis: prevalence, characteristics and therapeutic implications." *Eur J Intern Med* **22**(6): 554-560.
- Elvin, J., Cerundolo, V., *et al.* (1991). "A quantitative assay of peptide-dependent class I assembly." *Eur J Immunol* **21**(9): 2025-2031.
- English, L., Chemali, M., *et al.* (2009). "Autophagy enhances the presentation of endogenous viral antigens on MHC class I molecules during HSV-1 infection." *Nat Immunol* **10**(5): 480-487.
- Evans, D. M., Spencer, C. C., *et al.* (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**(8): 761-767.
- Eyre, S., Barton, A., *et al.* (2004). "Investigation of susceptibility loci identified in the UK rheumatoid arthritis whole-genome scan in a further series of 217 UK affected sibling pairs." *Arthritis Rheum* **50**(3): 729-735.
- Feder, J. N., Gnirke, A., *et al.* (1996). "A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis." *Nat Genet* **13**(4): 399-408.
- Feldtkeller, E., Khan, M. A., *et al.* (2003). "Age at disease onset and diagnosis delay in HLA-B27 negative vs. positive patients with ankylosing spondylitis." *Rheumatol Int* **23**(2): 61-66.
- Franke, A., McGovern, D. P., *et al.* (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." *Nat Genet* **42**(12): 1118-1125.
- Frazer, K. A., Ballinger, D. G., *et al.* (2007). "A second generation human haplotype map of over 3.1 million SNPs." *Nature* **449**(7164): 851-861.
- Furst, D. E., Keystone, E. C., *et al.* (2012). "Updated consensus statement on biological agents for the treatment of rheumatic diseases, 2011." *Ann Rheum Dis* **71 Suppl 2**: i2-45.
- Genomes (2010). "A map of human genome variation from population-scale sequencing." *Nature* **467**(7319): 1061-1073.
- Gery, I. & Egwuagu, C. E. (2002). "Central tolerance mechanisms in control of susceptibility to autoimmune uveitic disease." *Int Rev Immunol* **21**(2-3): 89-100.
- Goie The, H. S., Steven, M. M., *et al.* (1985). "Evaluation of diagnostic criteria for ankylosing spondylitis: a comparison of the Rome, New York and modified New York criteria in patients with a positive clinical history screening test for ankylosing spondylitis." *Br J Rheumatol* **24**(3): 242-249.
- Gottlieb, A., Menter, A., *et al.* (2009). "Ustekinumab, a human interleukin 12/23 monoclonal antibody, for psoriatic arthritis: randomised, double-blind, placebo-controlled, crossover trial." *Lancet* **373**(9664): 633-640.
- Hampe, J., Franke, A., *et al.* (2007). "A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1." *Nat Genet* **39**(2): 207-211.
- HapMap (2003). "The International HapMap Project." *Nature* **426**(6968): 789-796.
- HapMap (2005). "A haplotype map of the human genome." *Nature* **437**(7063): 1299-1320.
- Harvey, D., Pointon, J. J., *et al.* (2009). "Investigating the genetic association between ERAP1 and ankylosing spondylitis." *Hum Mol Genet* **18**(21): 4204-4212.

- Harvey, D., Pointon, J. J., *et al.* (2009). "Analysis of killer immunoglobulin-like receptor genes in ankylosing spondylitis." Ann Rheum Dis **68**(4): 595-598.
- Hill, A. V., Allsopp, C. E., *et al.* (1991). "HLA class I typing by PCR: HLA-B27 and an African B27 subtype." Lancet **337**(8742): 640-642.
- Hill, A. V., Allsopp, C. E., *et al.* (1991). "Common west African HLA antigens are associated with protection from severe malaria." Nature **352**(6336): 595-600.
- Hill, A. V., Yates, S. N., *et al.* (1994). "Human leukocyte antigens and natural selection by malaria." Philos Trans R Soc Lond B Biol Sci **346**(1317): 379-385.
- Huang, L., Li, Y., *et al.* (2009). "Genotype-imputation accuracy across worldwide human populations." Am J Hum Genet **84**(2): 235-250.
- Hugot, J. P., Chamaillard, M., *et al.* (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.
- Iwasaki, A. & Medzhitov, R. (2004). "Toll-like receptor control of the adaptive immune responses." Nat Immunol **5**(10): 987-995.
- Iwasaki, A. & Medzhitov, R. (2010). "Regulation of adaptive immunity by the innate immune system." Science **327**(5963): 291-295.
- Karaderi, T., Harvey, D., *et al.* (2009). "Association between the interleukin 23 receptor and ankylosing spondylitis is confirmed by a new UK case-control study and meta-analysis of published series." Rheumatology (Oxford) **48**(4): 386-389.
- Karaderi, T., Pointon, J. J., *et al.* (2012). "Evidence of genetic association between TNFRSF1A encoding the p55 tumour necrosis factor receptor, and ankylosing spondylitis in UK Caucasians." Clin Exp Rheumatol **30**(1): 110-113.
- Khan, M. A. (2010). "Remarkable polymorphism of HLA-B27: an ongoing saga." Curr Rheumatol Rep **12**(5): 337-341.
- Khan, M. A., Braun, W. E., *et al.* (1977). "HLA B27 in ankylosing spondylitis: differences in frequency and relative risk in American Blacks and Caucasians." J Rheumatol Suppl **3**: 39-43.
- Kidd, P. (2003). "Th1/Th2 balance: the hypothesis, its limitations, and implications for health and disease." Altern Med Rev **8**(3): 223-246.
- Kimura, A. & Kishimoto, T. (2010). "IL-6: regulator of Treg/Th17 balance." Eur J Immunol **40**(7): 1830-1835.
- Kobelt, G., Andlin-Sobocki, P., *et al.* (2004). "The burden of ankylosing spondylitis and the cost-effectiveness of treatment with infliximab (Remicade)." Rheumatology (Oxford) **43**(9): 1158-1166.
- Kollnberger, S., Bird, L., *et al.* (2002). "Cell-surface expression and immune receptor recognition of HLA-B27 homodimers." Arthritis Rheum **46**(11): 2972-2982.
- Kurts, C., Robinson, B. W., *et al.* (2010). "Cross-priming in health and disease." Nat Rev Immunol **10**(6): 403-414.
- Lan, C. C., Tsai, W. C., *et al.* (2004). "Psoriatic patients with arthropathy show significant expression of free HLA class I heavy chains on circulating monocytes: a potential role in the pathogenesis of psoriatic arthropathy." Br J Dermatol **151**(1): 24-31.
- Lapaquette, P., Brest, P., *et al.* (2012). "Etiology of Crohn's disease: many roads lead to autophagy." J Mol Med (Berl).

- Laval, S. H., Timms, A., *et al.* (2001). "Whole-genome screening in ankylosing spondylitis: evidence of non-MHC genetic-susceptibility loci." Am J Hum Genet **68**(4): 918-926.
- Lawrence, R., Day-Williams, A. G., *et al.* (2010). "CCRaVAT and QuTie-enabling analysis of rare variants in large-scale case control and quantitative trait association studies." BMC Bioinformatics **11**: 527.
- Leung, S., Liu, X., *et al.* (2010). "The cytokine milieu in the interplay of pathogenic Th1/Th17 cells and regulatory T cells in autoimmune disease." Cell Mol Immunol **7**(3): 182-189.
- Levine, B. & Kroemer, G. (2008). "Autophagy in the pathogenesis of disease." Cell **132**(1): 27-42.
- Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models." Genetics **49**(1): 49-67.
- Li, Y., Willer, C., *et al.* (2009). "Genotype imputation." Annu Rev Genomics Hum Genet **10**: 387-406.
- Lopez-Larrea, C., Blanco-Gelaz, M. A., *et al.* (2006). "Contribution of KIR3DL1/3DS1 to ankylosing spondylitis in human leukocyte antigen-B27 Caucasian populations." Arthritis Res Ther **8**(4): R101.
- MacKay, K., Eyre, S., *et al.* (2002). "Whole-genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 affected sibling pairs in the United Kingdom." Arthritis Rheum **46**(3): 632-639.
- Maddur, M. S., Miossec, P., *et al.* (2012). "Th17 cells: biology, pathogenesis of autoimmune and inflammatory diseases, and therapeutic strategies." Am J Pathol **181**(1): 8-18.
- Magi, R., Kumar, A., *et al.* (2011). "Assessing the impact of missing genotype data in rare variant association analysis." BMC Proc **5 Suppl 9**: S107.
- Maher, B. (2008). "Personal genomes: The case of the missing heritability." Nature **456**(7218): 18-21.
- Manolio, T. A., Collins, F. S., *et al.* (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.
- Marshak-Rothstein, A. & Rifkin, I. R. (2007). "Immunologically active autoantigens: the role of toll-like receptors in the development of chronic inflammatory disease." Annu Rev Immunol **25**: 419-441.
- Martinon, F., Mayor, A., *et al.* (2009). "The inflammasomes: guardians of the body." Annu Rev Immunol **27**: 229-265.
- McCappin, J., Harvey, D., *et al.* (2010). "No association of KIR3DL1 or KIR3DS1 or their alleles with ankylosing spondylitis." Tissue Antigens **75**(1): 68-73.
- McCarroll, S. A., Huett, A., *et al.* (2008). "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease." Nat Genet **40**(9): 1107-1112.
- Medzhitov, R. (2008). "Origin and physiological roles of inflammation." Nature **454**(7203): 428-435.
- Medzhitov, R. & Janeway, C., Jr. (2000). "Innate immunity." N Engl J Med **343**(5): 338-344.
- Medzhitov, R., Preston-Hurlburt, P., *et al.* (1997). "A human homologue of the Drosophila Toll protein signals activation of adaptive immunity." Nature **388**(6640): 394-397.
- Meucci, M. A., Marsh, S., *et al.* (2005). "CEPH individuals are representative of the European American population: implications for pharmacogenetics." Pharmacogenomics **6**(1): 59-63.

- Moll, J. M. & Wright, V. (1973). "New York clinical criteria for ankylosing spondylitis. A statistical evaluation." Ann Rheum Dis **32**(4): 354-363.
- Morris, A. P. & Zeggini, E. (2010). "An evaluation of statistical approaches to rare variant analysis in genetic association studies." Genet Epidemiol **34**(2): 188-193.
- Mosmann, T. R. & Coffman, R. L. (1989). "TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties." Annu Rev Immunol **7**: 145-173.
- Neale, B. M., Rivas, M. A., *et al.* (2011). "Testing for an unusual distribution of rare variants." PLoS Genet **7**(3): e1001322.
- Neefjes, J., Jongsma, M. L., *et al.* (2011). "Towards a systems understanding of MHC class I and MHC class II antigen presentation." Nat Rev Immunol **11**(12): 823-836.
- Ogura, Y., Bonen, D. K., *et al.* (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature **411**(6837): 603-606.
- Oukka, M. (2007). "Interplay between pathogenic Th17 and regulatory T cells." Ann Rheum Dis **66 Suppl 3**: iii87-90.
- Palm, N. W. & Medzhitov, R. (2009). "Pattern recognition receptors and control of adaptive immunity." Immunol Rev **227**(1): 221-233.
- Parcej, D. & Tampe, R. (2010). "ABC proteins in antigen translocation and viral inhibition." Nat Chem Biol **6**(8): 572-580.
- Peltonen, L., Palotie, A., *et al.* (2000). "Use of population isolates for mapping complex traits." Nat Rev Genet **1**(3): 182-190.
- Pillai, M. R., Collison, L. W., *et al.* (2011). "The plasticity of regulatory T cell function." J Immunol **187**(10): 4987-4997.
- Plenge, R. M., Padyukov, L., *et al.* (2005). "Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4." Am J Hum Genet **77**(6): 1044-1060.
- Pointon, J. J., Chapman, K., *et al.* (2008). "Toll-like receptor 4 and CD14 polymorphisms in ankylosing spondylitis: evidence of a weak association in Finns." J Rheumatol **35**(8): 1609-1612.
- Pointon, J. J., Harvey, D., *et al.* (2010a). "Elucidating the chromosome 9 association with AS; CARD9 is a candidate gene." Genes Immun **11**(6): 490-496.
- Pointon, J. J., Harvey, D., *et al.* (2010b). "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**(6): 1243-1246.
- Potts, W. K. & Slev, P. R. (1995). "Pathogen-based models favoring MHC genetic diversity." Immunol Rev **143**: 181-197.
- Price, A. L., Patterson, N. J., *et al.* (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet **38**(8): 904-909.
- Pritchard, J. K., Stephens, M., *et al.* (2000). "Association mapping in structured populations." Am J Hum Genet **67**(1): 170-181.
- Procko, E. & Gaudet, R. (2009). "Antigen processing and presentation: TAPping into ABC transporters." Curr Opin Immunol **21**(1): 84-91.
- Raine, T., Brown, D., *et al.* (2006). "Consistent patterns of expression of HLA class I free heavy chains in healthy individuals and raised expression in spondyloarthropathy patients point to physiological and pathological roles." Rheumatology (Oxford) **45**(11): 1338-1344.

- Remmers, E. F., Longman, R. E., *et al.* (1996). "A genome scan localizes five non-MHC loci controlling collagen-induced arthritis in rats." Nat Genet **14**(1): 82-85.
- Reveille, J. D. (2011). "The genetic basis of spondyloarthritis." Ann Rheum Dis **70 Suppl 1**: i44-50.
- Reveille, J. D. & Maganti, R. M. (2009). "Subtypes of HLA-B27: history and implications in the pathogenesis of ankylosing spondylitis." Adv Exp Med Biol **649**: 159-176.
- Reveille, J. D., Sims, A. M., *et al.* (2010). "Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci." Nat Genet **42**(2): 123-127.
- Rioux, J. D., Xavier, R. J., *et al.* (2007). "Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis." Nat Genet **39**(5): 596-604.
- Robbins, R. B. (1918). "Some Applications of Mathematics to Breeding Problems III." Genetics **3**(4): 375-389.
- Robinson, M. J., Osorio, F., *et al.* (2009). "Dectin-2 is a Syk-coupled pattern recognition receptor crucial for Th17 responses to fungal infection." J Exp Med **206**(9): 2037-2051.
- Robinson, W. P., van der Linden, S. M., *et al.* (1989). "HLA-Bw60 increases susceptibility to ankylosing spondylitis in HLA-B27+ patients." Arthritis Rheum **32**(9): 1135-1141.
- Romagnani, S. (1994). "Lymphokine production by human T cells in disease states." Annu Rev Immunol **12**: 227-257.
- Romagnani, S. (1996). "Th1 and Th2 in human diseases." Clin Immunol Immunopathol **80**(3 Pt 1): 225-235.
- Sambrook, P. N. & Geusens, P. (2012). "The epidemiology of osteoporosis and fractures in ankylosing spondylitis." Ther Adv Musculoskelet Dis **4**(4): 287-292.
- Sanseau, P., Agarwal, P., *et al.* (2012). "Use of genome-wide association studies for drug repositioning." Nat Biotechnol **30**(4): 317-320.
- Saveanu, L., Carroll, O., *et al.* (2005). "Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum." Nat Immunol **6**(7): 689-697.
- Schlosstein, L., Terasaki, P. I., *et al.* (1973). "High association of an HL-A antigen, W27, with ankylosing spondylitis." N Engl J Med **288**(14): 704-706.
- Sieper, J., Rudwaleit, M., *et al.* (2006). "Concepts and epidemiology of spondyloarthritis." Best Pract Res Clin Rheumatol **20**(3): 401-417.
- Sims, A. M., Barnardo, M., *et al.* (2007). "Non-B27 MHC associations of ankylosing spondylitis." Genes Immun **8**(2): 115-123.
- Sims, A. M., Timms, A. E., *et al.* (2008). "Prospective meta-analysis of interleukin 1 gene complex polymorphisms confirms associations with ankylosing spondylitis." Ann Rheum Dis **67**(9): 1305-1309.
- Snelgrove, T., Lim, S., *et al.* (2007). "Association of toll-like receptor 4 variants and ankylosing spondylitis: a case-control study." J Rheumatol **34**(2): 368-370.
- Stetson, D. B. (2009). "Connections between antiviral defense and autoimmunity." Curr Opin Immunol **21**(3): 244-250.
- Strange, A., Capon, F., *et al.* (2010). "A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1." Nat Genet **42**(11): 985-990.

- Strowig, T., Henao-Mejia, J., *et al.* (2012). "Inflammasomes in health and disease." Nature **481**(7381): 278-286.
- Thomson, G. T., DeRubeis, D. A., *et al.* (1995). "Post-Salmonella reactive arthritis: late clinical sequelae in a point source cohort." Am J Med **98**(1): 13-21.
- Thorsby, E. (2009). "A short history of HLA." Tissue Antigens **74**(2): 101-116.
- Todd, J. A., Aitman, T. J., *et al.* (1991). "Genetic analysis of autoimmune type 1 diabetes mellitus in mice." Nature **351**(6327): 542-547.
- Uhl, M., Kepp, O., *et al.* (2009). "Autophagy within the antigen donor cell facilitates efficient antigen cross-priming of virus-specific CD8+ T cells." Cell Death Differ **16**(7): 991-1005.
- van Dartel, S. A., Fransen, J., *et al.* (2012). "Difference in the risk of serious infections in patients with rheumatoid arthritis treated with adalimumab, infliximab and etanercept: results from the Dutch Rheumatoid Arthritis Monitoring (DREAM) registry." Ann Rheum Dis.
- van der Heijde, D., Landewe, R., *et al.* (2008). "Radiographic progression of ankylosing spondylitis after up to two years of treatment with etanercept." Arthritis Rheum **58**(5): 1324-1331.
- van der Heijde, D., Salonen, D., *et al.* (2009). "Assessment of radiographic progression in the spines of patients with ankylosing spondylitis treated with adalimumab for up to 2 years." Arthritis Res Ther **11**(4): R127.
- van der Linden, S., Valkenburg, H. A., *et al.* (1984). "Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria." Arthritis Rheum **27**(4): 361-368.
- van der Linden, S. & van der Heijde, D. (1998). "Ankylosing spondylitis: evidence for a non-HLA-B*27 protective effect." Ann Rheum Dis **57**(4): 263-264.
- van der Linden, S. M., Valkenburg, H. A., *et al.* (1984). "The risk of developing ankylosing spondylitis in HLA-B27 positive individuals. A comparison of relatives of spondylitis patients with the general population." Arthritis Rheum **27**(3): 241-249.
- Vander Cruyssen, B., Munoz-Gomariz, E., *et al.* (2010). "Hip involvement in ankylosing spondylitis: epidemiology and risk factors associated with hip replacement surgery." Rheumatology (Oxford) **49**(1): 73-81.
- Visscher, P. M., Brown, M. A., *et al.* (2012). "Five years of GWAS discovery." Am J Hum Genet **90**(1): 7-24.
- Vyas, J. M., Van der Veen, A. G., *et al.* (2008). "The known unknowns of antigen processing and presentation." Nat Rev Immunol **8**(8): 607-618.
- Wahl, S. M., Vazquez, N., *et al.* (2004). "Regulatory T cells and transcription factors: gatekeepers in allergic inflammation." Curr Opin Immunol **16**(6): 768-774.
- Wanders, A., Heijde, D., *et al.* (2005). "Nonsteroidal antiinflammatory drugs reduce radiographic progression in patients with ankylosing spondylitis: a randomized clinical trial." Arthritis Rheum **52**(6): 1756-1765.
- Wang, R. X., Yu, C. R., *et al.* (2012). "Novel IL27p28/IL12p40 cytokine suppressed experimental autoimmune uveitis (EAU) by inhibiting autoreactive Th1/Th17 cells and promoting expansion of regulatory T cells." J Biol Chem.
- Ward, M. M. (2002). "Functional disability predicts total costs in patients with ankylosing spondylitis." Arthritis Rheum **46**(1): 223-231.
- Wei, J. C., Tsai, W. C., *et al.* (2004). "HLA-B60 and B61 are strongly associated with ankylosing spondylitis in HLA-B27-negative Taiwan Chinese patients." Rheumatology (Oxford) **43**(7): 839-842.

- Wordsworth, B. P., Pearcy, M. J., *et al.* (1984). "In-patient regime for the treatment of ankylosing spondylitis: an appraisal of improvement in spinal mobility and the effects of corticotrophin." Br J Rheumatol **23**(1): 39-43.
- Wright, A. F., Carothers, A. D., *et al.* (1999). "Population choice in mapping genes for complex diseases." Nat Genet **23**(4): 397-404.
- WTCCC1 (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.
- Yamane, H. & Paul, W. E. (2012). "Memory CD4+ T cells: fate determination, positive feedback and plasticity." Cell Mol Life Sci **69**(10): 1577-1583.
- Yang, J., Benyamin, B., *et al.* (2010). "Common SNPs explain a large proportion of the heritability for human height." Nat Genet **42**(7): 565-569.
- Zhang, W., Ferguson, J., *et al.* (2012). "Effector CD4+ T cell expression signatures and immune-mediated disease associated genes." PLoS One **7**(6): e38510.
- Zhernakova, A., Festen, E. M., *et al.* (2008). "Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP." Am J Hum Genet **82**(5): 1202-1210.
- Zhou, X. & Stephens, M. (2012). "Genome-wide efficient mixed-model analysis for association studies." Nat Genet **44**(7): 821-824.
- Zhu, S. & Qian, Y. (2012). "IL-17/IL-17 receptor system in autoimmune disease: mechanisms and therapeutic potential." Clin Sci (Lond) **122**(11): 487-511.

CHAPTER 2

2. GENERAL METHODS

2.1. Study Populations

All patients with AS in this study were either members of the National Ankylosing Spondylitis Society (UK), attendees at the Nuffield Orthopaedic Centre (Oxford, UK), the Royal National Hospital for Rheumatic Disease (Bath, UK), or referrals from rheumatologists in the UK. All patients fulfilled the modified New York criteria for the diagnosis of AS (van der Linden *et al.* 1984) and were of British Caucasian origin (with grandparents originating from the UK or Ireland).

Ethnically matched healthy control subjects were blood donors recruited from the National Blood Service (Oxford, UK), or patients with osteoarthritis and their spouses. The study was approved by the research ethics committee board in the UK (MREC project number 98/5/23). All participants gave informed consent prior to enrolment and completed a comprehensive questionnaire which included data such as age of onset of AS, assessment of disease activity (Bath AS Disease activity index-BASDAI), disease severity (Bath AS functional index –BASFI) and the co-occurrence of psoriasis and inflammatory bowel disease (Calin *et al.* 1994; Garrett *et al.* 1994). The precise composition of the case and control groups varied between studies and is described in the relevant chapters.

2.2. DNA Extraction from Blood & Saliva

2.2.1. DNA Extraction from Blood

DNA was extracted from peripheral venous blood. Blood was collected in BD EDTA Vacutainer tubes (BD, USA) and stored at -20°C. Blood (10 mL) was mixed with cold lysis buffer (25 mL) (Section 2.9, reagents) and incubated for 10 minutes with intermittent mixing at room temperature. The mixture was centrifuged at 1000 x g in a Beckman Allegra centrifuge at 4°C for 10 minutes. The supernatant was discarded and the pellet was completely re-suspended in 20 mL of lysis buffer by vortexing. After centrifuging for 10-minute at 1000 x g, the pellet was re-suspended in 500 µL of sodium perchlorate (5M) and 2 mL of solution B were added (Section 2.9, reagents). The mixture was incubated at 65°C for one hour with intermittent mixing, cooled to room temperature, and extracted with 2 mL of chloroform by vigorous shaking to mix the layers. The mixture was centrifuged at 2000 x g for 5 minutes and the upper aqueous phase was separated and collected in a 15-mL tube. DNA was precipitated by the addition of 7 mL of cold 100% ethanol and mixed by gentle inversion. The precipitated DNA was spooled out, air dried and re-suspended in 500 µL of 1 x Tris-EDTA (TE) buffer (Section 2.9, reagents). The DNA was quantified using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, USA). DNA concentration was calculated using the conversion that double-stranded DNA with a concentration of 50 µg/mL has an absorbance of 1 OD in a 1 cm light path.

2.2.2. DNA Extraction from Saliva

Approximately 50% of the DNA samples used in this study were prepared from saliva samples collected using Oragene collection kits (DNA Genotek Inc., Ontario, Canada), according to manufacturer's instructions. Briefly, saliva was collected at least half an hour after eating, and 5 minutes after a water mouth wash and mixed with the preservative within 30 minutes to prevent the degradation of DNA. The samples were incubated at 50°C overnight and transferred to 15ml tubes. One-twenty-fifth of the sample volume of Oragene DNA purifier reagent was added and mixed by vortexing, incubated on ice for 10 minutes and centrifuged at 3500 g for 10 minutes. The clear supernatant was transferred to a new tube, DNA was allowed to precipitated for 10 minutes by the addition of an equal volume of 95-100% ethanol at room temperature. DNA was collected by centrifugation at 3500g at room temperature for 10 minutes and washed with 70% ethanol before dissolving in 500µL TE buffer over 2 days. DNA samples were stored at -20°C (short term) of -80°C (long term).

2.3. Polymerase Chain Reaction (PCR) & DNA Electrophoresis

2.3.1. Primer Design, Preparation & PCR

DNA primers for amplification of target regions were designed using genomic DNA sequence, available from genome browsers (Section 2.10, online resources).

Primer3, an online tool for primer design, was also employed to design primers of 18- 25 base pairs (Rozen & Skaletsky 2000). Primer 3 targets sequences with a GC content of the 50-60% with no internal palindromes, and selects non-

complementary primer pairs. The melting temperatures for a primer pair should be within 5°C of each other. Primers were checked for specificity in order to prevent the formation of non-specific PCR products, and for repetitive sequences using the Basic Local Alignment Search Tool (BLASTn) (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>).

Primers for exon amplification and sequencing (Chapter 4, table 4.1 for primer sequences) were generally designed to obtain PCR products of 400-500 base pairs, the optimum length for the 36-cm electrophoresis capillaries used in the in-house 3100 Genetic Analyser (Applied Biosystems, Warrington, UK) for DNA sequencing. Larger target regions were amplified with overlapping PCR products. Primers were positioned ~100 base pairs up- and down-stream of each exon to cleanly sequence the exons and the intron-exon boundaries.

The primers were synthesised and de-salted by reverse phase chromatography (cartridge) by Eurofins (Eurofins MWG Operon, Ebersberg, Germany). Primers were diluted to a stock solution of 100 µM using dH₂O according to the manufacturer's instructions and further diluted to working concentrations of 10 and 20 µM for amplification and sequencing PCRs respectively. Optimisation of the conditions for each PCR amplification reaction was carried out by testing a range of annealing temperatures (50°C-68°C) and magnesium chloride concentrations (MgCl₂; 1.0-3.0 mM) following a standard PCR protocol for templates up to about 500 base pairs (Table 2.1).

Standard PCR Protocol			
Temperature	Time	No. of Cycles	Stage
94°C	10 minutes	1	DNA denaturation & AmpliTaq activation
94°C	30 seconds	35	Denaturation
50 – 65°C	30 seconds		Annealing
72°C	30 seconds		Extension
72°C	10 minutes	-	Extension

Table 2.1. Standard PCR protocol details, including melting, annealing and extension conditions.

PCR reactions were carried out in 96-well skirted plates. Five μL of DNA samples (50 ng) were dispensed into 96-well PCR plates (Griener Bio-one) using an automatic microdispenser (Hydra96, Robbins Scientific, CA, USA). The PCR volume was 50- μL for sequencing templates unless otherwise stated (Table 2.2; section 2.5.1; Chapter 4, table 4.1) and 10- μL for restriction fragment length polymorphism (RFLP) genotyping (Table 2.2; section 2.4.1). PCR reactions were either carried out under mineral oil (for RFLP analysis of PCR products) or the plates were sealed with film lids and the PCR was done in a machine with heated lids (to produce sequencing templates). PCRs were performed using a MBS Satellite Thermal Cycler (Thermo Hybaid, Ashford, UK). PCR products were analysed by agarose gel electrophoresis. For products in the range 100-600 base pairs, 5 μL of PCR product was run on a 3% agarose gel in 0.5 x TBE buffer (Section

2.9, reagents) containing 0.2µg/ml ethidium bromide and viewed under short wave UV light (Section 2.3.2).

2.3.2. Agarose Gel Electrophoresis

Three per cent agarose gels were prepared by dissolving 15 g agarose (Sigma, Poole, UK) in 500 mL 0.5 x tris-borate-EDTA buffer (TBE) (Section 2.9, reagents) by boiling in a microwave oven to dissolve the agarose. Ethidium bromide (final concentration 10µg/ml), which becomes intercalated in the DNA, was added to the cooled agarose before pouring the gel into electrophoresis trays. Five µL of loading buffer (2.5mg/ml bromophenol blue and 2.5mg/ml xylene cyanol in 50% glycerol (Sigma, Poole, UK) and 50% 1x TBE) and 5 µL PCR product were mixed prior loading on the gel. 100-bp ladder (New England Biolabs, Hitchin, UK) was used as a size standard to determine the sizes of the PCR products. DNA fragments were separated by electrophoresis in 0.5 x TBE at 150-180 V for about 40 minutes (depending on the PCR product sizes) to resolve the DNA fragments. The PCR products were visualised under short wave UV light in a GelDoc-It TS transilluminator imaging system (UVP, UK) (Figure 2.1).

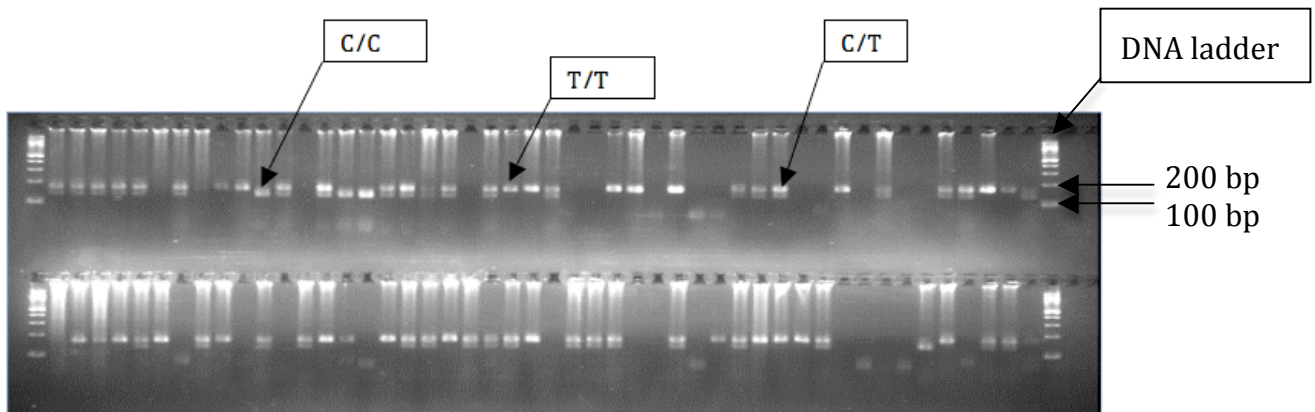


Figure 2.1. rs1126478 3% agarose gel picture for genotype counts. Restriction enzyme *HinfI* cuts where C allele exists. Then, the resulting DNA fragments are separated on an agarose gel. Genotypes are determined by looking at the lengths of these DNA fragments. The long DNA fragments (180 bp) are not digested by the restriction enzyme due presence of the T allele. Shorter DNA fragments are formed as a result of digestion by the restriction enzyme due to presence of the C allele. (bp: Base pairs).

PCR REACTION	For sequencing: Volume (μ L) per well			RFLP genotyping: Volume (μ L) per well			
AmpliTaq Gold Buffer (10X)	5			1			
8 mM dNTPs	1.3			0.5			
[MgCl ₂]	1 mM	2 mM	3 mM	1.5 mM	2.0 mM	2.5 mM	3.0 mM
25 mM MgCl ₂	2	4	6	0.6	0.8	1	1.2
Water	35.5	33.5	31.5	2.4	2.2	2	1.8
[Primer]	20 mM			10 mM			
Forward primer	0.5			0.2			
Reverse primer	0.5			0.2			
AmpliTaq Gold (10 unit/ μ L)	0.2			0.1			
DNA 10ng/ μ L	5			5			
TOTAL VOLUME	50			10			

Table 2.2. Typical PCR reaction details. (AmpliTaq Gold Buffer: 10mM Tris HCl, pH 8.3 at 25°C, 50mM KCl; dNTPs: Deoxynucleotide triphosphates (equimolar ATP, CTP, GTP and TTP); MgCl₂: Magnesium chloride; AmpliTaq Gold: AmpliTaq Gold DNA polymerase (Applied Biosystems, Warrington, UK)). MgCl₂ concentration varied depending on the optimum reaction conditions in both sequencing and RFLP genotyping PCR.

2.4. SNP Genotyping

2.4.1. Restriction Fragment Length Polymorphism

Restriction fragment length polymorphism (RFLP) assays were designed for the confirmation of possible novel SNPs identified in *TNFRSF1A* in Chapter 4 (Table 4.3 and 4.4). Primers were designed using Primer3 for RFLP assays with natural restriction sites (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). Restriction enzymes for RFLP assays were selected either using the NEBcutter (<http://tools.neb.com/NEBcutter2/index.php>) for natural restriction sites, or the dCAPS Finder v 2.0 (dCAPS, derived cleaved amplified polymorphic sequence; <http://helix.wustl.edu/dcaps/dcaps.html>) for engineered restriction sites where no natural site exists. The dCAPS Finder program introduces a single base change within a PCR primer which alters the sequence in the vicinity of polymorphism to enable it to be detected by the use of a restriction enzyme. The base change is near the 3' end of the dCAPS primer, 5' of the SNP which should not be within the dCAPS primer. The dCAPS primers and the PCR products were designed to be 25-30 bp and 200-300 bp long respectively, to enable detection of SNPs upon digestion by the restriction enzyme (Neff *et al.* 2002).

PCRs in 10- μ L were performed following standard procedures (Tables 2.1 and 2.2). For RFLP assays a 5- μ L mixture of the specific restriction enzyme and corresponding reaction buffer (<http://www.neb.com/nebecomm/EnzymeFinder.asp>; chapter 4, tables 4.3 and 4.4.) were added and digested overnight at the appropriate temperature. Digests were designed to produce DNA fragments of distinct lengths (e.g. 160 bp vs. 120

bp) to facilitate resolution in 3% agarose gels and the identification of alleles. Gels were visualised and photographed under UV light (Figure 2.1).

2.4.2. Tag SNP Selection & High Throughput Genotyping

Tag SNP selection for each study was carried out using the Haploview software (<http://www.broadinstitute.org>) (Barrett *et al.* 2005) and the most recent HapMap linkage disequilibrium data for the genetic regions of interest in European samples (<http://www.hapmap.org>). An LD (r^2) threshold of 0.8 was used to identify tagging SNPs in each gene region and the upstream and downstream flanking sequences (~50 kb) (See Chapters 3 and 4 for details of SNP selection) (Halperin & Stephan 2009).

Two higher throughput genotyping technologies, KASPar and iPLEX, were used in candidate gene analysis and in GWAS follow up studies involving low to mid range numbers (<3,000 samples and 100 SNPs). These SNP genotyping methods were used in case-control studies in chapters 3, 4 and 6. Genotyping by iPLEX technology was performed by the core facilities at the Wellcome Trust Centre for Human Genetics (WTCHG, Oxford, UK) and by KASPar technology by KBioscience (Hoddesdon, Herts, UK). For KASPar genotyping, genomic DNA concentration was adjusted to 10 – 20 ng/ μ L using OD_{260 nm} (NanoDrop Technologies, Wilmington, USA) (Ahn *et al.* 1996). For iPLEX genotyping, genomic DNA concentration was adjusted to 10 – 20 ng/ μ L using picogreen according to manufacturer's instructions (Quant-iT™ PicoGreen® dsDNA Reagent, Invitrogen, Paisly, UK; <http://probes.invitrogen.com/media/pis/mp07581.pdf>).

2.4.2.1. KASPar Genotyping

The KBioscience Competitive Allele-Specific PCR genotyping system (KASPar) is a fluorescent resonance energy transfer (FRET) based system (Cuppen 2007). Allele specific PCR along with FRET provides a practical uniplex method for SNP detection, requiring no specifically labelled oligonucleotides (Figure 2.2).

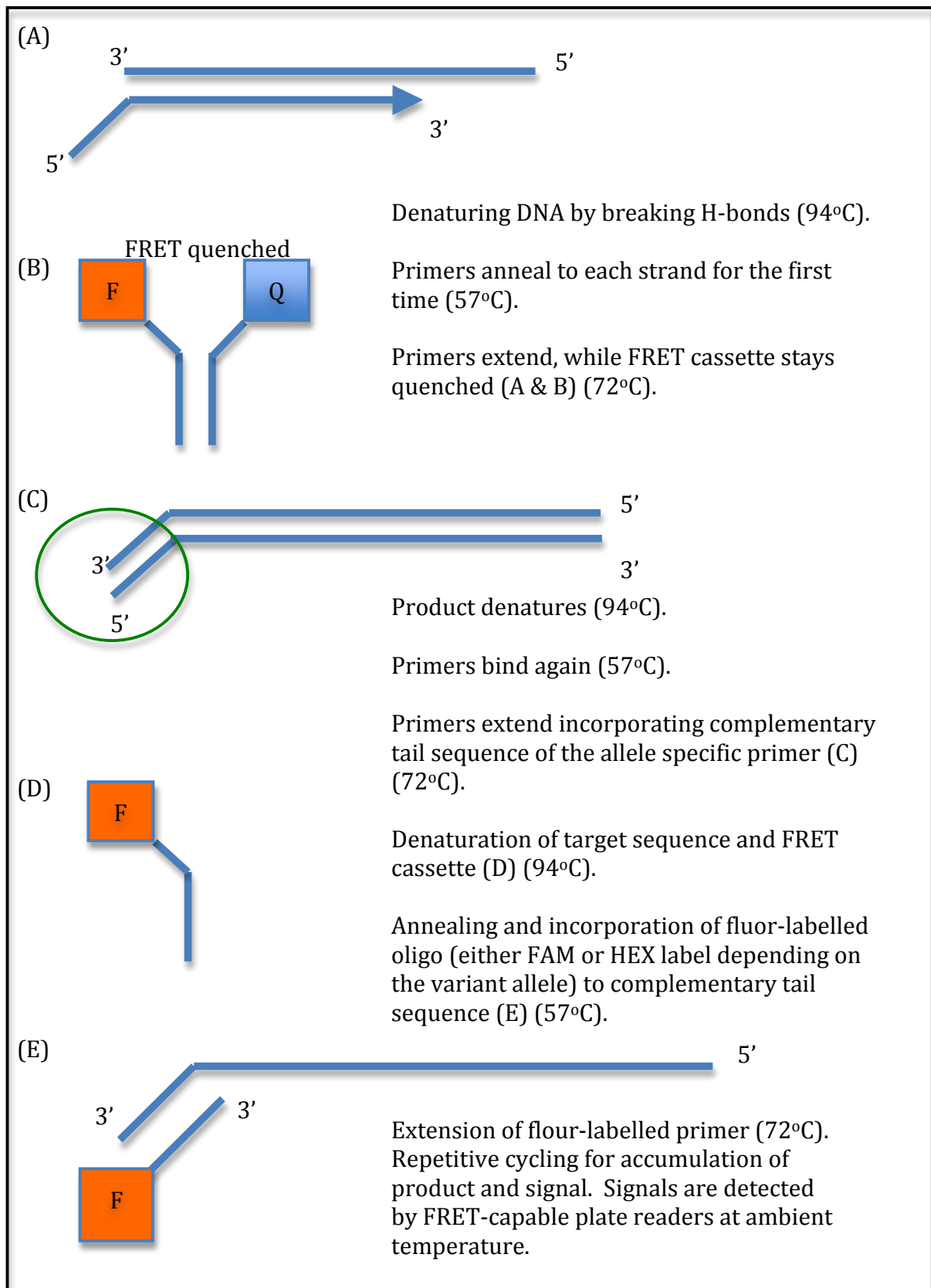


Figure 2.2. A summary of steps (A – E) in KASPar genotyping is illustrated.

Website: <http://www.kbioscience.co.uk/reagents/KASP/KASP.html> (Details are in the main text).

For KASPar genotyping, two allele-specific primers, each with a unique unlabelled tail sequence at the 5' end, and a common reverse primer are required. Two 5' fluor-labelled (F) primers, with complimentary sequences to the tail sequences of the allele-specific primers, separately labelled with FAM and HEX are also required. In addition two primers, complementary to 5' fluor-labelled primers, with 3'-bound quenchers (Q) are required. (Figure 2.2. B).

In the first round of PCR, the allele-specific primers and common reverse primer are extended (Figure 2.2. A). The fluor-labelled primer binds to its complimentary primer and the fluorescent signal is quenched. As the PCR proceeds the fluor-labelled primers are incorporated into the PCR product generating a signal (Figure 2.2. E). Homozygous individuals generate only one fluorescent signal (either FAM or HEX); heterozygous individuals generate both fluorescent signals (FAM and HEX).

2.4.2.2. iPLEX Genotyping

iPLEX technology is based on Sequenom MassARRAY platform which uses multiplex (up to 40 assays) allele-specific primer extension (Gabriel *et al.* 2009). It is a cost-effective and accurate (>99%) assay for SNP genotyping (Ragoussis 2009). iPLEX genotyping was performed at the WTCHG core facilities (Oxford, UK) by Mrs. Lyn-Louise Johnson. There are three main steps in iPLEX genotyping;

- 1) Amplification
- 2) Extension
- 3) Detection by mass spectrometry.

The initial locus-specific PCR reaction (step 1- amplification) is followed by single

base extension of an oligonucleotide primer which anneals immediately upstream of the SNP of interest with mass-modified dideoxynucleotide terminators (step 2 - extension). Each allele of the SNP is identified by the mass of the extended primer using matrix-assisted laser desorption ionisation-time-of-flight (MALDI-TOF) mass spectrometry (MS) (step 3 – detection) (Figure 2.3).

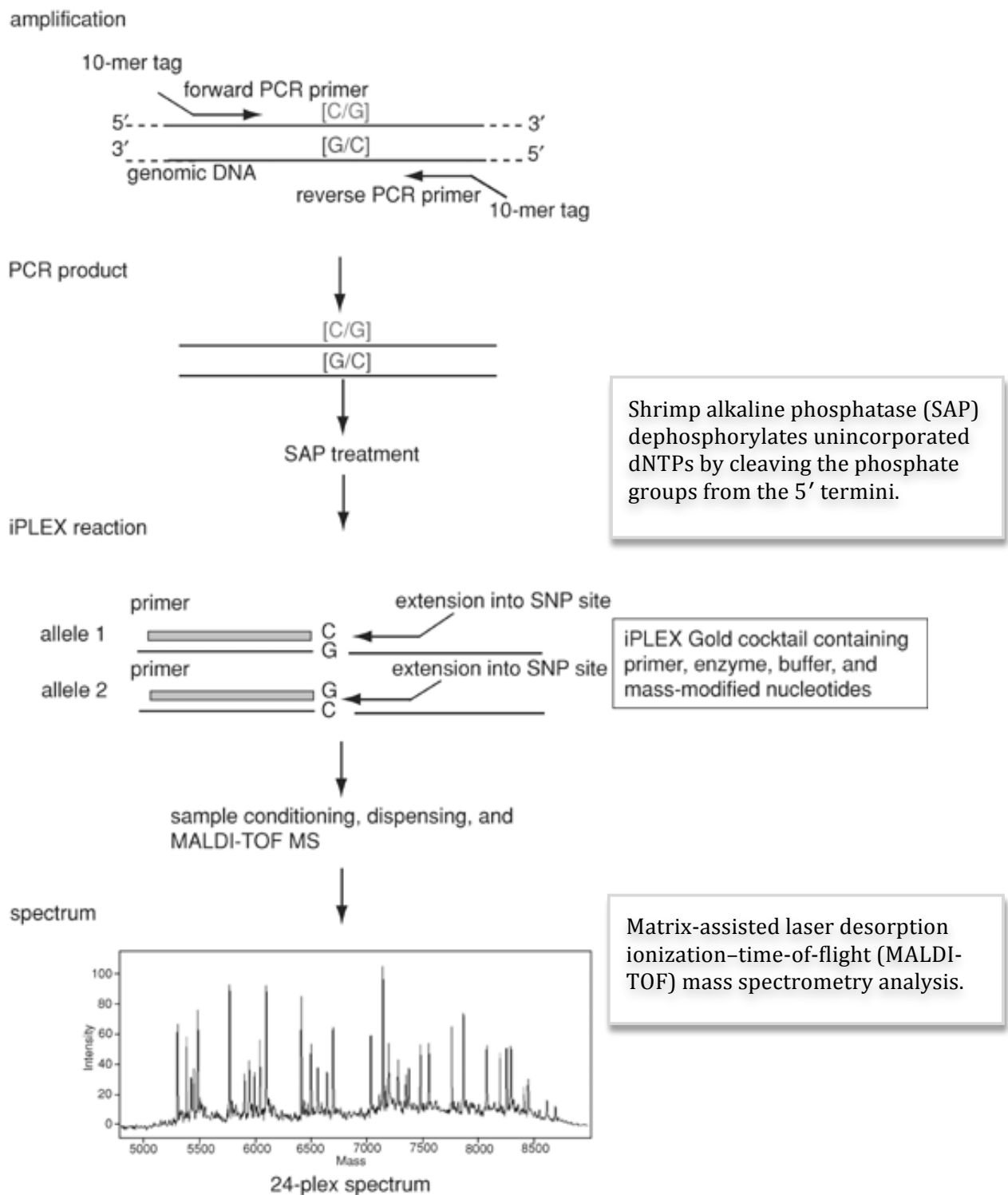


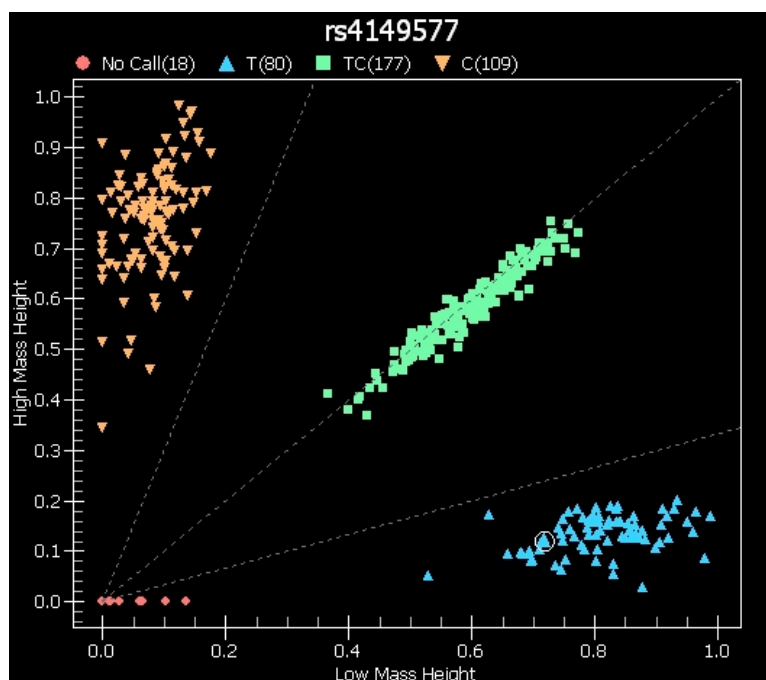
Figure 2.3. Steps of iPLEX genotyping for a C→G SNP: Amplification, extension & detection of different products by mass spectrometry (Adapted from (Gabriel *et al.* 2009)).

The assay provides two levels of specificity. A locus-specific PCR reaction is followed by a locus-specific primer extension reaction. In the iPLEX assay, the primer and amplified target DNA are incubated with mass-modified dideoxynucleotide (ddNTP) terminators. The primer extension occurs according to the sequence at the variant site of interest with a single complementary mass-modified base (Figure 2.3). MALDI-TOF MS is suitable for determining mass of single-stranded DNA in the range of 3-29 base (Ragoussis 2009). The mass of the single base extended primer specifies the alleles present at the site of interest. The mass of the observed primers are translated into genotypes with the software SpectroTYPER (Sequenom, San Diego, CA, USA) (Gabriel *et al.* 2009).

2.4.2.3. Quality Control

Assays for genotyping SNPs by iPLEX were checked on CEPH individuals (Centre d'Etude du Polymorphisme Humain, Utah residents of European descent; 90 samples and 5 replicates). Results were compared with data from the HapMap for concordance. If the genotype call rate was less than 80%, or the genotype discordance was high, then the assay was excluded. These iPLEX quality controls were performed by Mrs. Lynn Louise Johnson (WTCHG). Outlying genotype calls were also checked by observing genotype peaks in the MS spectrum of the iPLEX genotyping. Genotyping by KASPar technology was also checked on an in-house validation panel by the KBiosciences (n=48). Assays with non-specific amplification products were re-designed on the complementary strand, and genotype cluster plots were manually inspected for both genotyping methods (Figure 2.4). SNPs that did not exhibit three quite distinct genotype clusters were excluded from further analysis.

(A)



(B)

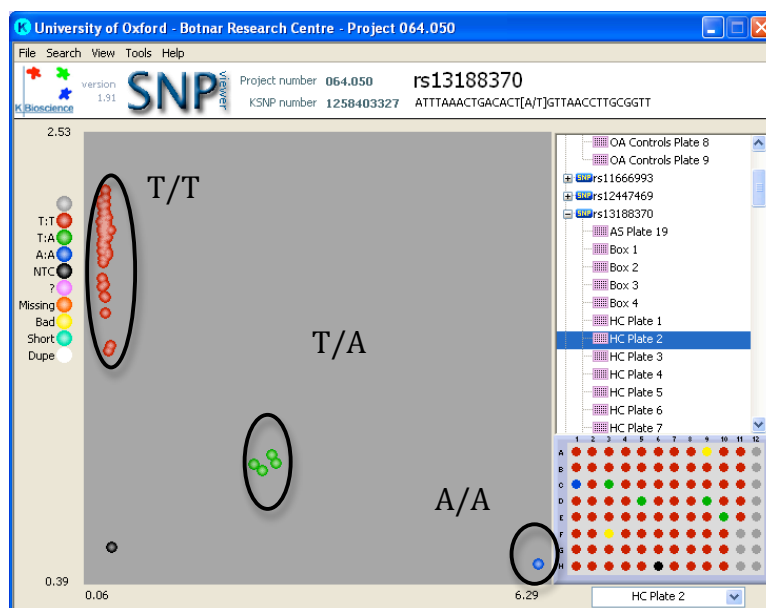


Figure 2.4. Examples of cluster plots from the (A) iPLEX and (B) KASPar genotyping systems showing properly segregated genotype clusters.

SNPs genotypes were tested for adherence to Hardy-Weinberg equilibrium (HWE) in controls using a Chi-square test. This test compares genotype frequencies expected under HWE with those actually observed where the expected values are as follows:

$$p^2 + q^2 + 2pq = 1$$

where p is the major allele frequency and q is the minor allele frequency.

A p-value > 0.05 renders the HWE assumption valid, and the null hypothesis of no difference between observed and expected allele frequencies is not rejected (Balding 2006).

2.5. DNA Sequencing & Detection of Novel SNPs

The complete sequence of the *TNFRSF1A* (NG_007506) and the exon locations provided by the REFSEQ database (<http://www.ncbi.nlm.nih.gov/RefSeq/>) were used to design primers using Primer3 (See Chapter 4, Table 4.1 for primer sequences). Sequencing reactions were carried out using BigDye terminators according to manufacturer's instructions (ABI, Warrington, UK). Products were resolved and detected using ABI 3100 Genetic Analyzer (ABI, Warrington, UK). Exons, intron-exon boundaries and upstream regions of *TNFRSF1A* were sequenced as follows:

2.5.1. PCR Amplification

The first step of sequencing involved PCR amplification of the target region using the designated primers and optimized PCR conditions (Chapter 4, table 4.1 for primers and PCR conditions; table 2.2). An aliquot of the PCR products was resolved on a 3% agarose gel and visualized under UV light to confirm amplification of the target region before proceeding to the next step (section 2.3.2).

2.5.2. Purification of the PCR Products

PCR products from the initial amplification were purified to remove unused reagents such as DNA polymerase, primers and dNTPs that would interfere with the sequencing reaction. PCR products were purified in Millipore MANU3050 96 well filter plates (Watford, UK). PCR products were transferred into Millipore filter plates on a vacuum manifold and allowed to dry for about 5 minutes. Plates were removed from the vacuum and 50 μ L double distilled water (ddH₂O) was pipetted into each well. The PCR products were allowed to re-suspend for about 20 minutes and then transferred to a clean 96-well plate.

2.5.3. Sequencing PCR

BigDye v3.1 (Applied Biosystems, Warrington, UK) sequencing reactions (20- μ L) were set up using the purified PCR product according to the manufacturer's instructions (Table 2.3). Sequencing PCR was performed following the sequencing PCR protocol (Table 2.4). Sequencing reactions were carried out in Abgene (Epsom, Surrey, UK) skirted plates sealed with Robbins cycle seal 96 well plate

sealers (ABgene). PCR was performed using a MBS Satellite Thermal Cycler (Thermo Hybaid, Ashford, UK) with a heated lid.

During the sequencing PCR, dideoxynucleotide triphosphate (ddNTP) terminators are incorporated into the template. Each of these ddNTPs is labelled with a different fluorescent tag for detection by laser photometry on separation.

BigDye v3.1 Reaction	Volume (μ L) per well
Ready reaction premix (2.5X)	2
BigDye sequencing buffer (5X)	2
3.2 μ M Primer	1
Template	5
Water	10
TOTAL VOLUME	20

Table 2.3. BigDye v3.1 reaction set up.

Sequencing PCR Protocol			
Temperature	Time	No. of Cycles	Stage
96°C	5 minutes	-	Denaturation & Activation
96°C	10 seconds	25	Denaturation
50°C	5 seconds		Annealing
60°C	4 minutes		Extension

Table 2.4. Typical sequencing PCR protocol.

2.5.4. Purification of the Sequencing PCR Products

The sequencing PCR products were purified using Sephadex G50 gel to remove unused reagents that may interfere with the analysis of the sequencing reaction products, in particular unincorporated fluorescent ddNTPs. Sephadex G50 (300-400 μ L) suspension in water was transferred into a filter plate, centrifuged at 4000g in centrifuge for 5 minutes, collecting the water in a plate placed under the filter plate to be discarded. The sample to be purified was loaded on the Sephadex and centrifuged at 4000g for 5 minutes collecting the purified PCR product into ABGENE non-skirted 96-well plate. 10 μ L of HiDi Formamide (Applied Biosystems, Warrington, UK) was added to the purified PCR product and the plate was sealed.

2.5.5. Analysis of DNA Sequencing Reaction Products

AB3100 Genetic Analyzer (Applied Biosystems, Warrington, UK) was used to resolve and analyse the products of the sequencing reactions. This process involves electrophoresis at 50oC through 36-cm capillaries filled with the ABI PRISM 3100 POP4 polymer. The negatively charged DNA moves towards the anode. Electrophoresis is performed under denaturing conditions so that the rate of migration of the PCR products depends entirely on their size with longer PCR products migrating at slower rates. The fluorescently labelled DNA products pass through a laser excitation beam in the detection cell. Sequences were analysed with polyphred for detection of SNPs.

2.5.6. Analysis of DNA Sequences with Polyphred for Detection of SNPs

Sequence data was analyzed by the software Polyphred (<http://droog.gs.washington.edu/polyphred/>) (Nickerson *et al.* 1997; Stephens *et al.* 2006). This software compares sequences with a reference sequence and with each other, detecting and scoring any variant site. Detailed instructions for obtaining the Polyphred package, installation and use are explained elsewhere (Montgomery *et al.* 2008). The sequence traces were displayed as electropherograms by the Consed tool supplied with the Polyphred package (See Chapter 4, Figure 4.3). Sequence variants identified by the software were confirmed by manual inspection of the electropherograms (See Chapter 4, Section 4.2).

2.6. Statistical Analysis of Directly Genotyped Variants

A variety of statistical analyses were performed to test for disease association in Chapters 3, 4, 5 and 6. These are detailed in the individual chapters. Data manipulation and statistical calculations were performed using statistical software packages and UNIX-based programs run on servers maintained by the Bioinformatics and Statistical Genetics group at the Wellcome Trust Centre for Human Genetics (Oxford, UK).

2.6.1. Statistical Power Calculations

Statistical power calculations for case-control studies were carried out using the program Quanto (version 1.2.3, May 2009) (<http://hydra.usc.edu/gxe/>)

(Gauderman 2002). Parameters used for power calculations are stated in each chapter. As the size of a study population, disease prevalence, minor allele frequency and odds ratio (OR) increase, the statistical power to detect an association between the variants and disease of interest increases.

2.6.1.1. Estimation of the Prevalence for Power Calculations

An estimate of the prevalence of AS is important for genetic studies as the statistical power calculations depend on the prevalence figure. However, there are no formal estimates of the prevalence of AS in the UK. Therefore, I have used an average European figure for the prevalence of AS in the UK. By taking the average of the European study results from Holland (0.2%) and Germany (0.55%), it is likely that the prevalence that is likely to be around 0.4% (van der Linden *et al.* 1984; Braun *et al.* 1998; Braun & Sieper 2007).

2.6.2. Genetic Association Analysis

In case-control studies and stratified analyses, 2 x 2 contingency tables and Pearson's χ^2 test (1 df) were used to identify associations between allele frequencies and AS. The Cochran-Armitage test of trend (1 df) was used to test a linear relationship between allelic dosage and risk of AS. Where relevant, a Bonferroni correction for multiple testing was applied by dividing the significance level (α) by the number tests carried out (Sasieni 1997; Perneger 1998; Lewis 2002; Balding 2006; Clarke *et al.* 2011).

The haplotype analyses reported in Chapters 4 and 5 were performed using PLINK (v1.07, 2009) (Purcell *et al.* 2007). Such analyses test for differences in haplotype

frequencies between cases and controls within a region of interest with low recombination. Fourteen SNPs were genotyped in Chapter 4, a sliding window method was used to analyse haplotypes containing between 2 and 14 SNPs. Differences in the frequencies of the predicted haplotypes between cases and controls were tested using parametric chi-square test (1 df) for calculating asymptotic p-values for common haplotypes (frequency>5%), and non-parametric permutation (number of permutations=1000) in order to calculate the empirical corrected p-values for low frequency haplotypes (frequency<5%). In Chapter 5, a two-SNP haplotype (rs30187(*ERAP1*)-rs4869313(*ERAP2*)) was analysed. Differences in the frequencies of the haplotypes between cases and controls were tested using chi-square test (1 df).

2.6.3. Meta-analysis

Meta-analyses increase statistical power to detect genetic associations by combining different datasets taking into account their relative sizes (i.e. each study weighted by its inverse variance). The StatsDirect statistical package (<http://www.statsdirect.com>) was used to perform the meta-analyses with either the Mantel-Haenszel odds ratio (fixed-effects model) or the DerSimonian-Laird odds ratio (random-effects model) pooling method. In a fixed-effects model, it is assumed that there is one true effect size underlying all studies and that any variation between studies in terms of effect size estimates is purely due to sampling variation. A random effects model assumes that effects detected by different studies may vary following a normal distribution. It has less statistical power to detect a given effect, and therefore requires larger study populations (Ades *et al.* 2005; Higgins *et al.* 2009).

Since it is important only to include comparable studies in a meta-analysis, it is necessary to identify possible heterogeneity across studies. To identify and quantify heterogeneity between studies in a meta-analysis, Cochran Q p-values and inconsistency (I^2 , %) values were calculated for each SNP to determine the combinability of the data. Cochran's Q is calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies, with the weights being those used in the pooling method. Q follows a chi-square distribution with $k-1$ (k =number of studies) degrees of freedom. When the number of studies is small, Cochran's Q has low statistical power as a test of heterogeneity but maybe over powered if the number of studies is large (Higgins *et al.* 2003). A Cochran Q p-value < 0.05 was considered significant, providing evidence for heterogeneity between studies. The inconsistency (I^2) statistic describes the percentage of variation across studies due to heterogeneity rather than chance. I^2 is not affected by the number of studies in the meta-analysis (Higgins *et al.* 2003). An inconsistency value larger than 50% was considered as evidence for substantial heterogeneity across studies.

Forest plots were produced for each SNP in the meta-analysis and provide visual information about study heterogeneity. When significant heterogeneity is not present between studies, confidence intervals are expected to overlap, with similar direction and magnitude of effects detected by each study (See Chapter 3, Figures 3.3.1-3.3.4).

2.6.4. Case-Control Rare Variant Analysis Tool (CCRaVAT)

CCRaVAT analysis was used to test for disease association with rare to low frequency variants (MAF<5%) within gene regions (<http://www.sanger.ac.uk/resources/software/rarevariant/>). Data on directly genotyped variants from the WTCCC2 study were used for the CCRaVAT analysis. CCRaVAT compared the presence or absence of rare variant minor alleles in cases and controls for each gene region in 2x2 contingency tables. Differences in the proportion of cases and controls carrying rare variant minor alleles were tested using a Pearson's chi-squared test or a Fisher's exact test when cell counts were small (<20). Thus, rare and low frequency variants (MAF<5%) within gene regions were combined to look for accumulation of minor alleles within genes, and tested for association with AS. This approach can increase power to detect association when several rare variants in a gene exist all predisposing to disease (Lawrence *et al.* 2010; Morris & Zeggini 2010).

CCRaVAT was run with the following command line analyzing each chromosome separately:

```
nohup perl ccravat.pl -gene -gPval=0.05 -graph
```

CCRaVAT collapsed and analysed the variants within regions including 50 kilobases up- and down-stream of the genes (gene build NCBI 36 (hg18)), producing Manhattan plots for each chromosome and highlighting results with p-values < 0.05. The maximum MAF cut off was 0.05 (by default).

2.7. Genotype Imputation

Genotype imputation (i.e. inferring genotypes at loci not physically genotyped) is described in detail in Chapters 1 and 5 (See Chapter 1, Section 1.7). Genome-wide SNP genotyping data from the Wellcome Trust Case-Control Consortium 2 (WTCCC2) discovery set (1782 cases and 5167 controls) were imputed using data from the 1000 Genomes Project pilot study (August 2009) as a reference set (Genomes 2010; Evans *et al.* 2011). In August 2009 when this work was undertaken, only 122 haplotypes from the CEU (Utah residents of Northern and Western European ancestry) population were available and these were used to infer genotypes in the WTCCC2 study population of the European descent (Evans *et al.* 2011). Genotypes were imputed using the Markov Chain based haplotyper (MACH 1.0) (Li *et al.* 2009; Li *et al.* 2010). Markov chains are used in many statistical models. In simple terms, the Markov chain tests all possible states (i.e. haplotypes) and assigns a probability to each determining the most likely state (Halperin & Stephan 2009). There were 4 main steps of genotype imputation; preliminary checks, estimating model parameters, imputation and disease association analysis using the imputed genotypes.

2.7.1. Input Data: Preliminary Checks & Data Filtering

All quality control measures of the WTCCC2 data were carried out as described in Evans *et al.* (2011) by Dr David Evans and members of the WTCCC2 team. These included checking SNP adherence to Hardy-Weinberg equilibrium ($p < 5 \times 10^{-20}$ in cases or controls), SNP genotype missingness ($> 2\%$), and checking individuals for genotype call rates, heterozygosity, ancestry (using ancestry informative markers), cryptic relatedness, gender and identity (Evans *et al.* 2011). The WTCCC2 data

were divided by chromosomes using the appropriate commands in PLINK (Purcell *et al.* 2007) and each chromosome was imputed separately. The reference CEU haplotypes (n=122) containing 9,232,128 markers from the 1000 Genomes Project were obtained from the MACH website (<http://www.sph.umich.edu/csg/abecasis/MACH>, UoM, 2009-08 release). These haplotypes only contained SNPs that were detected more than once.

2.7.2. MACH: Estimating Model Parameters

MACH 1.0 uses a hidden Markov Chain model and a 2-step process to impute genotypes (Halperin & Stephan 2009; Li *et al.* 2009; Li *et al.* 2010). The 2-step process makes imputation more efficient and less computationally demanding (Li *et al.* 2010). The initial step was to estimate the model parameters where a model was built to relate the samples in the study to the reference haplotypes.

Genotyping error rate for each SNP and a crossover rate, describing the recombination pattern in haplotypes shared by the study samples and the reference panel, were estimated. Factors affecting the estimation of the model parameters include the number of individuals used in the initial step and the number of iterations carried out to generate the model. MACH guidelines were followed, and 100 iterations using 250 cases and 250 controls were used to estimate parameters in the first stage of imputation.

2.7.3. MACH: Carrying Out Genotype Imputation

The second step of imputation involved using the estimated model parameters to impute all the SNPs in the reference panel in the WTCCC2 study population. This

step was performed following the MACH guidelines. Imputation of genotypes for the whole WTCCC2 study population was carried out separately for each chromosome.

2.8. Statistical Analysis of Imputed Genotypes

Various types of genetic association analysis were performed with the imputed genotypes and these are described in more detail in the next section and in Chapter 5. Data manipulation and statistical calculations were performed using freely available software packages and UNIX-based statistical programs run on servers maintained by the Bioinformatics and Statistical Genetics group at the Wellcome Trust Centre for Human Genetics (Oxford, UK).

2.8.1. Mach2Dat: Logistic Regression Analysis of Binary Data on Imputed Genotypes

2.8.1.1. Logistic Regression Analysis

Logistic regression analysis was carried out to test for association between the binary outcome of interest, (presence or absence of AS), and the imputed SNPs using the software Mach2dat (Bewick *et al.* 2005; Balding 2006; Li *et al.* 2009; Li *et al.* 2010). Specifically, case-control status was regressed on expected allelic dosage rather than best guess genotypes in order to incorporate uncertainty in the imputation. It is particularly important to model this uncertainty when analyzing less common SNPs (Li *et al.* 2009). Only SNPs with an RSQR imputation quality > 0.3 were analyzed.

The Wald chi-square test and the likelihood ratio test are asymptotically equivalent and give very similar results for large study populations (Engle, 1983; Bewick *et al.*, 2005). I used the Wald chi-square test statistic and p-value for reviewing the logistic regression analysis results. The Wald chi-square test statistic follows a chi-square distribution with 1 degree of freedom. PLINK was used to annotate the results with gene names and SNP features, such as synonymous coding and non-synonymous coding, following the online instructions (<http://pngu.mgh.harvard.edu/~purcell/plink/>) (Purcell *et al.* 2007).

2.8.1.2. Inflation Factor, Q-Q and Manhattan Plots

Genome-wide Manhattan plots and Q-Q plots were produced using R and freely available R packages. The R package qqman was installed from the source website (<http://www.StephenTurner.us/qqman.r>) and run following the online instructions (<http://gettinggeneticsdone.blogspot.com/2011/04/annotated-manhattan-plots-and-qq-plots.html>). Genome-wide Manhattan plots visualised the genetic association signals with $-\log_{10}(\text{p-value})$ on the y-axis and the position on the chromosomes on the x-axis. Q-Q plots showed the deviation of the observed chi-square distribution from the expected chi-square distribution with 1 degree of freedom. The inflation factor (λ) was then calculated using the command line below in R to estimate this deviation from the expected distribution that may due to population stratification, cryptic relatedness and/or genetic association (See Chapter 5) (Balding 2006).

Inflation factor = `Median(data$WALDCHISQ)/0.456`

where the median of the WALDCHISQ column was divided by the median for chi-square distribution with 1 degree of freedom.

2.8.1.3. Regional Association Plots

Association plots of regions of interest were produced with the online tool LocusZoom using the genome build NCBI 36 (hg18) and the 1000 Genomes August 2009 CEU LD data (<http://csg.sph.umich.edu/locuszoom/>) (See Chapter 5) (Pruim *et al.* 2010).

2.8.1.4. Conditional Logistic Regression Analysis

Conditional logistic regression analyses were carried out in regions of interest to determine if there were multiple associations. Imputed dosage data from the most strongly associated SNP was included as a covariate to look for any other independent associations within the region. This covariate data was added to the .ped file and mach2dat was run following the instructions in section 2.8.1.1.

2.8.1.5. HLA-B27-Stratified Logistic Regression Analysis

The SNP rs4349859 tags HLA-B27 with a sensitivity (98.0%) and specificity (99.0%) (Evans *et al.* 2011). Cases and controls were stratified as HLA-B27 positive (rs4349859 AA or AG) or HLA-B27 negative (rs4349859 GG). Logistic regression analyses were performed for each strata by chromosome using Mach2dat following the instructions in section 2.8.1.1.

2.8.2. GRANVIL: Gene- or Region-Based Analysis of Variants of Intermediate and Low Frequency

Gene- or region-based analysis of variants of intermediate and low frequency (GRANVIL) was carried out to identify genetic associations within gene regions

using the imputed dosages (a value between 0 - 2) of rare and low frequency variants (MAF<5%) (<http://www.well.ox.ac.uk/GRANVIL/>) (Morris & Zeggini 2010; Magi *et al.* 2011). Similar to the CCRaVAT analysis (See Section 2.6.4), combined dosages of rare and low frequency variants within genes were tested for disease association. The difference between the GRANVIL and the CCRaVAT methods was that CCRaVAT analysis used minor allele counts (0, 1 or 2) from the WTCCC2 discovery set, which contained much less dense SNP genotype data compared to the imputed data used in the GRANVIL analysis.

2.8.2.1. Data Re-formatting & GRANVIL Analysis

The GRANVIL v0.8 program required input files in the SNPTEST format. The perl script mach2snptest.pl was used to change the file format following the online instructions (http://www.openbioinformatics.org/gengen/tutorial_convert_mach.html). Input files in the SNPTEST format (.gen and .sample) were produced from MACH .mlinfo and .mlprob files. An exclusion file with imputed SNPs of low quality (rsq < 0.5) was produced and these SNPs were excluded from the analysis. The rsq threshold was increased for this analysis as imputation of rare and low frequency variants (MAF<5%) is less reliable than for common variants (MAF>5%). Low quality SNPs were removed from the input file (.gen) using GTOOL (v0.6.6) following the online instructions (<http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html>). Gene coordinates were extracted from the genome build NCBI 36 (hg18) for each chromosome to produce the necessary legend files for the analysis of gene regions.

GRANVIL v0.8 was run separately for each chromosome with the recommended command line:

```
nohup GRANVIL -g chr20_snptest.gen -s chr20_snptest.sample --pheno 1 --rare
0.05 --genmap chr20_genelistb36.txt --dosages -o chr20_granvil
```

GRANVIL v0.8 was run with the following command line with one covariate, HLA-B27 status (rs4349859 genotype), due to possible confounding by strong LD within the HLA region on chromosome 6 to identify any independent genetic associations:

```
nohup GRANVIL -g chr6_snptest.gen -s chr6_snptest_cov.sample --pheno 1 --cov 1 -
-rare 0.05 --genmap chr6_genelistb36.txt --dosages -o chr6_granvil_cov
```

GRANVIL used imputed allele dosages (0-2) from the genotype file to perform a logistic regression analysis. The phenotype was modelled in a regression framework as a function of the proportion of rare variants at which an individual carries a minor allele. Rare and low frequency variants (MAF<5%) were therefore combined to determine accumulation of minor alleles within gene regions, and tested for association with AS. This approach can increase power to detect association when several rare variants in a gene exist all of which predispose to disease (Bodmer & Bonilla 2008; Li & Leal 2008; Morris & Zeggini 2010; Magi *et al.* 2011).

2.8.3. Imputation Accuracy: Concordance between Imputed & Directly Genotyped SNPs

A confirmation study involving genotyping of selected SNPs (which were possibly associated with AS according the results of the imputed data analysis) was later carried out in an independent population (See Chapter 6). Within this study, ~350 individuals from the imputation project (WTCCC2 study population) were also genotyped to estimate the accuracy of imputation using the 1000 Genomes Project pilot data. The concordance rates (%) were determined by comparing the most likely (imputed) genotypes for these individuals with the direct genotyping results. The genotypes for each SNP were also divided into 3 strata (minor allele homozygote, heterozygote and major allele homozygote) to determine genotype concordances for each of these strata, and to find out if there is an effect of the minor allele frequency (MAF) on imputation accuracy.

2.9. Reagents

Lysis Buffer

Sucrose	109.4 g
1M Tris-HCl (pH 7.8)	1 mL
1.0M MgCl ₂	5.8 mL
Triton X-100	10 mL

Add ~500 mL of distilled water

Mix to dissolve sucrose.

Make up to 1 L with distilled water. Store at 4°C, use within 2 weeks.

Solution B

Tris HCl	48.44 g, or 333 mL 1M Tris HCl (pH 8.0)
EDTA	22.33 g
Sodium Chloride	8.766 g

H₂O to 600 mL

pH adjusted to 8.0 to dissolve EDTA dissolves, if not using 333 mL 1M Tris HCl pH

8.0.

H₂O to 900 mL

1% SDS Add 10% SDS 100 mL

5M Sodium Perchlorate

Sodium Perchlorate 72.05 g

H₂O to 100 mL

Do not autoclave.

1 x T.E.

10 mM Tris-HCL

1mM EDTA

Adjust pH to 8.0 using 40% NaOH.

1 x TBE

Tris Base 108 g

Boric Acid 55 g

0.5M EDTA (pH 8.0) 20 mL

Make up to 10 litres with dH₂O

2.10. Online Resources

A list of the various online resources that were used for obtaining genetic information, designing studies, statistical analyses and data manipulation is presented below (Table 2.5).

Resource	URL
Genome Browser	
National Centre for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/
University of California Santa Cruz (UCSC) Genome Browser	http://genome.ucsc.edu/
The International HapMap Project	http://www.hapmap.org/
Ensembl Genome Browser	http://www.ensembl.org/
1000 Genomes Project	http://www.genome.gov/
Genetic Study Design Tools	
Primer3	http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
NEBcutter V2.0	http://tools.neb.com/NEBcutter2/index.php
dCAPS Finder 2.0	http://helix.wustl.edu/dcaps/dcaps.html
Data Manipulation & Analysis	
Polyphred for SNP Detection	http://droog.gs.washington.edu/polyphred/
Imputation – Markov Chain Based Haplotyper (MACH 1.0)	http://www.sph.umich.edu/csg/abecasis/MACH/index.html
Binary Trait Analysis - Mach2Dat	http://www.sph.umich.edu/csg/abecasis/MACH/download/
PLINK: Whole Genome Association Analysis Toolset	http://pngu.mgh.harvard.edu/~purcell/plink/
GenGen – Imputation Helper	http://www.openbioinformatics.org/gengen/tutorial_convert_mach.html
GTOOL	http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html
Power/Sample Size Calculation for Logistic Regression with Binary Covariate(s)	http://www.dartmouth.edu/~eugened/power-samplesize.php
Gene- or Region-Based Analysis of Variants of Intermediate and Low Frequency (GRANVIL)	http://www.well.ox.ac.uk/GRANVIL/
Case-Control Rare Variant Analysis Tool (CCRaVAT)	http://www.sanger.ac.uk/resources/software/rarevariant/
SNP Annotation & Proxy Search (SNAP)	http://www.broadinstitute.org/mpg/snap/
LocusZoom	http://csg.sph.umich.edu/locuszoom/
MACH: 1000 Genomes Imputation Cookbook	http://genome.sph.umich.edu/wiki/MaCH:_1000_Genomes_Imputation_Cookbook

Table 2.5. Online resources used in this work with their Internet addresses.

2.11. References

- Ades, A. E., Lu, G., *et al.* (2005). "The interpretation of random-effects meta-analysis in decision models." *Med Decis Making* **25**(6): 646-654.
- Ahn, S. J., Costa, J., *et al.* (1996). "PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR." *Nucleic Acids Res* **24**(13): 2623-2625.
- Balding, D. J. (2006). "A tutorial on statistical methods for population association studies." *Nat Rev Genet* **7**(10): 781-791.
- Barrett, J. C., Fry, B., *et al.* (2005). "Haploview: analysis and visualization of LD and haplotype maps." *Bioinformatics* **21**(2): 263-265.
- Bewick, V., Cheek, L., *et al.* (2005). "Statistics review 14: Logistic regression." *Crit Care* **9**(1): 112-118.
- Bodmer, W. & Bonilla, C. (2008). "Common and rare variants in multifactorial susceptibility to common diseases." *Nat Genet* **40**(6): 695-701.
- Braun, J., Bollow, M., *et al.* (1998). "Prevalence of spondylarthropathies in HLA-B27 positive and negative blood donors." *Arthritis Rheum* **41**(1): 58-67.
- Braun, J. & Sieper, J. (2007). "Ankylosing spondylitis." *Lancet* **369**(9570): 1379-1390.
- Calin, A., Garrett, S., *et al.* (1994). "A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index." *J Rheumatol* **21**(12): 2281-2285.
- Clarke, G. M., Anderson, C. A., *et al.* (2011). "Basic statistical analysis in genetic case-control studies." *Nat Protoc* **6**(2): 121-133.
- Cuppen, E. (2007). "Genotyping by Allele-Specific Amplification (KASPar)." *CSH Protoc* **2007**: pdb prot4841.
- Evans, D. M., Spencer, C. C., *et al.* (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**(8): 761-767.
- Gabriel, S., Ziaugra, L., *et al.* (2009). "SNP genotyping using the Sequenom MassARRAY iPLEX platform." *Curr Protoc Hum Genet* **Chapter 2**: Unit 2 12.
- Garrett, S., Jenkinson, T., *et al.* (1994). "A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index." *J Rheumatol* **21**(12): 2286-2291.
- Gauderman, W. J. (2002). "Sample size requirements for association studies of gene-gene interaction." *Am J Epidemiol* **155**(5): 478-484.
- Genomes (2010). "A map of human genome variation from population-scale sequencing." *Nature* **467**(7319): 1061-1073.
- Halperin, E. & Stephan, D. A. (2009). "Maximizing power in association studies." *Nat Biotechnol* **27**(3): 255-256.
- Halperin, E. & Stephan, D. A. (2009). "SNP imputation in association studies." *Nat Biotechnol* **27**(4): 349-351.
- Higgins, J. P., Thompson, S. G., *et al.* (2003). "Measuring inconsistency in meta-analyses." *BMJ* **327**(7414): 557-560.
- Higgins, J. P., Thompson, S. G., *et al.* (2009). "A re-evaluation of random-effects meta-analysis." *J R Stat Soc Ser A Stat Soc* **172**(1): 137-159.
- Lawrence, R., Day-Williams, A. G., *et al.* (2010). "CCRaVAT and QuTie-enabling analysis of rare variants in large-scale case control and quantitative trait association studies." *BMC Bioinformatics* **11**: 527.

- Lewis, C. M. (2002). "Genetic association studies: design, analysis and interpretation." Brief Bioinform **3**(2): 146-153.
- Li, B. & Leal, S. M. (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." Am J Hum Genet **83**(3): 311-321.
- Li, Y., Willer, C., *et al.* (2009). "Genotype imputation." Annu Rev Genomics Hum Genet **10**: 387-406.
- Li, Y., Willer, C. J., *et al.* (2010). "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." Genet Epidemiol **34**(8): 816-834.
- Magi, R., Kumar, A., *et al.* (2011). "Assessing the impact of missing genotype data in rare variant association analysis." BMC Proc **5 Suppl 9**: S107.
- Montgomery, K. T., Iartchouck, O., *et al.* (2008). "PolyPhred analysis software for mutation detection from fluorescence-based sequence data." Curr Protoc Hum Genet **Chapter 7**: Unit 7 16.
- Morris, A. P. & Zeggini, E. (2010). "An evaluation of statistical approaches to rare variant analysis in genetic association studies." Genet Epidemiol **34**(2): 188-193.
- Neff, M. M., Turk, E., *et al.* (2002). "Web-based primer design for single nucleotide polymorphism analysis." Trends Genet **18**(12): 613-615.
- Nickerson, D. A., Tobe, V. O., *et al.* (1997). "PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing." Nucleic Acids Res **25**(14): 2745-2751.
- Perneger, T. V. (1998). "What's wrong with Bonferroni adjustments." BMJ **316**(7139): 1236-1238.
- Pruim, R. J., Welch, R. P., *et al.* (2010). "LocusZoom: regional visualization of genome-wide association scan results." Bioinformatics **26**(18): 2336-2337.
- Purcell, S., Neale, B., *et al.* (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.
- Ragoussis, J. (2009). "Genotyping technologies for genetic research." Annu Rev Genomics Hum Genet **10**: 117-133.
- Rozen, S. & Skaletsky, H. (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-386.
- Sasieni, P. D. (1997). "From genotypes to genes: doubling the sample size." Biometrics **53**(4): 1253-1261.
- Stephens, M., Sloan, J. S., *et al.* (2006). "Automating sequence-based detection and genotyping of SNPs from diploid samples." Nat Genet **38**(3): 375-381.
- van der Linden, S., Valkenburg, H. A., *et al.* (1984). "Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria." Arthritis Rheum **27**(4): 361-368.
- van der Linden, S. M., Valkenburg, H. A., *et al.* (1984). "The risk of developing ankylosing spondylitis in HLA-B27 positive individuals. A comparison of relatives of spondylitis patients with the general population." Arthritis Rheum **27**(3): 241-249.

CHAPTER 3

3. INTERLEUKIN 23 RECEPTOR & ANKYLOSING SPONDYLITIS

3.1. Introduction

IL23R, encoding a component of the interleukin 23 receptor, is one of several genes implicated in the pathogenesis of AS by genome-wide association studies (Burton *et al.* 2007; Reveille *et al.* 2010; Evans *et al.* 2011). Variants within *IL23R* have also been associated with inflammatory bowel disease (IBD) and psoriasis (PS) (Duerr *et al.* 2006; Cargill *et al.* 2007; Nunez *et al.* 2008; Franke *et al.* 2010; Strange *et al.* 2010; Anderson *et al.* 2011; Patsopoulos *et al.* 2011). *IL23R* therefore appears to be a common susceptibility locus for spondyloarthropathies and related conditions (Burton *et al.* 2007).

In the initial 2007 Wellcome Trust Case-Control/Australo-Anglo-American (Triple A) Spondylitis Consortia (WTCCC-TASC) study, variants within *IL23R* showed moderate evidence of association with AS (non-synonymous coding SNP rs11209026, OR=0.6, $p=1.7 \times 10^{-3}$). Further genotyping of cases and controls from the US replicated the association (rs11209026, OR=0.6, $p=0.014$; rs11209032, OR=1.3, $p=0.0013$) (Burton *et al.* 2007). This finding was subsequently replicated by other studies from Canada (Rahman *et al.* 2008) and Spain (Rueda *et al.* 2008).

The results presented here include a new UK case-control study, an extended UK study and a meta-analysis of all the published series (up to June 2008) which were performed in order to confirm and strengthen the evidence for association between AS and *IL23R*. Thus, study was aimed to serve as a confirmation of the *IL23R* association in the UK Caucasians and across other ethnicities after the WTCCC-TASC GWAS (2007). Furthermore, ~15% of AS patients also have IBD and/or psoriasis (Feldtkeller *et al.* 2003; El Maghraoui 2011). It was important to ascertain that the signal from *IL23R* with AS was not arising from concomitant psoriasis and/or IBD in a subgroups of patients with AS, since these inflammatory disorders are also associated with variants in *IL23R*. The association between *IL23R* variants and AS was also evaluated after excluding those AS cases with IBD and/or psoriasis. This stratified analysis was carried out as a part of the UK cases-control study in this chapter.

3.2. Patients, Materials & Methods

3.2.1. UK Case-Control Study & Extended Study of Association between *IL23R* & AS

In this study, 730 independent AS cases (See Chapter 2, Section 2.1) were compared to 1331 healthy controls from the UK 1958 British Birth Cohort (58BBC) previously genotyped by the WTCCC-TASC in its initial gene targeted GWAS (Burton *et al.* 2007). These new cases were then combined, in an extended study, with 1088 previously published cases from the WTCCC-TASC study (Burton *et al.* 2007) and compared to the same controls using a case-control design. None of the new UK cases had been included in previous studies of *IL23R*, but the same 58BBC controls as in the WTCCC-TASC study (Burton *et al.* 2007) were used as controls in

this study. The minor allele frequencies of the SNPs from the 58BBC controls were therefore compared with those from the HapMap population of the Western and Northern European ancestry to ensure that the observed associations were not due to bias in the selection of the 58BBC control population.

SNPs were genotyped in the new cases using iPLEX technology except for rs11209026, which was genotyped using KASPar technology (Figure 3.1; Figure 3.2 for SNPs in the study) (See Chapter 2, Section 2.4.2 for SNP selection, genotyping and quality control). All genotypes were checked for Hardy-Weinberg equilibrium. SNP genotyping of the WTCCC controls and cases had previously been performed with the Infinium I assay (Illumina, San Diego, CA), (Burton *et al.* 2007). Allelic associations were analysed using StatsDirect software (version 2.6.6 03/02/2008). The strength of association was tested using chi-square analysis of contingency tables (See Chapter 2, Section 2.6.2). A p-value < 0.05 was considered statistically significant, since there was a clear *a priori* hypothesis of positive association. The statistical power of the UK case-control study and the extended study to detect association with the variants in *IL23R* were 89% and 98% respectively, given an OR of 1.6, a MAF of 0.04, an alpha level of 0.05 (2-sided) and a disease prevalence of 4/1000 with a log-additive disease model (Quanto, version 1.2.4, May 2009) (See Chapter 2, Section 2.6.1). Previous OR estimates from the WTCCC-TASC GWAS were used for these power calculations (Burton *et al.* 2007).

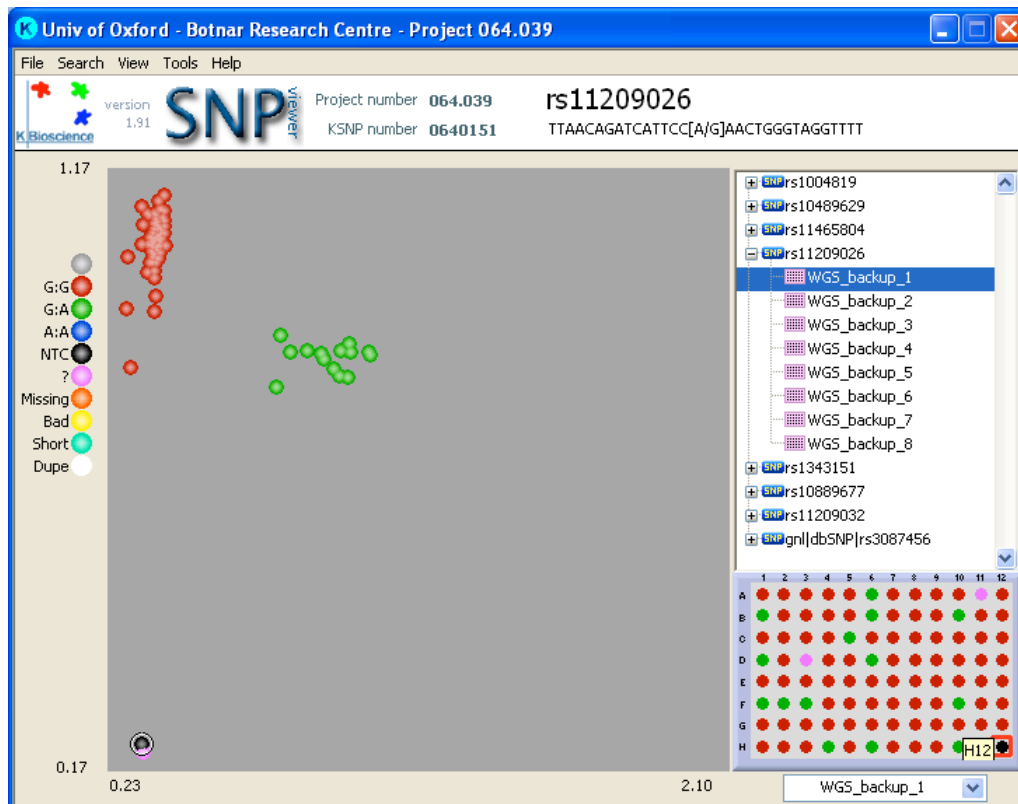


Figure 3.1. Genotype cluster plot for rs11209026 showing separate genotypes (KBiosciences SNPviewer Version 1.91). Red dots are GG homozygotes, green dots are GA heterozygotes; there are no AA homozygotes (rs11209026 minor allele frequency (MAF) ~ 0.05). Cluster plot for each SNP was checked for clearly separated genotypes as a quality control.

SNPs included in the Study

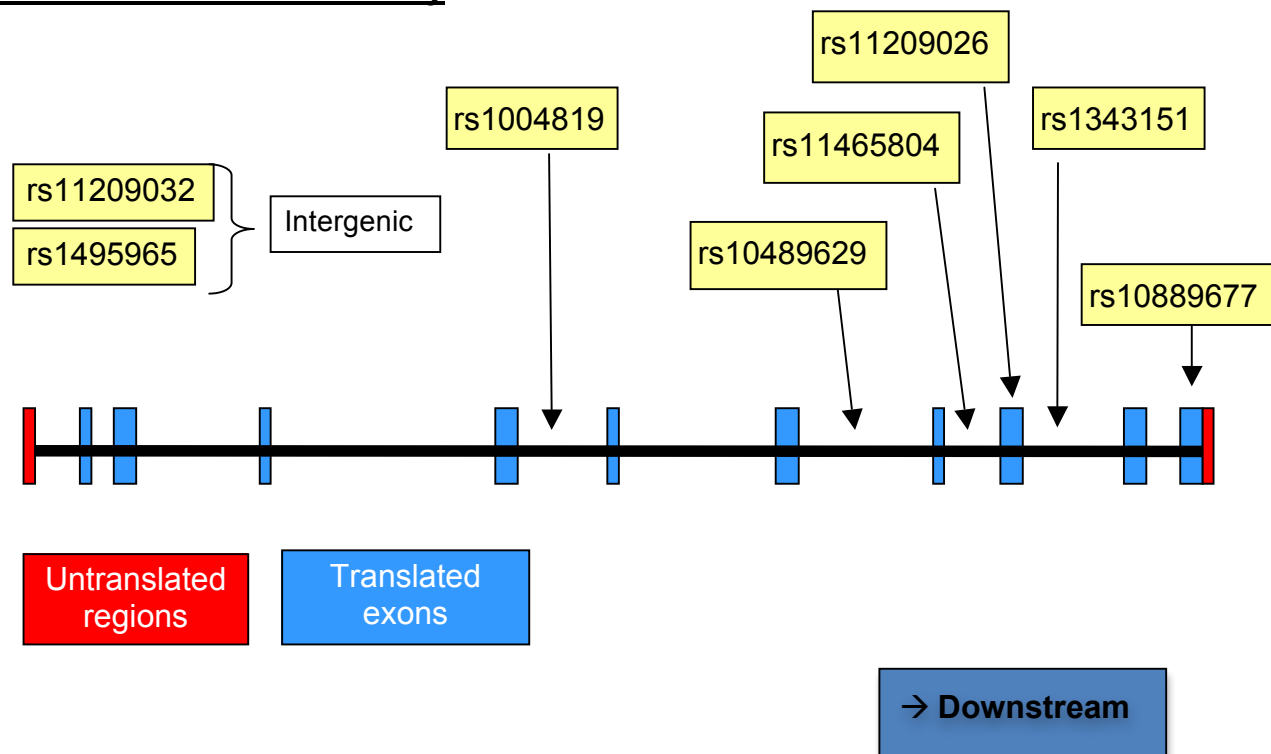


Figure 3.2. SNPs included in the *IL23R* study and their locations. *IL23R* is located at 1p31.3 and there are 10 exons of the gene. The intergenic SNPs (rs11209032 and rs1495965) are downstream of the gene.

3.2.2. Stratified Analysis

As *IL23R* is independently associated with both IBD and psoriasis, the analyses were performed both including and excluding AS cases known to have either of these conditions (excluded number of cases=90; cases remaining=640) (Franke *et al.* 2010; Strange *et al.* 2010; Anderson *et al.* 2011). In addition to the known AS cases with “clinical IBD”, it is also estimated that ~6% of cases with spondyloarhtropathies, including AS, have “subclinical IBD” (Mielants *et al.* 1996). However, such cases with “subclinical IBD” cannot not be determined without ileocolonoscopy/gut biopsy. Therefore, all AS cases without “clinical IBD” were included in this stratified analysis. The statistical power of the UK case-control study (number of cases=640) and the extended study (number of cases = 1728) slightly decreased to 86% and stayed at 98%, respectively, given an OR of 1.6, a MAF of 0.04, an alpha level of 0.05 (2-sided) and a disease prevalence of 4/1000 with a log-additive disease model.

3.2.3. Meta-analysis

A meta-analysis was conducted for *IL23R* based on the PubMed database using the terms “ankylosing spondylitis” and “interleukin-23R” (June 2008). Publications in English and conducted on populations of white European ancestry were included. The selection criteria were: 1) cases satisfying the modified New York criteria for AS; 2) controls derived from a population within the same geographic area and ethnic background as AS cases; 3) AS cases and controls not overlapping between studies.

Genotypes from up to 3482 cases and 3150 controls (depending on the numbers typed for particular SNPs in these studies) from this new series and previously published studies were combined in this meta-analysis (Burton *et al.* 2007; Rahman *et al.* 2008; Rueda *et al.* 2008). The pooled populations included in the meta-analysis originated from:

- Canada (Alberta, Newfoundland and Toronto as three separate strata presented in the original paper; cases=796, controls=742).
- Spain (cases=365, controls=500).
- UK (the Wellcome Trust case control consortium – WTCCC; UK cases combined=1703, controls=1331). The new UK data were combined with the UK cases from the original 2007 WTCCC-TASC study.
- USA (The Australo-Anglo-American Spondylitis Consortium – TASC; cases=618, controls=577).

Data were analysed using StatsDirect software. The DerSimonian-Laird test was used to calculate random effects pooled odds ratio, chi-square statistic and p-values in order to incorporate variation in true effects and to take into account the heterogeneity of the populations (See Chapter 2, Section 2.6.3). A p-value < 0.05 was considered statistically significant. Cochran Q and inconsistency (I^2) values were also reported as a measure of combinability and consistency of the studies. I^2 less than 50% indicates that less than half of the variation across the pooled studies is due to heterogeneity and they may therefore be combined. Similarly, an insignificant Cochran Q p-value (>0.05) indicates that the studies are combinable. The statistical power of the meta-analysis was 100% given an OR of 1.6, a MAF of 0.04, an alpha level of 0.05 (2-sided) and a disease prevalence of 4/1000 with a log-additive disease model.

3.3. Results

3.3.1. UK Case-Control Study & Stratified Analysis

The results of the new UK case-control study (Study I) and the extended UK case-control study (Study II) are shown in Tables 3.1 and 3.2. The results inclusive and exclusive of AS cases known to have concomitant IBD and/or psoriasis were similar. In the case-control study including only new cases and comparing them to the 58BBC controls, 4 of the 8 SNPs (rs11465804, rs11209026, rs10889677 and rs11209032) showed significant associations with AS, the strongest being with rs11209032 ($p=0.0037$, OR=1.2) when cases with IBD and/or psoriasis were excluded (Table 3.1, B). In the extended study, 7 of the 8 SNPs (rs1004819, rs11465804, rs11209026, rs1343151, rs10889677, rs11209032 and rs1495965) showed significant associations with AS, the strongest being with rs11209032 ($p=4.6 \times 10^{-6}$, OR=1.3) when cases with IBD and/or psoriasis were excluded (Table 3.2, D). For SNPs rs11465804 and rs11209026 excluding AS cases with IBD and/or psoriasis resulted in more highly significant associations with AS in both studies (Table 3.2).

	<i>A. Study I</i> <i>Case-control study:</i> <i>A. Including all new AS cases only (730)</i>							<i>B. Study II</i> <i>Extended study:</i> <i>C. Including all new AS cases (730) & WTCCC cases (1088)</i>					
<i>SNP</i>	<i>MINOR ALLELE</i>	<i>MAF CASES</i>	<i>MAF CONTROLS</i>	<i>CHI- SQUARE</i>	<i>P- VALUE</i>	<i>OR</i>	<i>95% CI</i>	<i>MAF CASES</i>	<i>MAF CONTROLS</i>	<i>CHI- SQUARE</i>	<i>P- VALUE</i>	<i>OR</i>	<i>95% CI</i>
rs1004819	A	0.33	0.3	3.12	0.077	1.13	0.99-1.30	0.33	0.3	9.83	0.0017	1.19	1.07-1.32
rs10489629	G	0.43	0.45	2.41	0.12	0.9	0.79-1.03	0.43	0.45	4.08	0.043	0.9	0.81-1.00
rs11465804	C	0.04	0.06	6.45	0.011	0.68	0.50-0.92	0.04	0.06	11.12	0.0009	0.68	0.54-0.85
rs11209026	A	0.05	0.06	5.41	0.02	0.71	0.53-0.95	0.05	0.06	12.74	0.0004	0.67	0.53-0.83
rs1343151	A	0.32	0.34	1.7	0.19	0.91	0.79-1.05	0.32	0.34	6.05	0.014	0.87	0.79-0.97
rs10889677	A	0.36	0.31	7.86	0.0051	1.21	1.06-1.39	0.36	0.31	12.91	0.0003	1.22	1.09-1.35
rs11209032	A	0.37	0.32	10.78	0.001	1.25	1.09-1.43	0.37	0.32	22.22	2.4x10 ⁻⁶	1.29	1.16-1.43
rs1495965	G	0.47	0.44	3.33	0.068	1.13	0.99-1.29	0.47	0.44	9.19	0.0024	1.17	1.06-1.29

Table 3.1. Results of the UK case-control study (study I), and the extended UK case-control study (study II) combining UK WTCCC cases and new UK cases. The combined sample of cases was compared with WTCCC (1958 Birth Cohort) controls (n=1331). (MAF: Minor allele frequency, OR: Odds ratio, CI: Confidence interval)

	C. Study I Case-control study: B. Excluding known cases with IBD and/or PS (640 remaining)							D. Study II Extended study: D. Excluding known cases with IBD and/or PS from the new cases only (1728 remaining)					
SNP	MINOR ALLELE	MAF CASES	MAF CONTROLS	CHI-SQUARE	P-VALUE	OR	95% CI	MAF CASES	MAF CONTROLS	CHI-SQUARE	P-VALUE	OR	95% CI
rs1004819	A	0.33	0.3	2.14	0.14	1.11	0.96-1.29	0.33	0.3	9.26	0.0023	1.18	1.06-1.32
rs10489629	G	0.43	0.45	1.56	0.21	0.92	0.80-1.05	0.43	0.45	3.56	0.059	0.91	0.82-1.00
rs11465804	C	0.04	0.06	7.88	0.005	0.63	0.45-0.87	0.04	0.06	12.21	0.0005	0.66	0.53-0.84
rs11209026	A	0.04	0.06	7.51	0.0061	0.64	0.46-0.88	0.04	0.06	14.66	0.0001	0.64	0.51-0.81
rs1343151	A	0.32	0.34	0.98	0.32	0.93	0.80-1.07	0.32	0.34	5.52	0.019	0.88	0.79-0.98
rs10889677	A	0.35	0.31	4.53	0.033	1.17	1.01-1.35	0.35	0.31	10.85	0.001	1.2	1.08-1.34
rs11209032	A	0.37	0.32	8.4	0.0037	1.23	1.07-1.42	0.37	0.32	20.99	4.6x10 ⁻⁶	1.28	1.15-1.43
rs1495965	G	0.47	0.44	1.68	0.19	1.1	0.95-1.26	0.47	0.44	8.01	0.0047	1.16	1.05 -1.28

Table 3.2. Results of the UK case-control study (study I) and the extended UK case-control study (study II) excluding cases with IBD and psoriasis. The combined sample of cases was compared with WTCCC (1958 Birth Cohort) controls (n=1331). (MAF: Minor allele frequency, OR: Odds ratio, CI: Confidence interval, PS: Psoriasis)

3.3.2. Meta-analysis

The meta-analysis showed significant associations between 7 *IL23R* SNPs and AS (Table 3.3). The strongest associations were seen with rs11209026 ($p=1.5 \times 10^{-9}$, OR=0.6) and rs11209032 ($p=4.06 \times 10^{-9}$, OR=1.2). However, the data for rs10889677 were non-combinable due to very different allele frequencies leading to an effect in the opposite direction (Cochran Q p-value <0.001 and inconsistency=75.6%) and have not therefore been presented in Table 3.3 (See Section 3.4.3 for a more detailed discussion of rs10889677). Forest plots showing meta-analysis odds ratios and 95% confidence intervals for each SNP are presented in Figures 3.3.1-3.3.4.

Table 3.3. Results of meta-analysis for *IL23R* polymorphisms, combining 4 published studies and new UK data. New UK case data and WTCCC case data were combined to form the UK stratum. Canadian data were included as 3 separate strata. No Spanish data were available for rs11465804. rs10889677 is not presented in the table due to non-combinability of the studies (See Section 3.4.6). (OR: Odds ratio, CI: Confidence interval, I²: Inconsistency, NS: Non-synonymous, #: number)

SNP	LOCATION	# OF CASES/ CONTROLS	MAF CASES	MAF CONTROLS	OR	95% CI	CHI SQUARE	P-VALUE	COCHRAN Q P-VALUE	I² (95% CI)
rs1004819	Intron 5	3482/3150	0.35	0.30	1.21	1.13 - 1.31	26.7	2.38 x 10 ⁻⁷	0.73	0% (0% - 61%)
rs10489629	Intron 7	3482/3150	0.42	0.46	0.85	0.77 - 0.94	9.4	0.0021	0.13	41.7% (0% - 75.5%)
rs11465804	Intron 8	3124/2650	0.04	0.06	0.62	0.50 - 0.76	20.4	6.3 x 10 ⁻⁶	0.29	18.2% (0% - 70.2%)
rs11209026	Exon 9 (NS)	3472/3150	0.04	0.06	0.61	0.52 - 0.72	36.6	1.5 x 10 ⁻⁹	0.55	0% (0% - 61%)
rs1343151	Intron 9	3478/3150	0.30	0.34	0.81	0.72 - 0.91	12.1	0.0005	0.08	49.1% (0% - 78%)
rs11209032	Intergenic	3483/3150	0.37	0.32	1.24	1.16 - 1.34	34.6	4.06 x 10 ⁻⁹	0.46	0% (0% - 61%)
rs1495965	Intergenic	3440/3150	0.48	0.44	1.19	1.10 - 1.29	18.8	1.5 x 10 ⁻⁵	0.32	15.2% (0% - 66.6%)

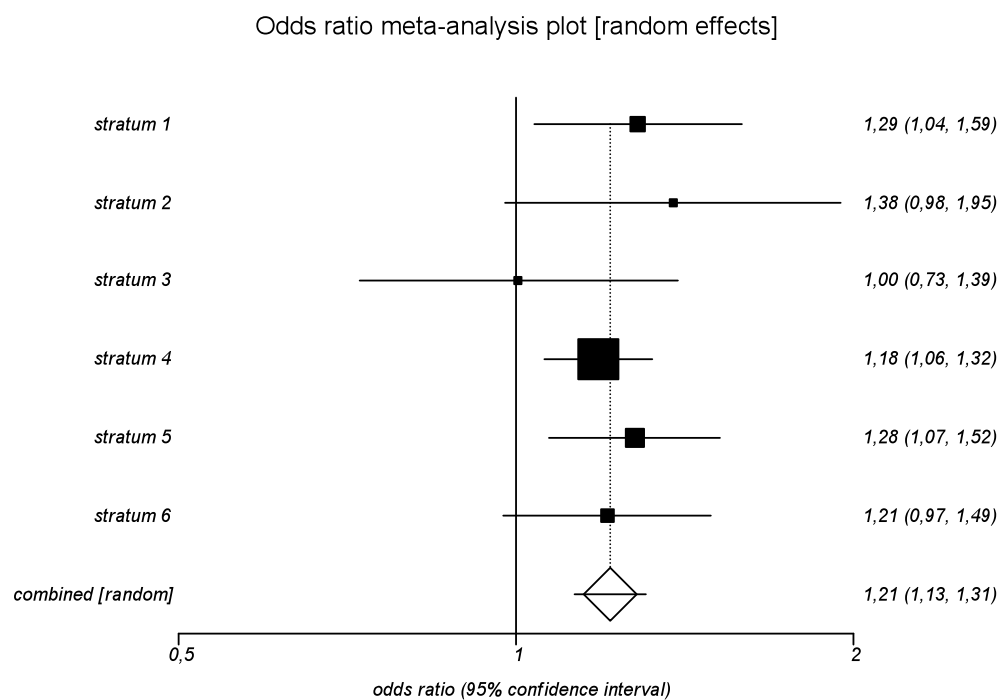
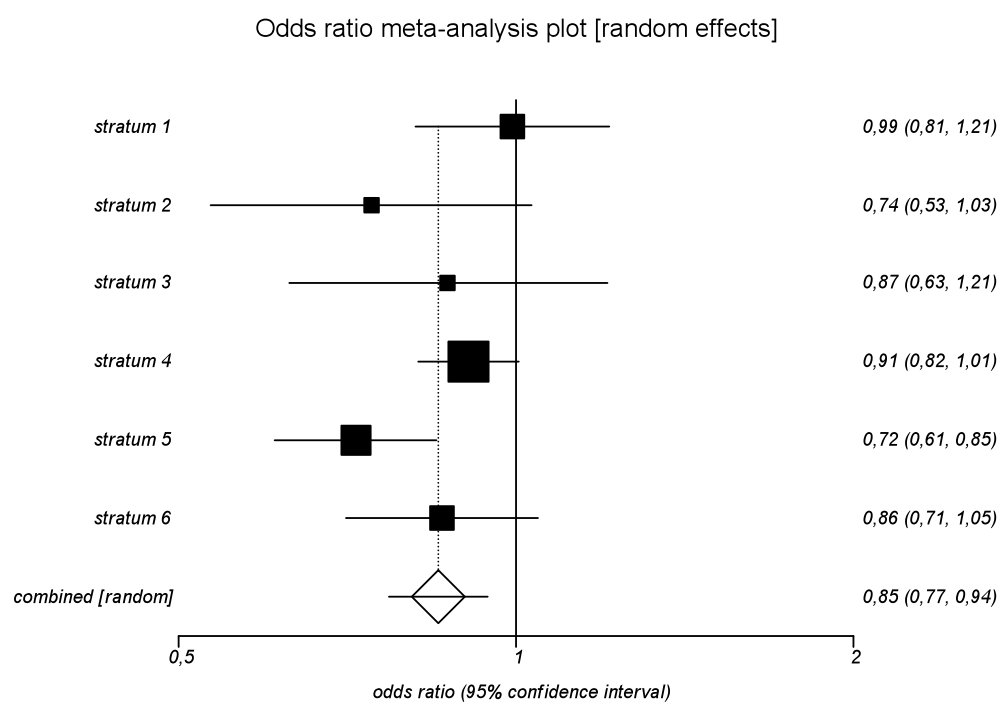
rs1004819**rs10489629**

Figure 3.3.1. Forest plots for the meta-analysis of *IL23R* SNPs. (Strata 1-6: Alberta, Newfoundland, Toronto, UK (case-control study and WTCCC I), USA, Spain, respectively).

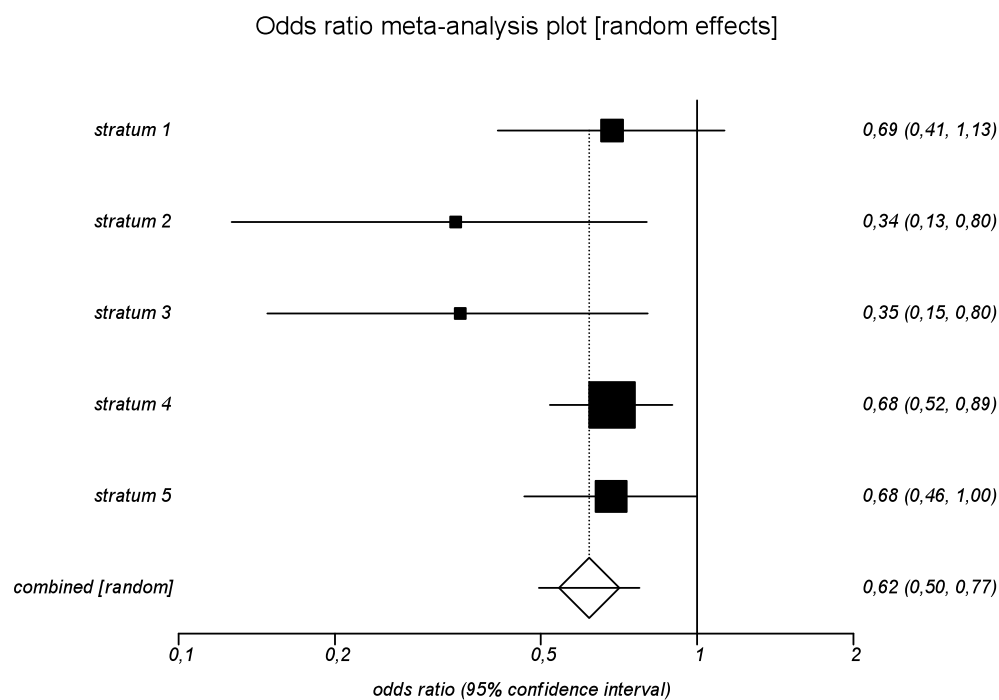
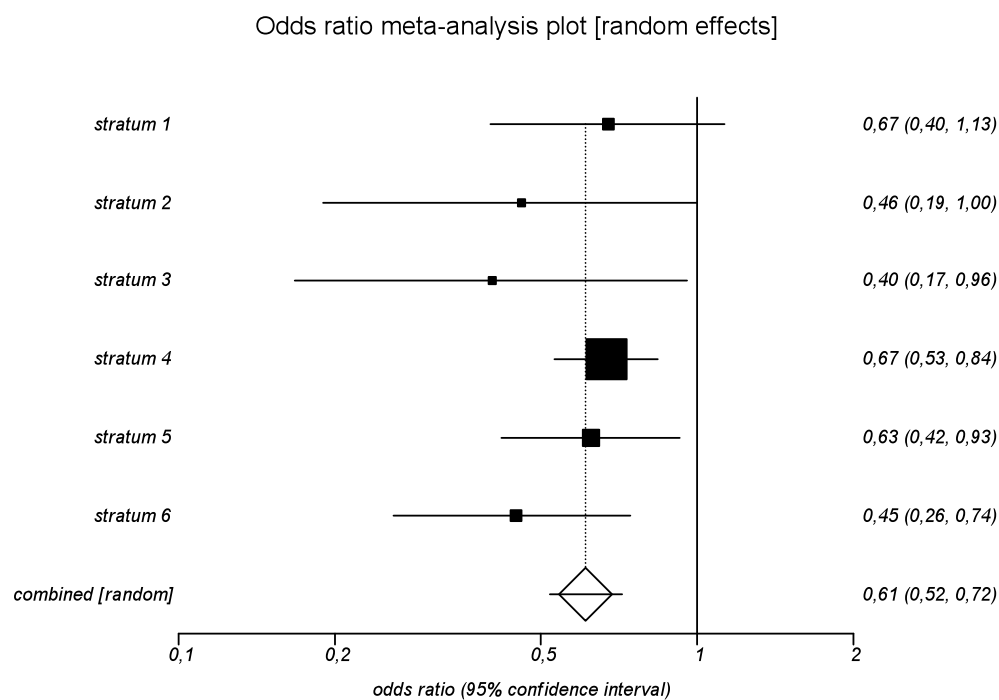
rs11465804**rs11209026**

Figure 3.3.2. Forest plots for the meta-analysis of *IL23R* SNPs. (Strata 1-6: Alberta, Newfoundland, Toronto, UK (case-control study and WTCCC I), USA, Spain, respectively).

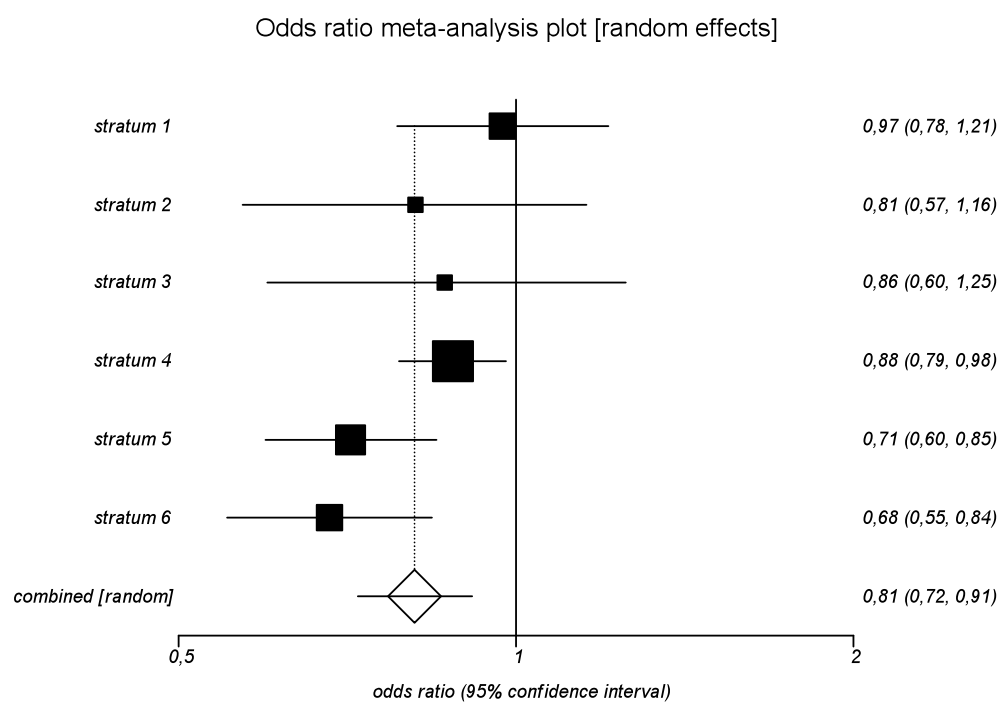
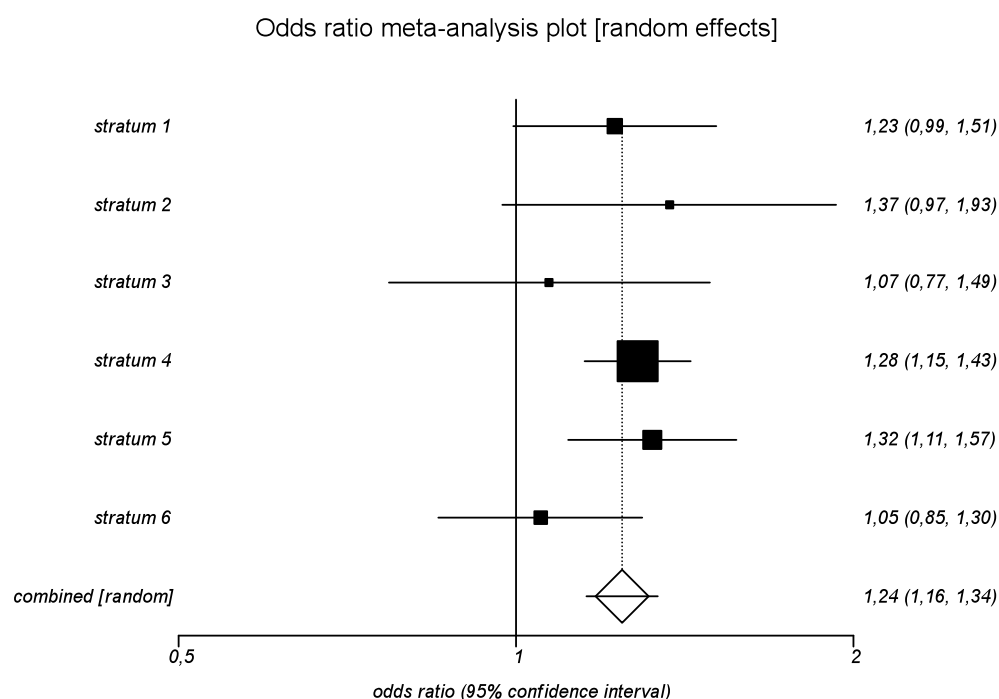
rs1343151**rs11209032**

Figure 3.3.3. Forest plots for the meta-analysis of *IL23R* SNPs. (Strata 1-6: Alberta, Newfoundland, Toronto, UK (case-control study and WTCCC I), USA, Spain, respectively).

rs1495965

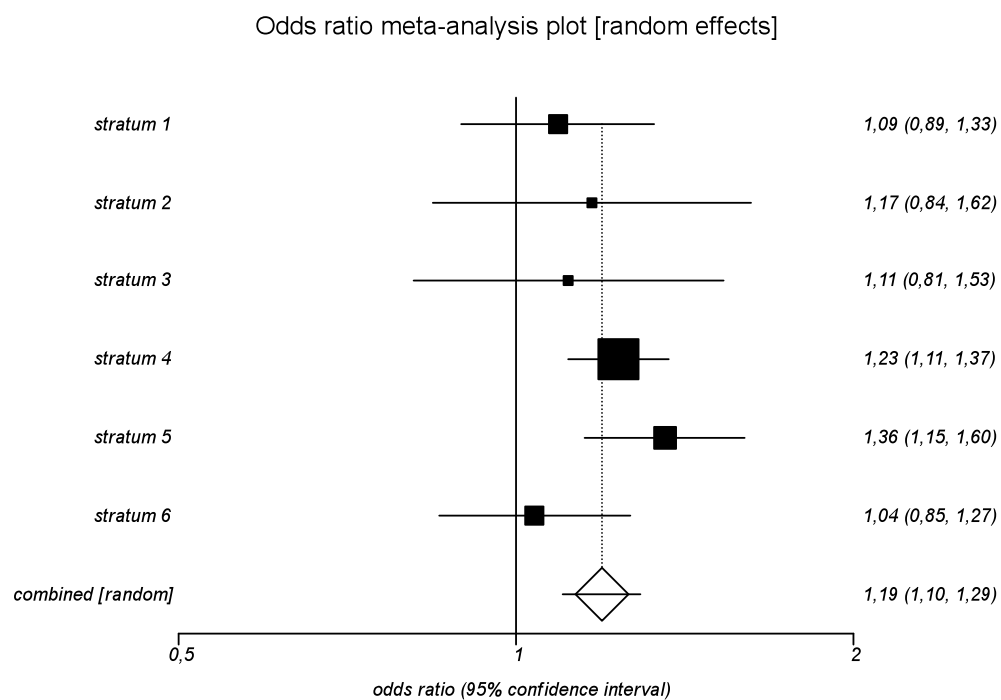


Figure 3.3.4. Forest plots for the meta-analysis of *IL23R* SNPs. (Strata 1-6: Alberta, Newfoundland, Toronto, UK (case-control study and WTCCC I), USA, Spain, respectively).

3.4. Discussion

3.4.1. UK Case-Control Study & Meta-analysis

In the relatively small replication study of 730 new cases compared to controls, I found weak associations with 4 SNPs in *IL23R* (rs11209032, OR=1.25, $p=0.001$). On combining the data with the WTCCC data for the UK cases in the extended study, all the SNPs tested showed significant associations, with rs11209032 being the most strongly associated (Table 3.1 B; OR=1.29, $p=2.4 \times 10^{-6}$). Compared to the results of the WTCCC-TASC study, there is a stronger evidence of association in the extended study in this chapter (WTCCC-TASC study: rs11209026, OR=1.6, $p=1.7 \times 10^{-3}$; extended study: rs11209026, OR=1.5, $p=0.0004$) (Table 3.1 B). In the UK studies reported here, the associations with rs11209026 and rs11465804 were actually stronger when AS cases with known IBD and psoriasis were excluded from the data set (Table 3.2 D; $p=0.0001$ and $p=0.0005$, respectively). This suggests that the association of AS with *IL23R* is not merely secondary to its association with IBD or psoriasis.

The meta-analysis followed the same trend as the WTCCC and TASC data sets with regard to the direction and magnitude of associations (Burton *et al.* 2007). This is perhaps unsurprising given the UK and TASC data sets were larger than the Spanish and Canadian data sets. Nevertheless, the forest plots show consistency in the associations of the most strongly associated SNPs in populations drawn from all four different geographical regions (Figures 3.3.1-3.3.4). Furthermore, evidence of association of this gene with AS is larger with lower p-values observed in the meta-analysis across 4 different populations. For instance, p-values for rs11209026 and rs11209032 were 1.5×10^{-9} (OR=0.61) and 4.06×10^{-9} (OR=1.24)

in the meta-analysis (Table 3.3). These are lower than those observed in the 2007 WTCCC-TASC GWAS (rs11209026, OR=0.61, $p=4.0 \times 10^{-6}$; rs11209032, OR=1.30, $p=7.5 \times 10^{-9}$) (Burton *et al.* 2007).

The uncommon allele (A) of the non-synonymous coding variant rs11209026 (Arg381Gln) is associated with protection against Crohn's disease (CD) and psoriasis (Duerr *et al.* 2006), as well as protection from AS (OR=0.61, Table 3.3). This SNP also showed one of the strongest associations in our extended UK study (Table 3.2 D, $p=0.0001$) along with rs11465804 (Table 3.2 D, $p=0.0005$), which is in linkage disequilibrium with rs11209026 ($r^2>0.8$). According to the Ensembl website, rs11209026 is predicted to be deleterious (SIFT score 0) and probably damaging (Polyphen score >0.980). rs11209026 (Arg381Gln) is within the cytoplasmic domain of IL23R, between the transmembrane domain and the JAK2 binding site. Therefore, it may influence the T_H17 response by affecting downstream signalling. In support of this is an *in vitro* study by Di Meglio *et al.* showing that T_H17 cells from 176 healthy donors of Western European descent carrying the protective allele A, produced less IL-17A and had reduced STAT3 phosphorylation in response to IL23 than the G allele carriers. There was no effect on T_H17 differentiation as might be expected from the primary role of IL23 in the expansion/maintenance rather than the differentiation of T_H17 cells (Di Meglio *et al.* 2011). Similar results were observed by Pidasheva *et al.*, who showed that in IL23-stimulated T cells from healthy Caucasian donors, the level of STAT3 phosphorylation was less in T cells from IL23R^{Gln381} individuals compared to IL23R^{Arg381} individuals, which carry the ancestral G allele. Likewise, in a study of rheumatoid arthritis (RA), it was concluded that the Arg381Gln substitution may

affect the serum IL17A concentrations, causing IL23R^{Gln381} RA cases to require higher serum IL23 concentrations to produce similar levels of IL17A compared to those in IL23R^{Arg381} cases (Hazlett *et al.* 2012). These studies suggest that the Arg381Gln is a loss-of-function mutation that confers a degree of protection against CD (Pidashveva *et al.* 2011), and potentially for AS as well. Higher levels of serum IL23 and IL17 have been detected in AS patients compared to healthy controls (Mei *et al.* 2011). These cytokine levels may also correlate with the Arg381Gln substitution. However, this conclusion requires further studies in AS.

3.4.2. Recent Studies on *IL23R* & AS

Associations at *IL23R*, previously identified by the WTCCC1 and TASC studies, were also confirmed in the 2010 TASC GWAS (rs11209026, $p=9.1 \times 10^{-14}$) (Burton *et al.* 2007; Reveille *et al.* 2010). In the most recent Wellcome Trust study (WTCCC2), confirmed association (genomewide $p < 5 \times 10^{-8}$) was seen with SNPs close to and within *IL23R* (Evans *et al.* 2011). Two independent signals were observed in *IL23R* in the WTCCC2 study after conditional logistic regression analysis. rs11209026 and rs11209032, which also showed the strongest associations in the meta-analysis (Table 3.3), were independently associated with AS in the WTCCC2 study. This independent association is also apparent from the weak linkage disequilibrium between these two SNPs ($r^2=0.032$ (HapMap release 22)) (See Chapter 5 for a more detailed analysis and discussion of the 2 independent signals in this gene).

In Han Chinese, there is apparently no association between AS and *IL23R* (Davidson *et al.* 2009; Davidson *et al.* 2011). However, this is unsurprising since

the most strongly associated SNP in Caucasians, rs11209026 (Arg381Gln substitution in exon 9 of *IL23R*) is monomorphic in Han Chinese. This finding supports the hypothesis that it is the common allele of this non-synonymous coding SNP that is largely responsible for the increased susceptibility to AS, and it is likely that rs11209026 is a functional variant. Subsequent to these studies, an association between *IL23R* and AS has also been reported in a small Portuguese study (rs1004819, OR~1.4, $p=0.0049$), further increasing the evidence that this gene has a role in AS (Pimentel-Santos *et al.* 2009).

3.4.3. Statistical Considerations

This new UK case-control study, the extended study and the meta-analysis had statistical power of 89%, 98% and 100% respectively, given an OR of 1.6, an alpha level of 0.05, a MAF of 0.04 and a disease prevalence of 0.4%. An advantage of meta-analysis is the increased statistical power gained by combining different studies. However, it is important to consider the possibility of heterogeneity between different studies and assess the results of the meta-analysis in the face of this heterogeneity. Furthermore, it is possible to take into account population heterogeneity at the time of data analysis and use appropriate statistical methods.

In the *IL23R* meta-analysis, the DerSimonian-Laird (DSL) test was used to calculate random effects pooled odds ratios, while Cochran Q and inconsistency (I^2) values were calculated for each SNP to determine combinability of the studies (Table 3.3). The DSL test does not make the assumption that all the individual studies were carried out under similar conditions with similar statistical powers and allows the effect size (OR) to vary. The varying sizes of study populations are presented in the

forest plots where sizes of the black boxes correspond to individual study population sizes (Figure 3.3.1-3.3.4). As a study population becomes smaller, confidence in the observed result is reduced leading to wider confidence intervals. Moreover, it is crucial to consider the direction and magnitude of associations, with overlapping confidence intervals, and make sure that they are in agreement for each study. This agreement is also presented in the meta-analysis forest plots for each SNP in this chapter (Figure 3.3.1-3.3.4). These and other important points to consider in meta-analysis and case-control studies have already been discussed in Chapter 2 (Section 2.6.3).

Meta-analysis result of rs10889677 was not presented in Table 3.3 due to non-combinability of the studies with a significant Cochran's Q p-value ($p=0.001$) and an I^2 value of 75.9%, indicating heterogeneity (Figure 3.4). Such heterogeneity occurs as a result of different allele frequencies across populations in the meta-analysis. Publication errors might also lead to such inconsistencies between allele frequencies reported. In the case of rs10889677, sample populations from Toronto and Spain had different allele frequencies resulting in ORs of 0.94 and 0.81, respectively. All allele frequencies from other sample populations lead to ORs greater than 1. As a result of this, random effects pooled OR was 1.14 with a p-value of 0.14.

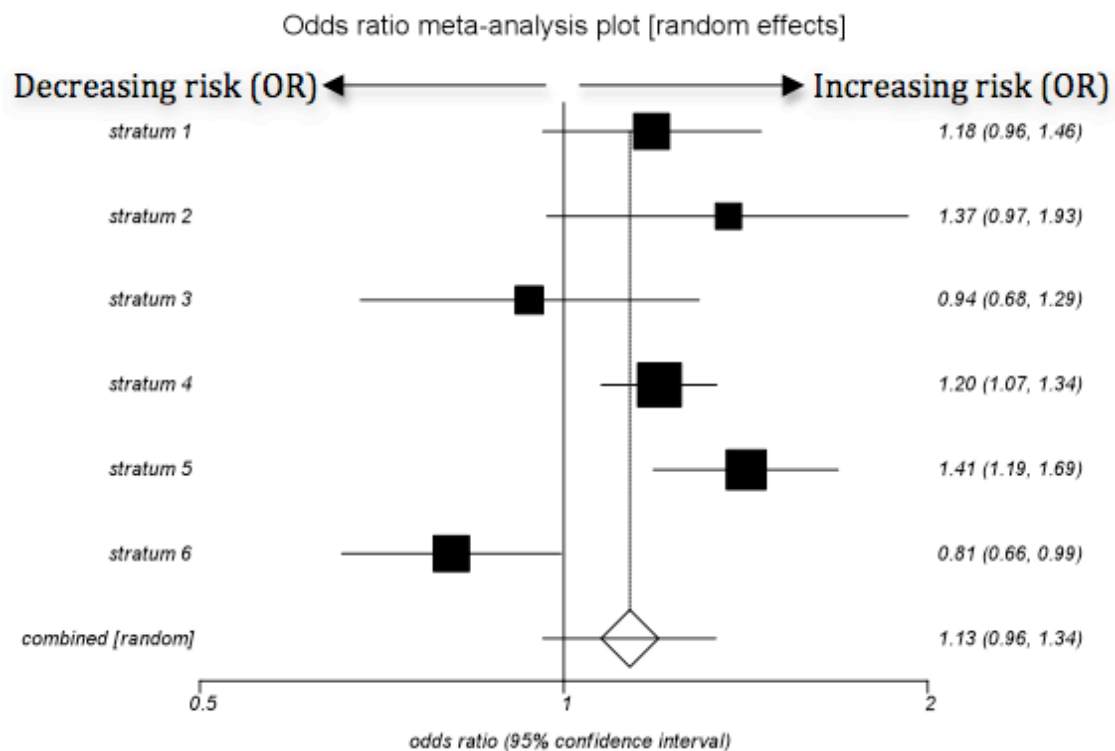


Figure 3.4. Forest plots for the meta-analysis of *IL23R* rs10889677 showing the heterogeneity across the studies. Direction of association (OR) for strata 3 (Toronto) and 6 (Spain) is reversed. (Strata 1-6: Alberta, Newfoundland, Toronto, UK (case-control study and WTCCC I), USA, Spain, respectively).

3.4.4. T_H17 T Cell Subset

IL23R is the *IL23R*-specific component of the heterodimeric receptor for the cytokine *IL23*. The other subunit, *IL12R β 1*, is also a component of the *IL12* receptor (Parham *et al.* 2002). The *IL23* receptor is expressed on the T_H17 lymphocyte subset that mediates inflammation in several animal models of autoimmunity (Cua *et al.* 2003; Murphy *et al.* 2003). T_H17 lymphocytes play an important role in defence against extracellular pathogens, such as *Candida* and

Klebsiella, and have a role in tissue damage, as demonstrated in experimental allergic/autoimmune encephalomyelitis (Steinman 2008; Maddur *et al.* 2012).

IL23R is also expressed on natural killer cells and, at lower levels, on monocytes, macrophages and dendritic cells (Parham *et al.* 2002; Zhu & Qian 2012).

The cytokines TGF β and IL6 are involved in early T_H17 differentiation, while IL21 and IL23 are mainly necessary for the amplification and maintenance of the T_H17 T cell subset, respectively (Luckheeram *et al.* 2012; Maddur *et al.* 2012). The effector cytokines produced by T_H17 T cells include IL21, IL22, IL17A/F and the chemokine CCL20 (Ouyang *et al.* 2008; Maddur *et al.* 2012). Retinoic-acid-receptor-related orphan receptor gamma-T (ROR γ t; RORC in humans) is the key transcription factor for the T_H17 T cell subset (Luckheeram *et al.* 2012).

Signal transducer and activator of transcription 3 (STAT3) is also a key signalling factor in the T_H17 pathway. Activation of IL23R by its ligand IL23 induces signaling through the JAK2/STAT3 pathway. JAK2/STAT3 are also activated by the IL6 and IL21 pathways, which are important for T_H17 T cell differentiation (Yang *et al.* 2007; Maddur *et al.* 2012). In IL23 signaling, JAK2 phosphorylates the IL23R subunit of the IL23 receptor at the tyrosine 705 residue upon ligand binding, recruiting STAT3 to the receptor complex. STAT3 is then phosphorylated by JAK2. Phosphorylated STAT3 homodimerizes and translocates to the nucleus to trigger the expression of cytokines IL17A, IL17F, IL22, IL21 and chemokine CCL20 in the T_H17 cells (Figure 3.5) (Altshuler *et al.* 2008; Ouyang *et al.* 2008). Furthermore, TYK2 binds to IL12R β 1 subunit of the IL23 receptor. In the T_H17 pathway, STAT3 phosphorylation by JAK2 dominates over STAT4 activation of TYK2 that is

essential for a T_H1 response (Ghoreschi *et al.* 2009). The fact that several of these factors, such as STAT3, IL12B and TYK2 have now been shown to be associated with AS and related conditions, such as IBD and psoriasis, highlights the pathogenic importance of T_H17 cells in these conditions. These are discussed in further detail below (See Section 3.4.5).

Regulation of expression of ROR γ t (RORC in humans) by STAT3 is also necessary for subsequent T_H17 T cell differentiation (Yang *et al.* 2007). STAT3 acts together with interferon regulatory factor 4 (IRF4) and Smad family of signal transduction proteins (SMADs) to promote expression of *RORC*. Moreover, RORC, in combination with STAT3 and runt-related transcription factor 1 (RUNX1), induces the expression of the effector cytokines *IL17A/F*, *IL21*, *IL22* and the chemokine *CCL20* (Figure 3.5) (Maddur *et al.* 2012). Thus, IL23R is an important player in T_H17 cytokine production, as well as in the proliferation and survival of the T_H17 T cell subset (Oppmann *et al.* 2000; Aggarwal *et al.* 2003; Harrington *et al.* 2005; Maddur *et al.* 2012).

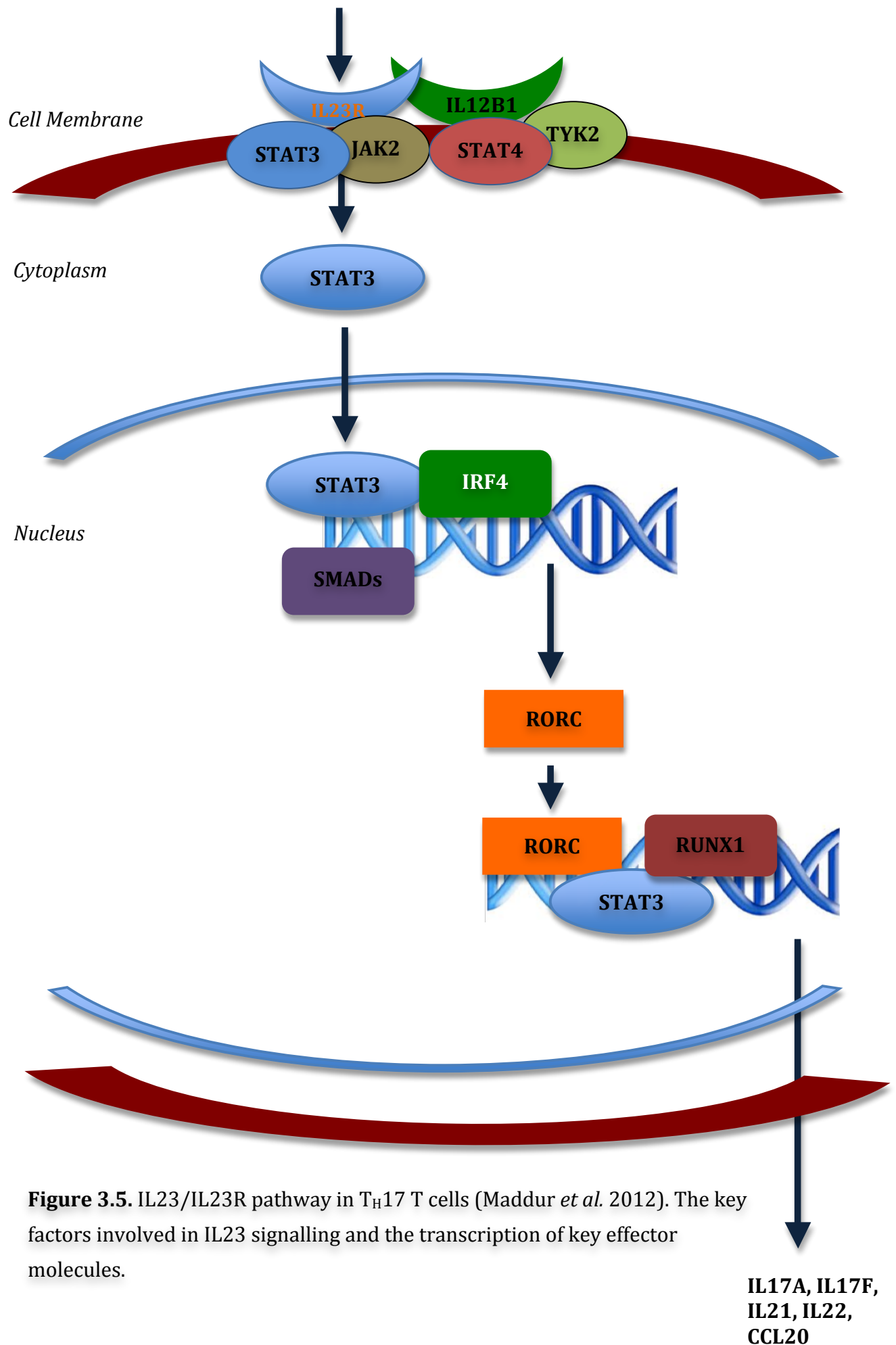


Figure 3.5. IL23/IL23R pathway in T_H17 T cells (Maddur *et al.* 2012). The key factors involved in IL23 signalling and the transcription of key effector molecules.

3.4.5. Related Inflammatory Diseases associated with the T_H17

Pathway: Psoriasis & Inflammatory Bowel Disease

About 15% of AS patients also have psoriasis and/or IBD. Recent studies of psoriasis and IBD (CD/UC) give insights into genetic factors shared with AS (Barrett *et al.* 2008; Barrett *et al.* 2009; Nair *et al.* 2009; Silverberg *et al.* 2009; Franke *et al.* 2010; Strange *et al.* 2010; Anderson *et al.* 2011; Cho & Brant 2011). Strong genetic associations have been observed between these inflammatory diseases and T_H17 pathway genes, such as *IL23R*, *IL12B*, *JAK2*, *TYK2*, *CCR6* and *STAT3* (Table 3.4) (Abraham & Cho 2009). The list in Table 3.4 includes 7 genes involved in the T_H17 pathway highlighted in recent genetic association studies of various related inflammatory diseases.

In addition to the studies listed in Table 3.4, the association between AS and *STAT3* has also been replicated in a study of Han Chinese (rs2293152 OR=0.81, p=0.0015) (Davidson *et al.* 2011). Furthermore, recently Cohen *et al.* published a sufficiently powered study (power>90% to detect an OR of 1.1) that also suggests no genetic association between AS and the RA-associated SNP rs3093024 in *CCR6* (Cohen *et al.* 2012).

Gene	SNP	Disease	P-value	OR	Genetic Study
IL23R (Chr1)	rs11209026	AS	4.2×10^{-7}	1.60	GWAS; WTCCC 2, 2011
	rs11209026	PS	7.13×10^{-7}	1.49	GWAS; WTCCC 2, 2010
	rs11209026	UC	3.0×10^{-10}	NR	GWAS; WTCCC 2, 2009
	rs11465804	CD	1.0×10^{-64}	2.66	GWAS meta-analysis; Franke <i>et al.</i> 2010
IL12B (Chr5)	rs6556416	AS	1.0×10^{-4}	1.18	GWAS; WTCCC 2, 2011
	rs3213094	PS	4.93×10^{-11}	1.39	GWAS; WTCCC 2, 2010
	rs1368438	UC	0.0039	NR	GWAS; WTCCC 2, 2009
	rs6556412	CD	5.37×10^{-14}	1.18	GWAS meta-analysis; Franke <i>et al.</i> 2010
STAT3 (Chr17)	rs744166	AS	2.6×10^{-5}	0.84	Danoy <i>et al.</i> 2010 (UK-USA-Australia)
	-	PS	-	-	GWAS: WTCCC 2, 2010
	rs744166	UC	0.0025	NR	GWAS; WTCCC 2, 2011
	rs11871801	CD	2.51×10^{-8}	1.15	GWAS meta-analysis; Franke <i>et al.</i> 2010
CCR6 (Chr6)	-	AS	-	-	GWAS; WTCCC 2, 2011
	-	PS	-	-	GWAS; WTCCC 2, 2010
	-	UC	-	-	GWAS; WTCCC 2, 2009
	rs415890	CD	2.51×10^{-12}	1.17	GWAS meta-analysis; Franke <i>et al.</i> 2010
JAK2 (Chr9)	rs10758669	AS	2.9×10^{-4}	1.14	Danoy <i>et al.</i> 2010 (UK-USA-Australia)
	-	PS	-	-	GWAS: WTCCC 2, 2010
	rs10974914	UC	1.5×10^{-5}	NR	GWAS; WTCCC 2, 2011
	rs10758669	CD	1.00×10^{-13}	1.18	GWAS meta-analysis; Franke <i>et al.</i> 2010
TYK2 (Chr19)	-	AS	-	-	GWAS; WTCCC 2, 2011
	rs12720356	PS	8.82×10^{-7}	1.40	GWAS; WTCCC 2, 2010
	NR	UC	$<5 \times 10^{-4}$	NR	Review by Lees <i>et al.</i> , 2011
	rs12720356	CD	9.20×10^{-10}	1.12	GWAS meta-analysis; Franke <i>et al.</i> 2010
IL23A (Chr12)	-	AS	-	-	GWAS; WTCCC 2, 2011
	rs2066808	PS	2.49×10^{-7}	1.49	GWAS; WTCCC 2, 2010
	-	UC	-	-	GWAS; WTCCC 2, 2009
	-	CD	-	-	GWAS meta-analysis; Franke <i>et al.</i> 2010

Table 3.4. Genes involved in the T_H17 pathway that have been associated with AS, PS, UC and CD in various genetic association studies. (Chr: chromosome; NR: not reported)

It is still not apparent which of these disease-associated SNPs are truly functional. Functional studies are necessary to determine and confirm truly causal variants. In a functional study of IBD, it has been shown that rs10889677 (C→A), which is located in the 3' untranslated region (UTR) of *IL23R*, affects both the mRNA and protein levels. The mutant A allele leads to loss of regulation by microRNAs, Let-7e and Let-7f, and thus, results in higher levels of *IL23R*, which may cause increased inflammation (Zwiers *et al.* 2012). rs10889677 was also associated with AS in this

case-control study (Table 3.1; $p=0.0003$) and the A allele of this SNP increased the risk of AS (OR=1.2), similar to IBD. This functional study of IBD has provided clues as to which of the AS-associated variants in *IL23R* are truly causal. However, it is essential to replicate these findings, which may lead to development of a new treatment, in AS as well.

3.4.6. T_H17 Pathway as a Therapeutic Target

IL23R is particularly relevant to the expansion and maintenance of the T_H17 subset of CD4⁺ lymphocytes (Stritesky *et al.* 2008). Monoclonal antibodies, capable of blocking the IL17 produced by these T_H17 cells, have been fast-tracked in clinical trials in AS (now also in Phase II trials for MS, CD, rheumatoid arthritis, psoriasis) (Lubberts *et al.* 2004; Balague *et al.* 2009; Zhu & Qian 2012), partly as a result of the genetic association of *IL23R* with AS. Thus, it has already become apparent that more precise definition of the genetic associations of AS can potentially lead to relatively rapid translation to new forms of therapy targeting this pathway.

The signalling pathways involving the pro-inflammatory cytokine IL23 and its receptor IL23R are important potential therapeutic targets in AS, as indicated by studies in mouse models of collagen-induced arthritis and experimental autoimmune encephalomyelitis (Cua *et al.* 2003; Murphy *et al.* 2003). A monoclonal antibody against the p40 subunit of both IL23 and IL12 is effective in treatment of psoriasis, psoriatic arthritis and CD (Sandborn *et al.* 2008; Gottlieb *et al.* 2009; Griffiths *et al.* 2010; Ryan *et al.* 2010). This promising therapy has not been tried in AS yet. Furthermore, other members of the T_H17 pathway such as IL17R, IL6, IL6R, CCR6 and CCL20 are all potential therapeutic targets in AS and

other autoimmune diseases. Blocking differentiation and expansion of the T_H17 T cells via IL6R, IL6 and IL21, inhibiting/neutralising the cytokines/chemokines produced by the T_H17 cells and their receptors such as IL17R, CCL20 and its receptor CCR6, as well as inhibiting T_H17 T cell-specific transcription factors such as RORC could be possible methods for targeting this pathway (Maddur *et al.* 2012; Zhu & Qian 2012). For instance, IL6 is important for the balance between T_H17 T cells and regulatory T (T_{REG}) cells (Kimura & Kishimoto 2010). Therefore, using anti-IL6 molecules may be useful in inhibiting the T_H17 response and promoting the T_{REG} response. This therapy has been approved for use in rheumatoid arthritis. However, there have been no concluding results from experiments in AS yet (Rajalingham & Das 2012).

There is now substantial confirmatory evidence that variants in *IL23R* are associated with AS and other inflammatory diseases across populations. However, understanding the functional consequences of these observed associations still require investigation. One of the AS associations is with a low frequency non-synonymous coding variant (rs11209026), for which there is evidence that causes loss of function and therefore, protection against AS. But, the possible functional consequence of the second strong signal (rs11209032) in *IL23R* is less clear. rs11209032 may be tagging another lower frequency coding SNP that has not been included in any of the genetic studies yet, or may have a regulatory function that has not been identified. Thus, substantial analysis and possible re-sequencing of *IL23R*, its regulatory sequences and flanking regions to identify novel variants that are tagged by SNPs associated with AS is important for better understanding of the functional consequences of these associations.

3.5. References

- Abraham, C. & Cho, J. H. (2009). "IL-23 and autoimmunity: new insights into the pathogenesis of inflammatory bowel disease." *Annu Rev Med* **60**: 97-110.
- Aggarwal, S., Ghilardi, N., *et al.* (2003). "Interleukin-23 promotes a distinct CD4 T cell activation state characterized by the production of interleukin-17." *J Biol Chem* **278**(3): 1910-1914.
- Altshuler, D., Daly, M. J., *et al.* (2008). "Genetic mapping in human disease." *Science* **322**(5903): 881-888.
- Anderson, C. A., Boucher, G., *et al.* (2011). "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47." *Nat Genet* **43**(3): 246-252.
- Balague, C., Kunkel, S. L., *et al.* (2009). "Understanding autoimmune disease: new targets for drug discovery." *Drug Discov Today* **14**(19-20): 926-934.
- Barrett, J. C., Hansoul, S., *et al.* (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease." *Nat Genet* **40**(8): 955-962.
- Barrett, J. C., Lee, J. C., *et al.* (2009). "Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region." *Nat Genet* **41**(12): 1330-1334.
- Burton, P. R., Clayton, D. G., *et al.* (2007). "Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants." *Nat Genet* **39**(11): 1329-1337.
- Cargill, M., Schrodi, S. J., *et al.* (2007). "A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes." *Am J Hum Genet* **80**(2): 273-290.
- Cho, J. H. & Brant, S. R. (2011). "Recent insights into the genetics of inflammatory bowel disease." *Gastroenterology* **140**(6): 1704-1712.
- Cohen, C. J., Karaderi, T., *et al.* (2012). "A CCR6 variant strongly associated with rheumatoid arthritis in two populations is not associated with ankylosing spondylitis." *Rheumatol Int*.
- Cua, D. J., Sherlock, J., *et al.* (2003). "Interleukin-23 rather than interleukin-12 is the critical cytokine for autoimmune inflammation of the brain." *Nature* **421**(6924): 744-748.
- Davidson, S. I., Liu, Y., *et al.* (2011). "Association of STAT3 and TNFRSF1A with ankylosing spondylitis in Han Chinese." *Ann Rheum Dis* **70**(2): 289-292.
- Davidson, S. I., Wu, X., *et al.* (2009). "Association of ERAP1, but not IL23R, with ankylosing spondylitis in a Han Chinese population." *Arthritis Rheum* **60**(11): 3263-3268.
- Di Meglio, P., Di Cesare, A., *et al.* (2011). "The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced Th17 effector response in humans." *PLoS One* **6**(2): e17160.
- Duerr, R. H., Taylor, K. D., *et al.* (2006). "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene." *Science* **314**(5804): 1461-1463.
- El Maghraoui, A. (2011). "Extra-articular manifestations of ankylosing spondylitis: prevalence, characteristics and therapeutic implications." *Eur J Intern Med* **22**(6): 554-560.

- Evans, D. M., Spencer, C. C., *et al.* (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**(8): 761-767.
- Feldtkeller, E., Khan, M. A., *et al.* (2003). "Age at disease onset and diagnosis delay in HLA-B27 negative vs. positive patients with ankylosing spondylitis." *Rheumatol Int* **23**(2): 61-66.
- Franke, A., McGovern, D. P., *et al.* (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." *Nat Genet* **42**(12): 1118-1125.
- Ghoreschi, K., Laurence, A., *et al.* (2009). "Janus kinases in immune cell signaling." *Immunol Rev* **228**(1): 273-287.
- Gottlieb, A., Menter, A., *et al.* (2009). "Ustekinumab, a human interleukin 12/23 monoclonal antibody, for psoriatic arthritis: randomised, double-blind, placebo-controlled, crossover trial." *Lancet* **373**(9664): 633-640.
- Griffiths, C. E., Strober, B. E., *et al.* (2010). "Comparison of ustekinumab and etanercept for moderate-to-severe psoriasis." *N Engl J Med* **362**(2): 118-128.
- Harrington, L. E., Hatton, R. D., *et al.* (2005). "Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages." *Nat Immunol* **6**(11): 1123-1132.
- Hazlett, J., Stamp, L. K., *et al.* (2012). "IL-23R rs11209026 polymorphism modulates IL-17A expression in patients with rheumatoid arthritis." *Genes Immun* **13**(3): 282-287.
- Kimura, A. & Kishimoto, T. (2010). "IL-6: regulator of Treg/Th17 balance." *Eur J Immunol* **40**(7): 1830-1835.
- Lubberts, E., Koenders, M. I., *et al.* (2004). "Treatment with a neutralizing anti-murine interleukin-17 antibody after the onset of collagen-induced arthritis reduces joint inflammation, cartilage destruction, and bone erosion." *Arthritis Rheum* **50**(2): 650-659.
- Luckheeram, R. V., Zhou, R., *et al.* (2012). "CD4(+)T Cells: Differentiation and Functions." *Clin Dev Immunol* **2012**: 925135.
- Maddur, M. S., Miossec, P., *et al.* (2012). "Th17 cells: biology, pathogenesis of autoimmune and inflammatory diseases, and therapeutic strategies." *Am J Pathol* **181**(1): 8-18.
- Mei, Y., Pan, F., *et al.* (2011). "Increased serum IL-17 and IL-23 in the patient with ankylosing spondylitis." *Clin Rheumatol* **30**(2): 269-273.
- Mielants, H., De Vos, M., *et al.* (1996). "The role of gut inflammation in the pathogenesis of spondyloarthropathies." *Acta Clin Belg* **51**(5): 340-349.
- Murphy, C. A., Langrish, C. L., *et al.* (2003). "Divergent pro- and antiinflammatory roles for IL-23 and IL-12 in joint autoimmune inflammation." *J Exp Med* **198**(12): 1951-1957.
- Nair, R. P., Duffin, K. C., *et al.* (2009). "Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways." *Nat Genet* **41**(2): 199-204.
- Nunez, C., Dema, B., *et al.* (2008). "IL23R: a susceptibility locus for celiac disease and multiple sclerosis?" *Genes Immun*.
- Oppmann, B., Lesley, R., *et al.* (2000). "Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12." *Immunity* **13**(5): 715-725.
- Ouyang, W., Kolls, J. K., *et al.* (2008). "The biological functions of T helper 17 cell effector cytokines in inflammation." *Immunity* **28**(4): 454-467.

- Parham, C., Chirica, M., *et al.* (2002). "A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rbeta1 and a novel cytokine receptor subunit, IL-23R." *J Immunol* **168**(11): 5699-5708.
- Patsopoulos, N. A., Esposito, F., *et al.* (2011). "Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci." *Ann Neurol* **70**(6): 897-912.
- Pidasheva, S., Trifari, S., *et al.* (2011). "Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q." *PLoS One* **6**(10): e25038.
- Pimentel-Santos, F. M., Ligeiro, D., *et al.* (2009). "Association of IL23R and ERAP1 genes with ankylosing spondylitis in a Portuguese population." *Clin Exp Rheumatol* **27**(5): 800-806.
- Rahman, P., Inman, R. D., *et al.* (2008). "Association of interleukin-23 receptor variants with ankylosing spondylitis." *Arthritis Rheum* **58**(4): 1020-1025.
- Rajalingham, S. & Das, S. (2012). "Antagonizing IL-6 in ankylosing spondylitis: a short review." *Inflamm Allergy Drug Targets* **11**(4): 262-265.
- Reveille, J. D., Sims, A. M., *et al.* (2010). "Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci." *Nat Genet* **42**(2): 123-127.
- Rueda, B., Orozco, G., *et al.* (2008). "The IL23R Arg381Gln non-synonymous polymorphism confers susceptibility to ankylosing spondylitis." *Ann Rheum Dis*.
- Ryan, C., Thrash, B., *et al.* (2010). "The use of ustekinumab in autoimmune disease." *Expert Opin Biol Ther* **10**(4): 587-604.
- Sandborn, W. J., Feagan, B. G., *et al.* (2008). "A randomized trial of Ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with moderate-to-severe Crohn's disease." *Gastroenterology* **135**(4): 1130-1141.
- Silverberg, M. S., Cho, J. H., *et al.* (2009). "Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study." *Nat Genet* **41**(2): 216-220.
- Steinman, L. (2008). "A rush to judgment on Th17." *J Exp Med* **205**(7): 1517-1522.
- Strange, A., Capon, F., *et al.* (2010). "A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1." *Nat Genet* **42**(11): 985-990.
- Stritesky, G. L., Yeh, N., *et al.* (2008). "IL-23 promotes maintenance but not commitment to the Th17 lineage." *J Immunol* **181**(9): 5948-5955.
- Yang, X. O., Panopoulos, A. D., *et al.* (2007). "STAT3 regulates cytokine-mediated generation of inflammatory helper T cells." *J Biol Chem* **282**(13): 9358-9363.
- Zhu, S. & Qian, Y. (2012). "IL-17/IL-17 receptor system in autoimmune disease: mechanisms and therapeutic potential." *Clin Sci (Lond)* **122**(11): 487-511.
- Zwiers, A., Kraal, L., *et al.* (2012). "Cutting edge: a variant of the IL-23R gene associated with inflammatory bowel disease induces loss of microRNA regulation and enhanced protein production." *J Immunol* **188**(4): 1573-1577.

CHAPTER 4

4. INVESTIGATION OF *TNFRSF1A* AS A CANDIDATE SUSCEPTIBILITY LOCUS IN ANKYLOSING SPONDYLITIS

4.1. Introduction

TNFRSF1A encodes the type I (p55) tumour necrosis factor receptor (TNFR1). Its potential importance in the pathogenesis of inflammatory diseases, including AS, has been supported by many studies. Evidence for the involvement of TNFR1 comes from several sources:

- 1) TNFR1 has a pivotal role in TNF-related pathways leading to NF κ B-mediated activation of pro-inflammatory genes and FADD-mediated apoptosis (Yang *et al.* 2001);
- 2) TNF blockade with recombinant TNF receptor/Ig fusion protein (etanercept) is a highly effective treatment in AS (Davis *et al.* 2008);
- 3) TNF has been detected in the sacroiliac joints of patients with AS, and disease activity correlates with increased serum levels of TNF (Braun *et al.* 1995; Lange *et al.* 2000).
- 4) Tissue-specific expression of TNFR1 is critical to the development of disease in a mouse model of AS in which inflammatory bowel disease (IBD) and sacroiliitis similar to that seen in AS are observed after over-expression of TNF (Armaka *et al.* 2008);

5) There is also accumulating evidence of genetic association between *TNFRSF1A* and a number of inflammatory disorders, including AS, multiple sclerosis, Crohn's disease and ulcerative colitis (Waschke *et al.* 2005; Lappalainen *et al.* 2008; De Jager *et al.* 2009; Reveille *et al.* 2010; Davidson *et al.* 2011). A well-powered GWAS in AS (2053 cases vs. 5140 controls) reported by the TASC showed suggestive association of AS with rs1800693 (OR=0.85, $p=6.9 \times 10^{-5}$) at *TNFRSF1A* (Reveille *et al.* 2010). A weak association (OR= 0.79, $p=8.2 \times 10^{-4}$) was also reported at *TNSFRSFA1* with rs4149577 in Han Chinese in a somewhat lower powered study (775 cases vs. 1587 controls) (Davidson *et al.* 2011).

TNFRSF1A was re-sequenced in a sample of 48 AS cases to identify novel variants in this functional candidate gene and to investigate this possible association further. Previously a possible association was described between the *TNFRSF1A* region and AS (Reveille *et al.* 2010), but this required further independent replication and refinement. Therefore, re-sequencing was followed by a systematic re-evaluation of the AS association in an independent case-control study. In this study, a low frequency novel variant and 13 other SNPs were genotyped in an independent UK Caucasian population sample containing 988 cases and 1017 controls. Then, an extended study utilising historical controls was performed to increase the statistical power to detect an association. Finally, a meta-analysis of the independent replication study and the TASC GWAS (2010) was undertaken. In addition, stratified analyses were also carried out to identify possible associations between the *TNFRSF1A* variants and subgroups of AS patients, such as those with uveitis, psoriasis and later age of onset.

4.2. Patients, Materials & Methods

4.2.1. Re-sequencing of *TNFRSF1A*

Nine exons, exon-intron boundaries and 2100 bp of 5' flanking sequence of *TNFRSF1A* were sequenced in 48 individuals with AS (96 chromosomes) (See Chapter 2, Section 2.1). The complete sequence of the gene (NG_007506) and the exon locations provided by the REFSEQ database were used to design primers using the program Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) (Table 4.1) (See Chapter 2, Section 2.3). DNA sequencing was carried out using BigDye terminators version 3.1 according to manufacturer's instructions (ABI, Warrington, UK) (See Chapter 2, Section 2.5). Products were resolved and detected using ABI 3100 or ABI 3700 Genetic Analyser (Botnar Research Centre and Wellcome Trust Centre for Human Genetics, Oxford, respectively).

Table 4.1. PCR primers, sequencing primers and PCR conditions for *TNFRSF1A* exons, exon/intron boundaries and upstream sequences.

Primer	Sequence	Sequencing Primer (if different)	Annealing Temperature	Number of Cycles	[Mg ²⁺]	Rxn V	Product size (bp)
Exon1F	CTTGGTGTGTTGGTTGGGAGT	(same primers)	55°C	30	2.5 mM	25 µL	628
Exon1R	GCAGTGGCTGAGGTTAGGAC						
Exon2F	AGATTGTATGGCCCCAACTG	(same primers)	55°C	30	2.5 mM	25 µL	485
Exon2R	GCCGATTCCCTGAAGTCTCT						
Exon3F	TTTTCTCCCTCCCTCCTCTC	(same primers)	52°C	32	3.0 mM	25 µL	366
Exon3R	ATCCATGCAGTGTCACCA						
Exon4F	GGTGGGACACTGCATGGAT	(same primers)	56°C	40	2.5 mM	50 µL	398
Exon4R	GCCAGAGAGGAGTTGGTTGT						
Exon5F	TCCCCCTCCTGTATTCTGTG	(same primers)	53°C	32	2.5 mM	25 µL	391
Exon5R	GAAGAGCAAGGGGAAGGACT						
Exon6F	TGCATCAGTTCCCAAAACAA		55°C	40	1.4 mM	25 µL	450
Exon6R	AATGTATCGCTACCAACGGTG	CTGACCAACACCTGCTTT					400
Exon7F	TGAGGACTCAGGTGAGGAGAA	(same primers)	53°C	32	2.5 mM	25 µL	386
Exon7R	GTCCCCAGCGGTATGAACT						
Exon8F	TGTCGAGCTGCATCAGTAGC	(same primers)	53°C	33	3.0 mM	25 µL	300
Exon8R	TGGGACTTAGAGGGAATGAGG						
Exon9F	Not sequenced						
Exon9R	Not sequenced						
Exon10_1_F	Not sequenced						
Exon10_1_R	Not sequenced						
Exon10_2_F	AGGCGCTTTGCGGCCCCGCCGCC		68°C	32	2.0 mM	25 µL	763
Exon10_2_R	GCAACAGATGACGGTGTTCA	GTTTCTATTAGTGTAACATGATTG					
Exon10_3_F	CTCACCAGCAAGGCTGCTCGGGG	(same primers)	52°C	35	3.0 mM	25 µL	430
Exon10_3_R	CAGATGACGGTGTTTCAGAAC						

Table 4.1 (Continued). PCR primers, sequencing primers and PCR conditions for *TNFRSF1A* exons, exon/intron boundaries and upstream sequences. The technical reasons for not obtaining sequences for exon 9 and part of exon 10 are discussed in Section 4.4.2. ([Mg⁺²]: Magnesium chloride concentration; Rxn V: Reaction volume; bp: Base pair).

Primer	Sequence	Sequencing Primer (if different)	Annealing Temperature	Number of Cycles	[Mg ⁺²]	Rxn V	Product size (bp)
Up(1-2)_1_F	TACTGGAAACTTGCACTTGATTC	(same primers)	52°C	35	3.0 mM	25 µL	377
Up(1-2)_1_R	CTGCTGAGGCTGCTAGAAATGCA						
Up(1-2)_2_F	TGCATTTCTAGCAGCCTCAGCAG	(same primers)	55°C	32	2.5 mM		
Up(1-2)_2_R	AGGTCCAGAAACCCAATGTG						
Up(3-4)_1_F	CACATTGGGTTTCTGGACCT		62°C	32	2.0 mM	25 µL	751
Up(3-4)_1_R	-	CCGCTCCGTTTCCCAGGCTGGAG					391
Up(3-4)_2_F	-	CCTAGGGGGTTGTAGTCGATGATC					517
Up(3-4)_2_R	AGTCCCAACCCTCGTTTTCT						
Up(5-6)_1_F	TGGGGTGAGGAATGGCAGGCACC	(same primers)	52.5°C	34	2.5 mM	25 µL	328
Up(5-6)_1_R	TTAGCTAAGAATGTGTCTTGGAC						
Up(5-6)_2_F	GTCCAAGACACATTCTTAGCTAA	(same primers)	52.5°C	34	2.5 mM	25 µL	328
Up(5-6)_2_R	GGAGGGGAAGAGTGAGGCAGTGT						

4.2.2. Detection of SNPs

Polyphred software package was used to detect SNPs (<http://droog.gs.washington.edu/polyphred/>) (Stephens *et al.* 2006) (See Chapter 2, Section 2.5.6). This software compares individual sequences to each other as well as to a reference sequence and assigns a score for each polymorphic site. Each score corresponds to the probability that a true polymorphic site has been detected (Table 4.2). The Consed tool in the Polyphred package allows visual inspection of the electropherograms. In the analysis, the score threshold was set at 70 to obtain a set of “possible” polymorphic sites (Table 4.2 for threshold probabilities).

The locations of possible SNPs and the flanking sequences were analysed to determine the existence of any known SNPs in *TNFRSF1A* and its flanking sequences, and to exclude known SNPs from further RFLP analysis (<http://www.ensembl.org>). Novel SNPs were identified by comparison with known sequences and SNPs, (<http://www.ensembl.org>, GRCh37; <http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp>, GRCh37, genome build 37.1; <http://genome.ucsc.edu/>, GRCh37/hg19).

Score	True Positive Rate
99	97%
95 – 98	75%
90 – 94	62%
70 – 89	35%
50 – 69	11%
0 – 49	1%

Table 4.2. Polyphred scores and corresponding probabilities of detecting a true polymorphism as supplied by the software designers.

4.2.3. Genotyping of Possible Novel SNPs

In addition to visual inspection of the sequence electropherograms to exclude any spurious “new SNPs”, restriction fragment length polymorphism (RFLP) assays were used to confirm novel SNPs with a polyphred score ≥ 70 . RFLP assays were designed using NEBcutter and Primer3 (<http://tools.neb.com/NEBcutter2/index.php>; http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) for natural restriction sites, or dCAPS finder (<http://helix.wustl.edu/dcaps/dcaps.html>) for engineered restriction sites where no natural site exists (Tables 4.3 and 4.4) (See Chapter 2, Section 2.4.1). Alleles were determined by digestion of PCR products with restriction endonucleases, products were separated by electrophoresis on 3% agarose gels and visualised under UV light.

SNP	Enzyme	Digestion Temp.	Buffer	Incubation	Restriction Site (allele)	Product sizes (bp)
Novel 6	BsRBI	37°C	1X NEB 4	16 hours	CCGCTC (C)	Uncut 400 Cut 75 & 325

Table 4.3. Restriction enzyme and conditions used in the RFLP analysis of a “high probability” novel SNP with natural restriction sites (NEB New England Biolabs).

SNP	Forward & Reverse Primers	PCR conditions Temperature, No. of cycles, [Mg ²⁺]	Enzyme	NEB Buffer	Digest conditions	Restriction site (allele)	Product sizes (bp)
Novel 1	F primer: AGGTGGGAGGATTACTTG R primer: TTTTCATTTTGTGTTTTC <u>G</u> CTT	54°C, 35 cycles, 2.5 mM, 10% DMSO	DdeI	3 (1X)	37°C, 16 hours	CTNAG (A)	Uncut 145 Cut 27 & 118
Novel 7	F primer: GAGCTGCATCAGTAGCTCTC R primer: GTGTCGATTTC <u>CC</u> CAAACCTGA	64°C, 38cycles 2.5 mM, 10% DMSO	BslI	3 (1X)	55°C, 16 hours	CCNNNNNNNGG (G)	Uncut 165 Cut 24 & 141
Novel 8	F primer: TCAGATGAGGCTGCGCCC R primer: CTATGTACATCGAGGGGTTAGC <u>CC</u> CA	67°C, 38 cycles, 2.5 mM, 10% DMSO	BsaI	4 (1X)	60°C, 16 hours	CCNNGG (G)	Uncut 150 Cut 28 & 122

Table 4.4 RFLP analyses for the detection of “high probability” novel SNPs without natural digestion sites. The base underlined and in bold is the engineered base used to create a restriction site. (BSA: Bovine serum albumin; DMSO: Dimethyl sulfoxide; NEB New England Biolabs)

4.2.4. Case-Control Study

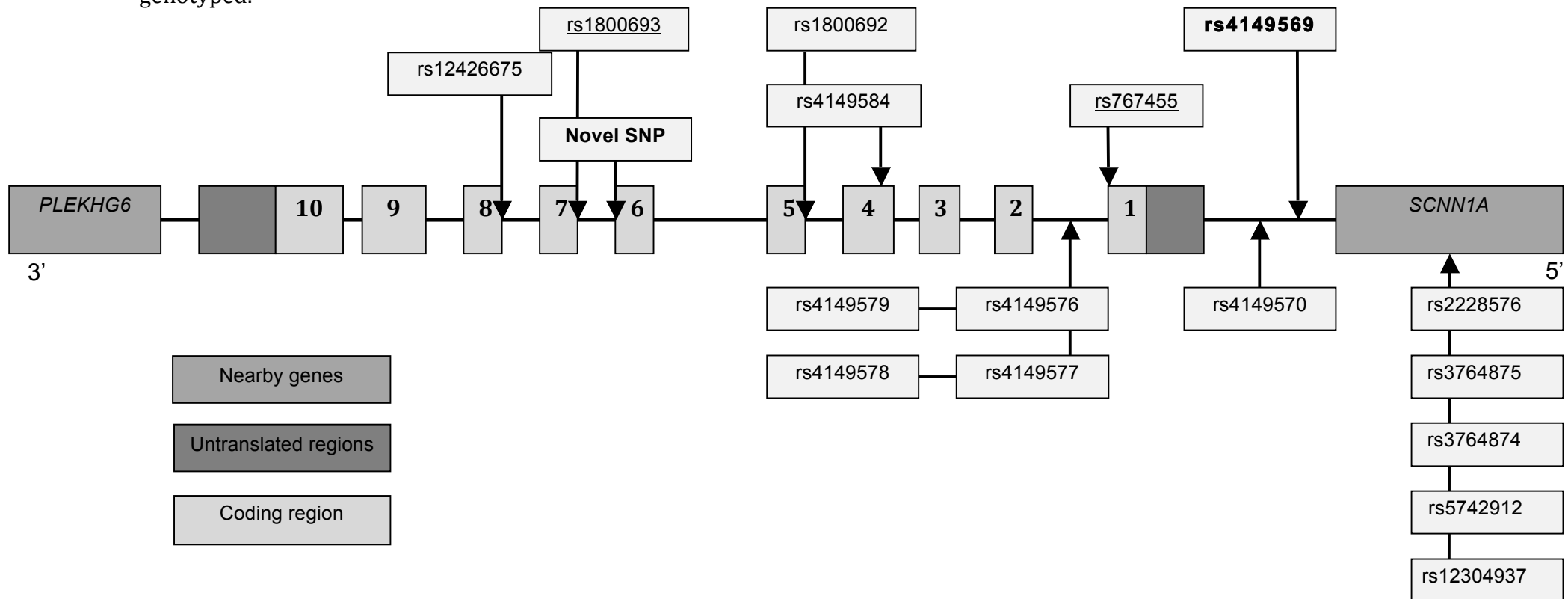
4.2.4.1. Study A

In the first study, “Study A”, 14 SNPs were genotyped in a new sample of 1116 UK Caucasians with AS and 1019 ethnically-matched controls. (See Chapter 2, Section 2.1 for study population). None of these individuals were included in the previous WTCCC-TASC GWAS (Reveille *et al.* 2010). The 14 SNPs included one novel SNP in *TNFRSF1A* detected in the re-sequencing of the gene, 8 tagging SNPs in *TNFRSF1A* ($r^2 > 0.8$, HapMap release 28), 3 of which were detected in the re-sequencing, and 5 tagging SNPs in the final 3 kilobases of the upstream gene *SCNN1A* which encodes a sodium channel, non-voltage-gated 1 alpha subunit (Figures 4.1 and 4.2). Ten SNPs were genotyped by iPLEX technology (MassArray, Sequenom) and four SNPs (rs4149576, rs1800692, rs12426675, novel SNP IVS6-32C→A) by KASPar technology (KBiosciences, Hoddesdon, Herts, UK) (See Chapter 2, Section 2.4.2).

Examples of the genotyping cluster plots are shown in Figure 4.3. Genotypes were tested for adherence to Hardy-Weinberg equilibrium (HWE); p-values < 0.05 were considered statistically significant deviations from HWE (See Chapter 2, Section 2.4.2.3). The Cochran-Armitage test of trend was used for case-control analyses; p-values < 0.05 were considered statistically significant (See Chapter 2, Section 2.6.2). The statistical power of the independent UK case-control study to detect association with the variants in the *TNFRSF1A* region were between 15% and 78%, given an OR of 1.1-1.2, a MAF of 0.09-0.40, an alpha level of 0.05 (2-sided) and a disease prevalence of 4/1000 with a log-additive disease model (Quanto, version

1.2.4, May 2009) (See Chapter 2, Section 2.6.1). Previous OR estimates from the TASC GWAS were used for these power calculations (Reveille *et al.* 2010).

Figure 4.1. Structure of the *TNFRSF1A* locus with detected known and novel SNPs with their relative locations (above). SNPs tagged by rs4149576 ($r^2 > 0.8$), which were and not genotyped are underlined. rs4149569 (in bold) is within a repeat sequence and was not genotyped.



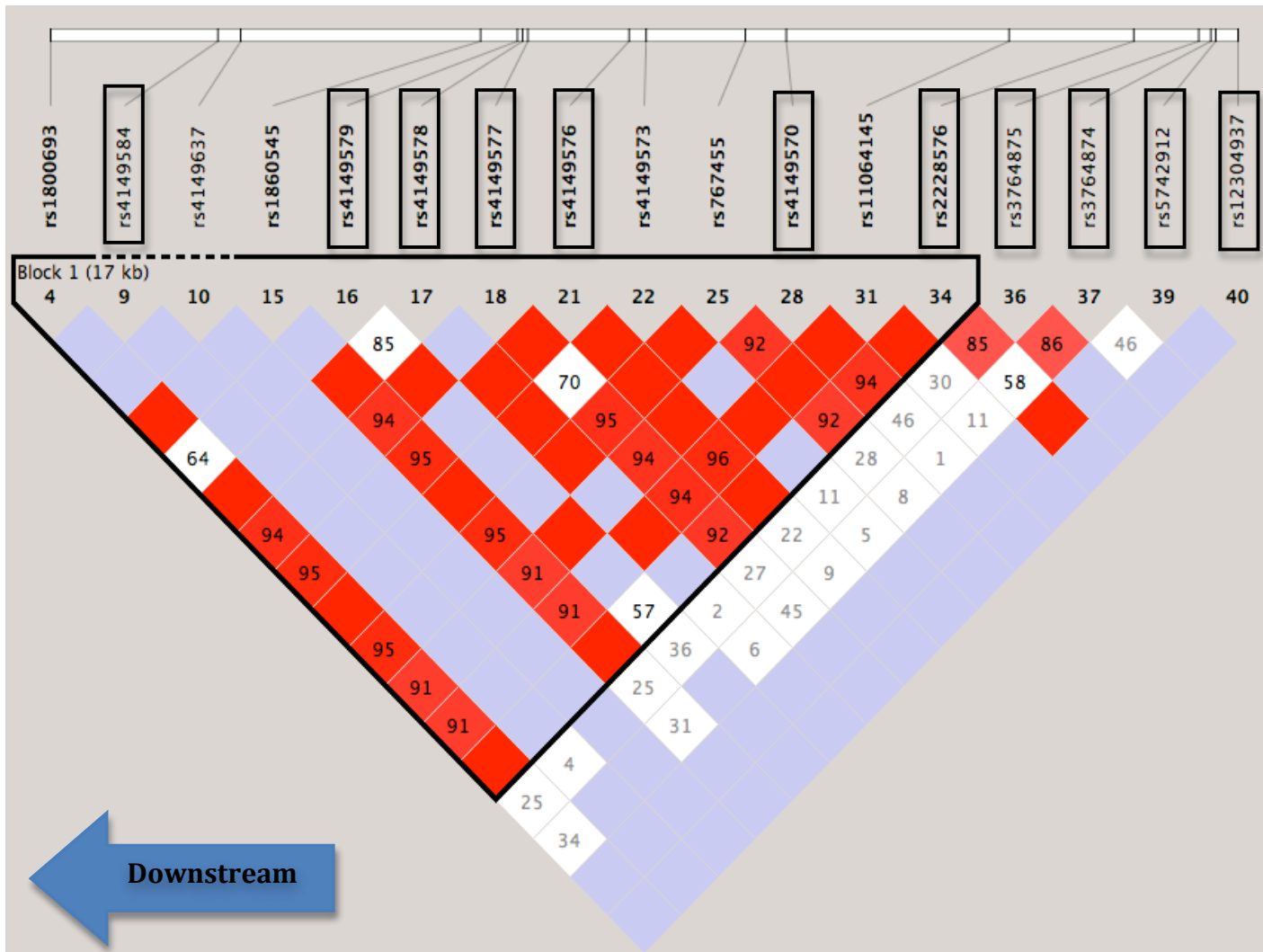
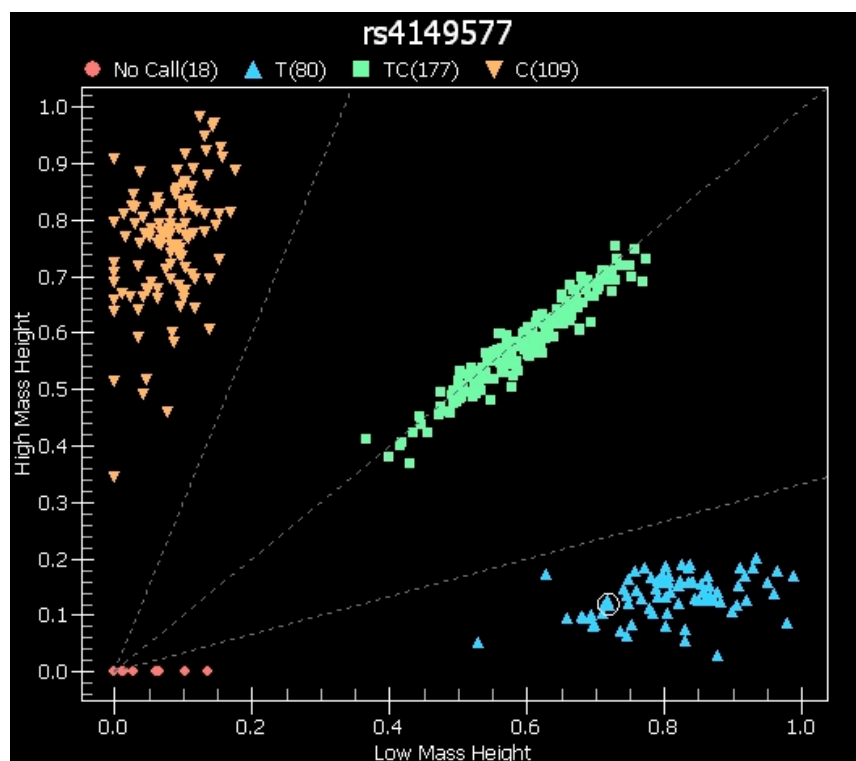


Figure 4.2. *TNFRSF1A* region D' LD plot (HapMap release 28). D' plots show the LD pattern within the region. D' between two SNPs is low when there is recombination between the SNPs. Red colours represent high LD, lighter colours represent lower LD. The SNPs used in the study are shown in boxes (SNPs rs12426675, novel SNP and rs1800692 are not shown).

(1)



(2)

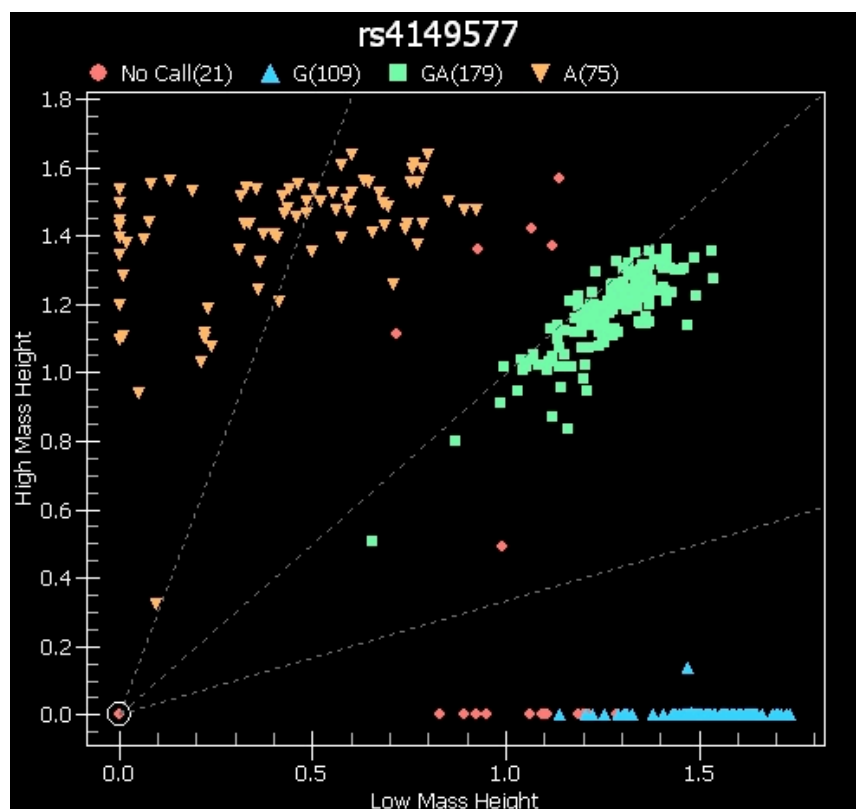


Figure 4.3. iPLEX cluster plots for rs4149577 showing (1) clearly segregated 3 genotype clusters (2) overlapping clusters. Plots were visually checked. Samples with ambiguous genotypes were excluded (in red) (See Chapter 2, Section 2.4.2.3).

4.2.4.2. Study B

This study was then extended (“Study B”; 1116 cases vs. ~2419 controls) by including control data on between 1400 and 1475 controls (depending on the SNP) from the 1958 British Birth Cohort (BBC) (<http://www.b58cgene.sgul.ac.uk/October 2008>). This increase in the number of controls was carried out in order to increase the statistical power of the study. The statistical power of the extended UK case-control study to detect association with the variants in the *TNFRSF1A* region were 19% and 91% respectively, given an OR of 1.1-1.2, a MAF of 0.09-0.40, an alpha level of 0.05 (2-sided) and a disease prevalence of 4/1000 with a log-additive disease model (Quanto, version 1.2.4, May 2009) (See Chapter 2, Section 2.6.1).

4.2.4.3. Stratified Analyses

Further analyses were undertaken after stratification for age of symptom onset (age cutoff=25), presence of peripheral joint disease, inflammatory bowel disease (Crohn’s disease and ulcerative colitis), uveitis and for HLA-B27 status using the program PLINK (PLINK v1.07, August 2009). These stratified analyses were done using 1116 replication cases and 1019 controls from “Study A”, and 485 cases from the TASC 2010 GWAS (Reveille *et al.* 2010), for whom all individual genotype data was available for the PLINK analyses. For the stratified analyses, after the application of a conservative Bonferroni correction for multiple testing (14 SNPs x 5 strata=70 tests), a P-value < 0.0007 (0.05/70) was considered statistically significant.

4.2.5. Meta-analysis

Data were available for only two *TNFRSF1A* SNPs (rs4149577 and rs4149578) from the 2010 TASC case-control study (2053 cases and 5140 controls) (Reveille *et al.* 2010). “Study A” data (1116 cases and 1019 controls) and the TASC discovery set data for these two SNPs were combined in a meta-analysis. The Mantel-Haenszel test was used to calculate fixed-effects pooled odds ratio, chi-square and p-values using the StatsDirect software (version 2.6.6 03/02/2008). Cochran Q p-values were calculated to establish the combinability of the studies. An insignificant p-value (>0.05) indicated low evidence of heterogeneity between studies, validating the use of a fixed-effects meta-analysis approach (See Chapter 2, Section 2.6.3). The statistical power of the meta-analysis to detect association with the variants in the *TNFRSF1A* region were 43% and 100% respectively, given an OR of 1.1-1.2, a MAF of 0.09-0.40, an alpha level of 0.05 (2-sided) and a disease prevalence of 4/1000 with a log-additive disease model (Quanto, version 1.2.4, May 2009) (See Chapter 2, Section 2.6.1).

4.2.6. Haplotype Analysis

It is possible that none of the SNPs genotyped in this experiment represent a true functional SNP responsible for the association between genetic variants in this region and AS. In this case, a more powerful strategy might be to perform a haplotype analysis across this region, which might better capture the relationship between an untyped variant and the disease risk. In addition, interactions between variants on the same haplotype are also taken into account using this approach, which may not be tested effectively using single-locus tests of association.

The haplotype analyses were done using 1116 replication cases and 1019 controls from “Study A”, and 485 cases from the TASC 2010 GWAS (Reveille *et al.* 2010), for whom all individual genotype data was available for the PLINK analyses. The program PLINK was used to derive haplotypes from the genotype data (SNPs rs12426675, novel SNP, rs1800692, rs4149584, rs4149579, rs4149578, rs4149577, rs4149576, rs4149570, rs2228576, rs3764875, rs3764874, rs5742912 and rs12304937). A sliding window method was used to analyse haplotypes consisting of between 2 and 14 SNPs. Differences in the frequencies of the predicted haplotypes between cases and controls were tested using chi-square test for common haplotypes (frequency>5%), as well as permutation (number of permutations=1000) in order to calculate the empirical corrected p-value for low frequency haplotypes (frequency<5%) (PLINK v1.07, 2009) (See Chapter 2, Section 2.6.2).

4.3. Results

4.3.1. SNP Detection

Sixteen possible SNPs were detected (Table 4.5). Comparison with the Ensembl database (<http://www.ensembl.org>) revealed that 6 of these were already known (rs4149569, rs767455, rs4149584, rs1800692, rs1800693 and rs12426675) (Figure 4.3). To confirm the existence of other possible SNPs (Polyphred score ≥ 70 ; novel SNPs 1-10, Table 4.5), visual inspection of electropherograms and RFLP analyses were performed. However, only one was a novel SNP in intron 6 (IVS6-32C $\rightarrow$ A). Nine other possible SNPs proved to be monomorphic. Novel SNPs 2, 3, 4, 5 and 9 were excluded from RFLP analysis following visual inspection of the electropherograms as well as because of their lower polyphred probability scores (score <90) (Table 4.5) (See Appendix 1 for additional electropherograms).

#	ID	Location	Base Change (ancestral)	Location within the gene/region	Sequence	# of cases (heterozygous)	# of cases (homozygous)	MAF	# of sequences analysed	Polyphred Score (0-99)
1	rs4149569	6452468	C/G (G)	Upstream	AGCACTTTAG (C/G) AGGCTGAGGT	C/G 4	C/C 6 G/G 9	C 0.42	19 cases	98
2	Novel 1	6452233	A/T	Upstream	GTTTTTGT (A/T) GAGGGTCCGC	A/T 13	T/T 25	A 0.17	38 cases	93
3	rs767455	6450945	A/G (A)	Exon1 (SYN Coding)	TGCTGCTGCC (A/G) CTGGTGAGAC	A/G 8	G/G 5 A/A 9	G 0.41	22 cases	93
4	Novel 2	6443530	C/G	Intronic	GAGTGGCGTG (C/G) CCTGGGCGGA	C/G 7	G/G 11	C 0.19	18 cases	83
5	Novel 3	6443298	G/T	Exon 2	ATTTTGAGGG (G/T) GGATATATTT	-	T/T 43 G/G 1	G 0.02	44 cases	80
6	rs4149584	6442633	A/G (G)	Exon 4 (Arg121Gln)	ACAGTGGACC (A/G) GGACACCGTG	A/G 1	G/G 42	A 0.01	43 cases	97
7	rs1800692	6442346	C/T (C)	Intronic	TGCAGGCGCT (C/T) CTCCTTTAGC	C/T 4	T/T 4 C/C 2	C 0.40	10 cases	97
8	rs1800693	6440009	C/T (T)	Intronic	ACCAGGTCAC (C/T) TCTCCTCACC	C/T 6	T/T 21	C 0.11	27 cases	97
9	Confirmed Novel SNP IVS6-32C→A	6439909	A/C	Intronic	GTCAGAGGAG (A/C) GGGCAGAGGG	-	C/C 5 A/A 1	A 0.17	6 cases	98
10	Novel 4	6439325	A/T	Intronic	TTAGAGGGAA (A/T) GAGGCGAGCC	A/T 14	T/T 23 A/A 2	A 0.23	39 cases	72
11	Novel 5	6439440	C/T	Exon 8	CACCTCTTTT (C/T) CAGGTGTCGA	C/T 6	T/T 39	C 0.07	45 cases	89
12	Novel 6	6439454	A/C	Exon 8	GTGTTCGATTT (A/C) CCACAACTG	A/C 1	C/C 44	A 0.01	45 cases	79
13	Novel 7	6439466	G/T	Intronic	CACAAACTGA (G/T) GAAAAAGAAA	G/T 9	G/G 36	T 0.10	45 cases	96
14	rs12426675	6439470	A/G (G)	Intronic	AACTGAGGAA (A/G) AAGAAAGAAA	A/G 13	A/A 32	G 0.14	45 cases	94
15	Novel 8	6438362	C/T	Exon 10	AGCCGCCTAC (C/T) TGGTGCTAAC	C/T 3	T/T 39	C 0.04	42 cases	91
16	Novel 9	6437904	A/C	3' UTR	CCCACCTCCC (A/C) TCCCGACAGA	A/C 5	A/A 34	C 0.06	39 cases	71

Table 4.5. Known and possible novel SNPs detected in *TNFRSF1A*. (Chr.: chromosome, SYN: synonymous). The chromosomal locations, sequences and minor allele frequencies are also shown.

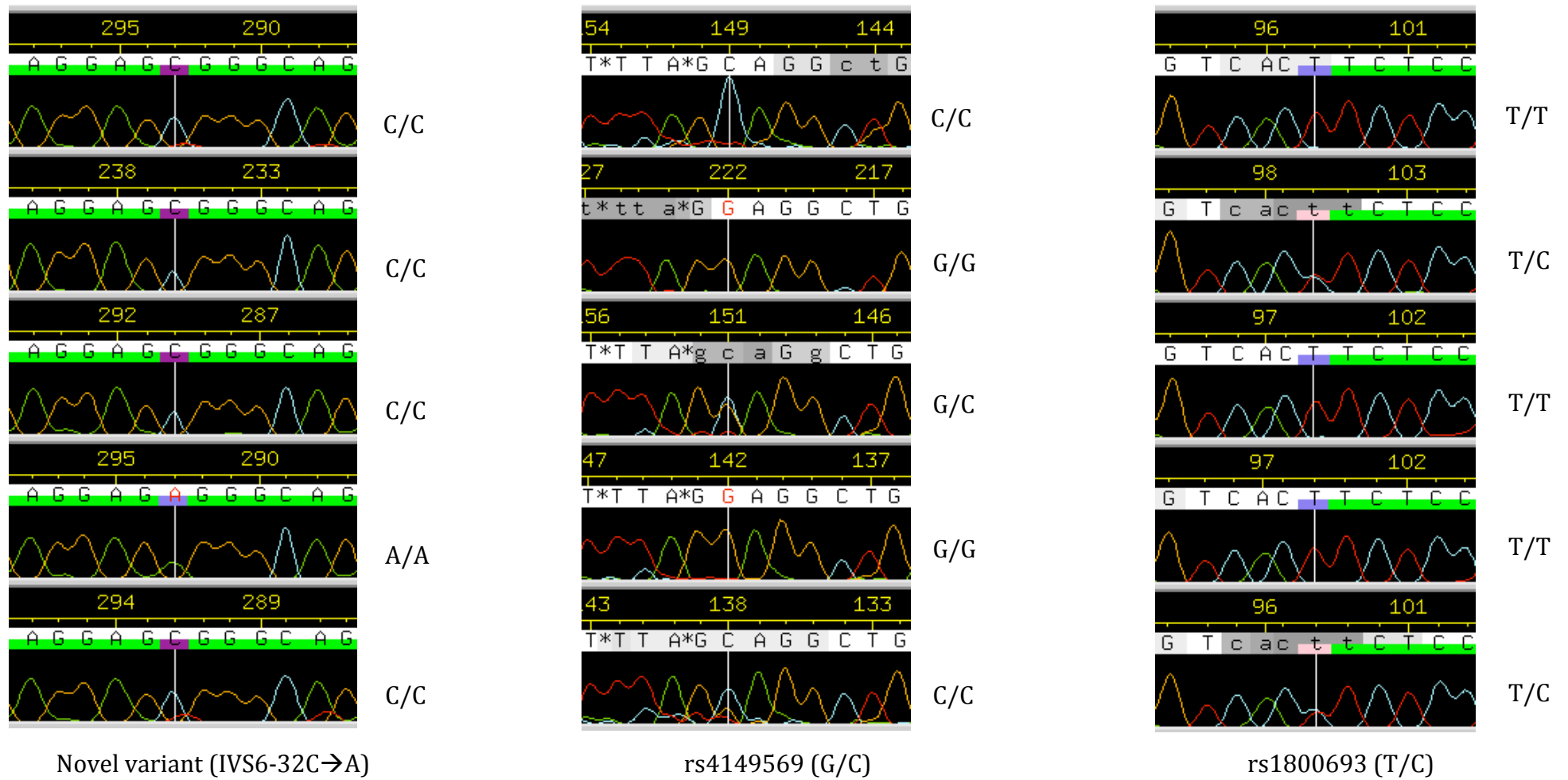


Figure 4.3. Electropherograms for the novel variant (IVS6-32C→A), and two known SNPs (rs4149569 and rs1800693) in 5 individuals. Genotypes for each individual are shown.

4.3.2. Case-Control Study

4.3.2.1. Study A

The results of this case-control study are shown in Table 4.6. There were no significant associations (i.e. $p < 0.05$) between AS and any of the SNPs in *TNFRSFA1* or *SCNN1A*.

4.3.2.2. Study B

Increasing the power of the study by including control data from the 1958 BBC revealed weak associations with one low frequency marker, rs4149584 ($p = 0.01$, OR=1.58, 95% CI 1.11-2.26). However, no associations were observed with the other SNPs in this study.

4.3.2.3. Stratified Analyses

No additional significant associations were apparent after stratifying for the presence of peripheral joint disease, inflammatory bowel disease, uveitis or for HLA-B27 status. After stratifying on the mean of age of onset (25 years), a borderline statistically significant association was observed with rs4149570 and age of onset older than or equal to 25 ($p = 0.0006$, OR=1.4, 95% CI 1.1-1.6, number of cases with age of onset older than or equal to 25=382). After conditioning on rs4149570 in the logistic regression analysis, no residual association was observed with age of onset (≥ 25) and the other SNPs in the region.

STUDY			(Study A) UK CASE - CONTROL STUDY (Cases~1116, controls~1019)					(Study B) EXTENDED UK CASE - CONTROL STUDY (Including 1958 BBC controls~1400, total controls~ 2419)		
SNP	GENE	POSITION	P	OR (95% CI)	MAF CONTROLS	MAF CASES	ALLELE	P	OR (95% CI)	MAF CONTROLS
rs12426675	<i>TNFRSF1A</i>	Chr12:6439470	0.16	1.12 (0.96-1.31)	0.19	0.21	G			
Novel SNP	<i>TNFRSF1A</i>	Chr12:6439909	0.05	0.15 (0.02-1.27)	0.003	0.0005	T			
rs1800692	<i>TNFRSF1A</i>	Chr12:6442346	0.5	0.95 (0.84-1.10)	0.43	0.41	T			
rs4149584	<i>TNFRSF1A</i>	Chr12:6442643	0.06	1.55 (0.99-2.43)	0.016	0.02	T	0.01	1.58 (1.11-2.26)	0.015
rs4149579	<i>TNFRSF1A</i>	Chr12:6447357	0.3	1.13 (0.90-1.42)	0.07	0.08	T	0.4	1.08 (0.90-1.30)	0.08
rs4149578	<i>TNFRSF1A</i>	Chr12:6447437	0.1	1.19 (0.97-1.48)	0.08	0.10	T	0.4	1.07 (0.90-1.27)	0.09
rs4149577	<i>TNFRSF1A</i>	Chr12:6447522	0.6	0.97 (0.85-1.10)	0.5	0.49	A	0.1	0.92 (0.82-1.02)	0.51
rs4149576	<i>TNFRSF1A</i>	Chr12:6449115	0.4	0.95 (0.83-1.10)	0.42	0.40	T	0.2	0.94 (0.84-1.04)	0.42
rs4149570	<i>TNFRSF1A</i>	Chr12:6451590	0.8	1.02 (0.90-1.15)	0.42	0.42	A	0.4	1.04 (0.94-1.16)	0.41
rs2228576	<i>SCNN1A</i>	Chr12:6457062	0.3	0.93 (0.82-1.06)	0.36	0.34	T	0.3	0.95 (0.85-1.06)	0.36
rs3764875	<i>SCNN1A</i>	Chr12:6458084	0.3	1.06 (0.94-1.21)	0.355	0.37	C	0.3	1.05 (0.95-1.17)	0.36
rs3764874	<i>SCNN1A</i>	Chr12:6458274	0.2	1.10 (0.95-1.27)	0.22	0.24	C			
rs5742912	<i>SCNN1A</i>	Chr12:6458350	0.5	1.15 (0.74-1.78)	0.019	0.022	G			
rs12304937	<i>SCNN1A</i>	Chr12:6458703	0.2	0.73 (0.42-1.24)	0.016	0.01	A			

Table 4.6. The UK case-control study ("Study A") and the extended study ("Study B") results for SNPs in the *TNFRSF1A* region. SNPs

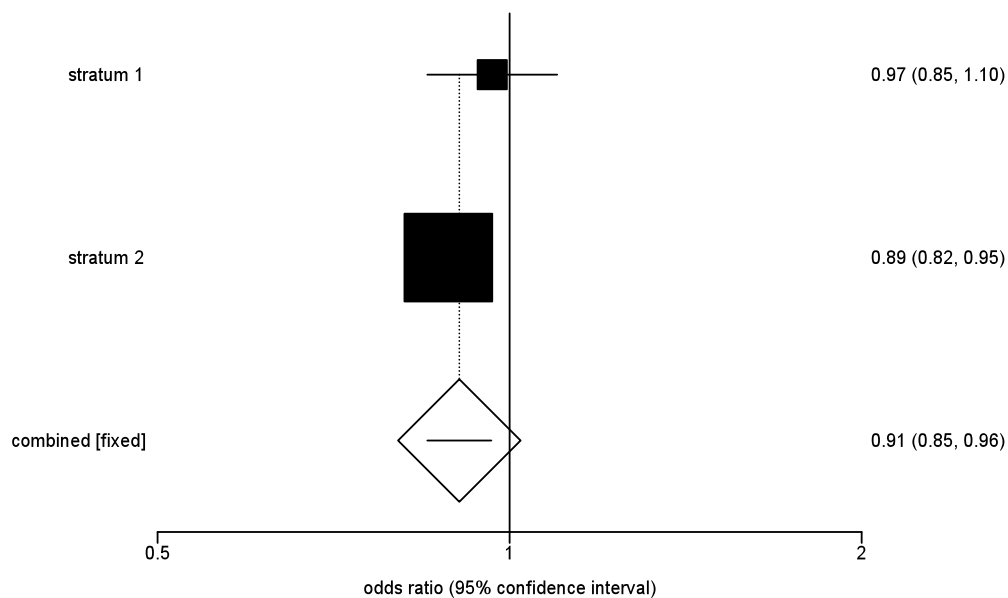
showing association ($p < 0.05$) are shown in bold. (OR: odds ratio, CI: confidence interval, MAF: minor allele frequency, BBC: British birth cohort (<http://www.b58cgene.sgul.ac.uk/>); Position: Genome build 37.3)

4.3.3. Meta-analysis

When data for the two SNPs (rs4149577 and rs4149578) common to “Study A” and the previously published TASC study were combined in a meta-analysis, a somewhat stronger association was observed with rs4149577 ($p=0.002$, $OR=0.91$, 95% CI 0.85-0.96) and rs4149578 ($p=0.015$, $OR=1.14$, 95% CI 1.03-1.27). The magnitude and direction of the associations are in agreement with the meta-analysis with very narrow confidence intervals for the fixed effects pooled odds ratios for rs4149577 and rs4149578 (Figure 4.4).

(A) rs4149577

Odds ratio meta-analysis plot [fixed effects]

**(B) rs4149578**

Odds ratio meta-analysis plot [fixed effects]

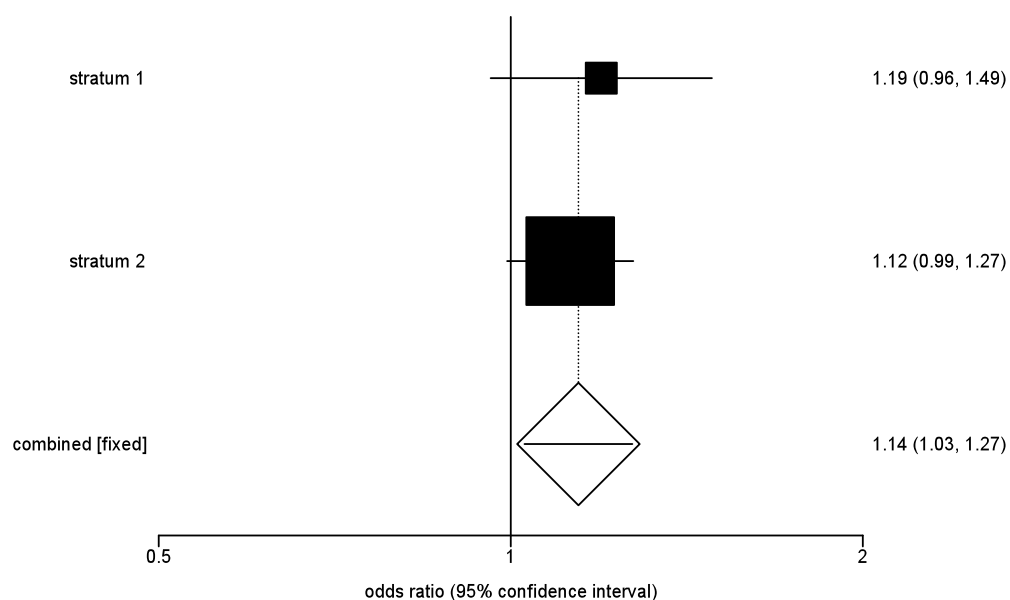


Figure 4.4. Forest plots for the meta-analyses of *TNFRSF1A* SNPs (A) rs4149577 and (B) rs4149578. (Stratum 1 – Study A; stratum 2 – TASC study)

4.3.4. Haplotype Analysis

Comparison of patient and control haplotypes revealed no associations in the sliding window analyses except for one very low frequency 8-SNP haplotype (frequency=0.012) with corrected $p < 0.05$ (Discussed in Section 4.4.1.4);

rs12426675, novel SNP, rs1800692, rs4149584, rs4149579, rs4149578,
rs4149577, rs4149576;

AGCCCCGT

Haplotype frequencies in cases and controls were 0.018 and 0.004, respectively, with an OR of 10.8 (Chi-square=19, $p = 1.3 \times 10^{-5}$). After performing 1000 permutations, the corrected p-value was 0.001.

4.4. Discussion

4.4.1. Case-Control Study, Meta-analysis & Haplotype Analysis

4.4.1.1. Case-Control Study & Extended Study

At the time of these studies were undertaken, there was limited evidence of association between AS and *TNFRSF1A* that required investigation. It was unclear whether variants within the gene *TNFRSF1A* were associated with AS. To clarify the possible association between variants at this locus and AS, I sequenced the *TNFRSF1A* gene in a limited number of cases (n=48), but was able to detect only one new SNP (IVS6-32C→A) with a very low minor allele frequency (Discussed in Section 4.4.3) and to confirm the existence of six other SNPs in *TNFRSF1A* with minor allele frequencies in excess of 0.01. My initial case-control study (Study A) of 14 SNPs did not reveal any significant associations with AS, except for a borderline association with the novel SNP (IVS6-32C→A, $p=0.05$, OR=0.15) (Table 4.6, Study A). Such an association with a very rare variant (MAF=0.003 in cases, MAF=0.0005 in controls) requires replication to be confirmed and considered true. It is well recognised that the power to detect association in studies of complex traits and diseases is low. In this study, some of the SNPs were of low MAF (rs4149584, rs4149579, rs4149578, rs5742912 and rs12304937) and had small effect sizes implying low statistical power < 50%, which might explain why no associations were detected.

In the second part of the study (Study B), I was able to increase the study power for some SNPs by including additional controls from the 1958 BBC, which allowed the detection of a strong association with a low frequency SNP

(rs4149584, $p=0.01$, $OR=1.58$). For this SNP, the study power was increased from 51% (Study A) to 65% (Study B) (Table 4.6), and confidence that this is a true association was also increased as the 95% CIs were reduced (95% CI 0.99-2.43 (Study A); 95% CI 1.11-2.26 (Study B)) for both the associated SNP rs4149584 and non-associated SNPs (Table 4.6, Study B).

This low frequency non-synonymous coding SNP (rs4149584, $G \rightarrow A$, $Arg \rightarrow Gln$) was associated with atherosclerosis (Poirier *et al.* 2004), tumour necrosis factor receptor associated periodic syndrome (TRAPS) (Aksentijevich *et al.* 2001; Kumpfel *et al.* 2008) and multiple sclerosis (MS) (De Jager *et al.* 2009; Swaminathan *et al.* 2010), all of which are inflammatory conditions (Discussed in Sections 4.4.3 and 4.4.3.1). Interestingly, there were two independent signals in *TNFRSF1A* in a meta-analysis study of MS (rs4149584, $OR=1.6$, 95% CI 1.15-2.17, $p=5.25 \times 10^{-6}$; rs1800693, $OR=1.2$, 95% CI 1.10-1.31, $p=1.50 \times 10^{-11}$) (De Jager *et al.* 2009). In the current study, SNP rs4149576 tags rs1800693 ($r^2=0.8$, $D'=0.96$, HapMap release 22; Figure 4.1 & 4.1.2), which showed the strongest association with AS (rs1800693, $p=6.9 \times 10^{-5}$) in the TASC GWAS as well (Reveille *et al.* 2010). Even though this SNP (rs1800693) was not formally typed here, as this case-control study was carried out in parallel to the TASC GWAS, my data did not replicate the TASC finding by testing the tagging SNP rs4149576. According to power calculations based on the TASC GWAS results, both “Study A” and “Study B” were sufficiently powered to detect associations (Power>80%). However, since the SNPs rs1800693 and rs4149576 are not in complete linkage disequilibrium ($r^2<1.0$), it might be possible to miss an association by testing only rs4149576.

4.4.1.2. Stratified Analyses

A significant association was observed only with age of symptom onset. A total of 990 cases had data on the age of symptom onset, with the average age being 25 years. Patients were grouped according to age at onset into those under 25 years (608 cases), and 25 years and older at onset (382 cases). There was a nominal association between rs4149570 and age of onset ≥ 25 (OR=1.4 (allele A), 95% CI 1.1-1.6, $p=0.0006$). Even though this subgroup had fewer cases, the power to detect an association might be increased by the sub-grouping strategy if the variant differentially affects individuals according to age. The observed effect might mean that *TNFRSF1A* rs4149570 association may be influential in AS cases with an older age of symptom onset (≥ 25).

rs4149570 genotype	Statistic	Age of onset	t test p-value
AA	Mean	25.90	AA vs. CC 0.002
	Mode	20	
	Median	24	
CC	Mean	22.96	CC vs. CA 0.015
	Mode	18	
	Median	21	
CA	Mean	24.81	CA vs. AA 0.226
	Mode	21	
	Median	22	

Table 4.7. Mean, mode and median age of onset for the three rs4149570 genotypes. T-test was employed to compare the means (2-tailed, unequal variance, 2-sample test).

When the mean, mode and median were calculated for three different genotypes of rs4149570, a tendency towards an older age of disease onset was observed in AA homozygotes compared to CC homozygotes with heterozygotes showing an intermediate age when cases are subgrouped by mean and median (Table 4.7). This observation can only be considered robust after further detailed analysis and replication in an independent dataset.

Interestingly, rs4149570, located in the promoter region, has been shown to be associated with the level of expression of *TNFRSF1A* in hepatocellular carcinoma (Kim *et al.* 2008). The A allele of rs4149570 was associated with repression of *TNFRSF1A* expression. In the context of AS, one could speculate that lower expression of TNFR1 might lead to a later age of symptom onset, or a less severe form of the disease because of a reduction in the pro-inflammatory effect of this receptor.

4.4.1.3. Meta-analysis

The meta-analysis of rs4149577 and rs4149578 was consistent with a real association between AS and *TNFRSF1A* (rs4149577, $p=0.002$, OR=0.91, 95% CI 0.85-0.96; rs4149578, $p=0.015$, OR=1.14, 95% CI 1.03-1.27). This meta-analysis provided further evidence for the association between these SNPs and AS as observed p -values are lower i.e. more significant compared to the individual p -values from the combined studies (Meta-analysis: rs4149577 $p=0.002$, rs4149578 $p=0.015$; this study: rs4149577 $p=0.2$, rs4149578 $p=0.8$; TASC GWAS: rs4149577 $p=0.0047$, rs4149578 $p=0.17$). Another recent support for *TNFRSF1A* as a susceptibility locus comes from the detection of an association

with AS in a different ethnic group, the Han Chinese (rs4149577, $p=8.2 \times 10^{-4}$, OR=0.79, 95% CI 0.69-0.91) (Davidson *et al.* 2011). rs4149577 also exhibited the strongest association in this meta-analysis study (Figure 4.4). The data from the previous Chinese study was not included in the meta-analysis because of the differences between the distinct populations i.e. ancestry and allele frequencies (Davidson *et al.* 2011). However, the direction of association for rs4149577 was the same in all three studies (Study A [OR=0.97], the TASC [OR=0.89] and Chinese [OR=0.79]). Furthermore, the requirement for larger datasets to achieve sufficient statistical power to detect smaller effect sizes ($1.1 > \text{OR} > 1.3$) reliably is clearly emphasised by this meta-analysis ($43\% < \text{power} < 100\%$).

Even though the identified effect sizes are small, these apparent associations could identify potential pathways involved in disease pathogenesis. In turn, this may assist in identifying particular components of a pathway that are amenable to manipulation. For example, the weak but robust association between rheumatoid arthritis (RA) and *CTLA4*, increasing the risk of disease by about 8 per cent (Stahl *et al.* 2010), is indicative of the potential for therapeutic manipulation of T cell activation through the CD28 pathway. This has been amply demonstrated by the efficacy of recombinant CTLA4 immunoglobulin fusion protein (abatacept) in RA (Kremer *et al.* 2003). It is also possible that such genetic associations could assist in identifying individual patient's therapeutic response. For example, an association between individual patient responses to anti-TNF therapy (infliximab) in Crohn's disease (CD) and the *TNFRSF1A* SNP rs767455 has been described in the Japanese (Matsukura *et al.* 2008). If replicated in this and other populations, this observation may either help

identify patients that will most benefit from the therapy, or determine those who might require increased doses to achieve the desired therapeutic effect.

4.4.1.4. Haplotype Analysis

The positive associations with an 8-SNP haplotype should be considered in the context of its low frequencies. This haplotype has a very low frequency (1.8% in cases and 0.4% in controls); consequently the observed association may reflect the presence of low frequency variant(s), which is tagged by this haplotype and was not detected in the re-sequencing of the gene. Alternatively, it is possible that an interaction between the SNPs in the haplotype, which cannot be observed in single-locus tests, is causing either increased or decreased disease risk leading to a statistically significant association (OR=10.8, corrected $p=0.001$). Before these results can be considered robust, an independent replication is required in order to ensure that the observations are not merely artefacts of correlation within the gene region. Upon replication, it may be suitable to re-sequence the specific haplotype in AS cases to scan the region for any novel SNPs that might be associated with AS. Alternatively, it may be possible to impute untyped SNPs in the region to look for associations with these additional variants that might be tagged by this haplotype.

4.4.2. Problems with Re-sequencing of *TNFRSF1A*

4.4.2.1. A Brief Summary of the Sequencing Process

DNA sequencing involves six main steps:

4. Initial amplification of the region to be sequenced by polymerase chain reaction (PCR).
5. Detection and purification of the amplification products.
6. Sequencing of purified PCR products with fluorescently-labelled dideoxy nucleotides (ddNTPs) using the chain termination method.
7. Purification of dideoxy-labelled products.
8. Separation of and detection of dideoxy-labelled products.
9. Detection of DNA variants.

Each of these steps is prone to errors that may cause problems with the re-sequencing process. Two of these problems arose in this study.

4.4.2.2. Re-sequencing Problems due to the Initial Amplification Step

TNFRSF1A has 10 exons; exons 9 and 10 are the two largest (Figure 4.1). Eight of the exons, exon-intron boundaries and upstream sequences were fully re-sequenced (Table 4.1). Similarly, primers to amplify exons 9 and 10 were designed and tested under different PCR conditions to obtain specific PCR products. Various strategies to remove long non-specific PCR products were also undertaken. These included:

- 1- Increasing annealing temperature
- 2- Decreasing annealing time

- 3- Decreasing extension time
- 4- Decreasing extension temperature to 62-68°C
- 5- Increasing potassium chloride concentration ([KCl]) to 1.2x-2x, while keeping [Mg²⁺] at 1.5-2.0 mM
- 6- Increasing [Mg²⁺] up to 3-4.5 mM, while keeping deoxy nucleotide (dNTP) concentration constant
- 7- Using less primer
- 8- Using less DNA template
- 9- Using less Taq polymerase
- 10- Checking for repeat sequences in the primers
- 11- Using additives such as DMSO (10%) and glycerol (5-10%)

Exons 9 and 10 were also divided into 2 or 3 sections in order to facilitate PCR reactions. Two sections of exon 10 were sequenced upon division, whereas it was not possible to sequence exon 9 and section 1 of exon 10 (Table 4.1).

Furthermore, published primers were tested for exons 9 and 10 (McDermott *et al.* 1999; Bazzoni *et al.* 2000; Aksentijevich *et al.* 2001). None of these worked successfully.

These downstream sections (exon 9 and exon 10 section 1) may contain novel mutations that can provide insights into the disease aetiology. However, these mutations might be in LD with the SNPs tested in the case-control study. Thus, it may still be possible to detect their effects indirectly via linkage disequilibrium (Figure 4.2 for the LD plot).

4.4.2.3. Re-sequencing Problems due to Purification of the Sequencing Reaction

Polyphred software was used for SNP detection. It aligns and compares the sequences for each individual, and reports locations of possible polymorphisms. Thus, it is important to have high quality sequences for accurate sequence alignment. Regions of low quality sequences can arise as a result of failure of the second purification step (step 4) to remove all unused ddNTP, which can obscure the first 100 or so bases of the sequence as they exit the electrophoresis capillary as “dye blobs”. There may also be loss of sequence products in the second purification step reducing the signal to noise ratio making true variants harder to detect. Consequently, it was necessary to visually inspect the electropherograms and eliminate the low scoring (<90) i.e. low probability SNPs from further analysis (Table 4.5).

4.4.3. Implications for the Variants of *TNFRSF1A* & their Associations with AS

TNFR1, the primary p55 TNF receptor, is sequestered in the Golgi apparatus in resting cells and interacts with TNFR-associated-death-domain protein (TRADD) (Jones *et al.* 1999; Aggarwal 2003). Upon stimulation, TNFR1 binds to cytoplasmic TRADD, which initiates a signalling cascade leading to NF- κ B activation, translocation to the nucleus and initiation of gene transcription (Yang *et al.* 2001; Blonska *et al.* 2005). Some of the genes under NF- κ B regulation enhance cell survival and pro-inflammatory signals such as interferon- γ and interleukin 6 (Sica *et al.* 1997; Son *et al.* 2008). TNFR1 can also initiate cell death by binding of Fas-associated DD (FADD) to TRADD. The FADD-TRADD complex

recruits pro-caspase 8, which is activated via autocatalytic activity leading to activation of pro-caspase 3 that causes apoptosis (Micheau *et al.* 2001).

The *TNFRSF1A* gene is known to be highly polymorphic, specifically in the cysteine-rich extra-cellular domains. The cysteine-rich extracellular domains, the transmembrane domain and the cytoplasmic domains are encoded by exons 1 – 7, 23 residues of exon 7 and the remaining residues of exons 7 – 10, respectively (Schall *et al.* 1990). *TNFRSF1A* has 16 splice variants with 9 protein coding transcripts, and three of these protein coding transcripts undergo nonsense-mediated decay (Ensemble: ENSG00000067182). The novel SNP (IVS6-32C>A, or NM_001065.3:c.626-32C>A) in intron 6 (32 base pairs 5' of exon 7, 100 base pairs 3' of rs1800693) described here may play a role in gene splicing and/or influence gene expression. As it is located in intron 6, it might theoretically affect all three domains of the protein. The extracellular domain is crucial for interactions with the ligand cytokine TNF; the cytoplasmic domain is important for signal transduction by interacting with the death domain-containing proteins, such as TRADD; and the transmembrane domain is also necessary for signal transduction as well as protein stability.

Examples of splice site variations in *TNFRSF1A* have been identified previously. A mutation (c. 193-14G>A) creating a splice acceptor site (AG) upstream of exon 3, leading to a transcript with 4 additional extracellular amino acids, has been previously detected in a family of Scottish descent and a family of Arab descent with TRAPS (Aksentijevich *et al.* 2001). Interestingly, the novel SNP (IVS6-32C>A) detected in this study also leads to formation of an “AG” site, but this

observation requires further statistical analyses in larger study populations and functional experiments before drawing any conclusions. Furthermore, another *TNFRSF1A* splice site variation (intron 4; c. 472+1G>A) was discovered in a mother and son suffering from TRAPS (Churchman *et al.* 2008). This variation leads to insertion of 45 nucleotides of intronic DNA into mRNA causing the addition of 15 amino acids and deletion of a cysteine residue from the TNFR1 protein. This structural change causes decreased levels of soluble and cell surface TNFR1 with increased levels of pro- and anti-inflammatory cytokines, and increased NF- κ B activation.

Furthermore, this novel SNP was also detected by the 1000 Genomes Project exome sequencing (Phase I, 20110521 data freeze) and added to the dbSNP database (rs200415260, entry date: 10/02/2012). The 1000 Genomes Project data also support the genotyping results observed in this study, with a MAF (A) of 0.001 in 2256 sampled chromosomes. Interestingly, the minor allele was only detected in the Tuscan population in the 1000 Genomes Project. Moreover, two other very rare SNPs (rs199601025, MAF (A)=0.001; rs4149643, MAF (A)=0.002) were identified 3 base pairs upstream and downstream of the novel SNP (rs200415260) in the Southern Han Chinese and African (African-American/Yoruba) populations, respectively. The SNP rs4149643 observed in African populations is known to be present in two *TNFRSF1A* splice variants that undergo nonsense-mediated decay (Ensemble: ENST00000534885 and ENST00000543995). However, there is no further published information on the more recently discovered SNPs (rs200415260 and rs199601025).

A weak association between *TNFRSF1A* and AS was also identified in parallel in the TASC GWAS (Reveille *et al.* 2010). The most strongly associated SNP was rs1800693 ($P = 6.9 \times 10^{-5}$, OR=0.85, 95% CI 0.79-0.91), which is in LD ($r^2 > 0.8$, Figure 4.1 & 4.2) with rs4149576, a SNP that was not associated with AS in this study, possibly requiring a larger study population. The population attributable risk (PAR) for the most strongly associated *TNFRSF1A* SNPs in the meta-analysis, rs4149577 (PAR 6%), is substantially lower than those for *ERAP1* (rs2287987 OR~0.7, PAR~26%) or *IL23R* (rs11209032 OR~1.3, PAR~9%). This is unsurprising given the lower OR (0.91), but it is similar to the -6.5% that was estimated for rs1800693 from the published TASC data (Burton *et al.* 2007; Reveille *et al.* 2010). However, although the PAR values are low, this does not diminish the potential biological significance of this association. Given the striking therapeutic effect of anti-TNF therapy, there is already strong functional evidence that signalling through this pathway is likely to be of major importance in AS (Davis *et al.* 2008).

4.4.3.1. *TNFRSF1A* Variants & Inflammatory Diseases

It is of considerable interest that mutations in *TNFRSF1A* are implicated in the auto-inflammatory disease TRAPS. Study of TRAPS mutations has provided useful information about the relationship between the protein structure and function. Most mutations (>90%, ~17) are located in the first two cysteine-rich subdomains (exons 2-4). For example, 9 of the mutations associated with TRAPS disrupt the conserved extracellular disulfide bonds leading to reduced activation-induced receptor shedding and increased surface expression (McDermott *et al.* 1999; Galon *et al.* 2000; Aksentijevich *et al.* 2001). Even

though the SNP rs4149584 (also associated with AS in Study B) is also located in exon 4 (Arg92Gln), it is not causing lower receptor shedding in monocytes from TRAPS patients implying that there are different disease mechanisms (Savic *et al.* 2012). The presence of such different disease mechanism may explain why some TRAPS patients do not respond well to anti-TNF therapy, and such findings may also be translated into treatment of AS.

TNFRSF1A associations in other inflammatory diseases, such as IBD, have also been previously identified (Waschke *et al.* 2005; Lappalainen *et al.* 2008; De Jager *et al.* 2009). In a Finnish study, a weak association between familial ulcerative colitis (UC), and *TNFRSF1A* synonymous coding variant rs767455 and non-coding variant rs1800693 was observed ($p=0.007$; $p=0.042$, respectively) (Lappalainen *et al.* 2008). *TNFRSF1A* variants also showed a trend towards association with familial clustering of UC and rs767455 was also associated with CD in a Caucasian study ($p=0.0019$) (Waschke *et al.* 2005). Furthermore, TNFR1, but not TNFR2, was required for development of ileitis and arthritis in a mouse model in which TNF α was constitutively over-expressed (Kontoyiannis *et al.* 1999; Armaka *et al.* 2008).

In MS, two independent associations were observed, one with the common SNP rs1800693 and the second with the low frequency non-synonymous coding SNP rs4149584 (also associated with AS in Study B) in a large genome-wide meta-analysis study (De Jager *et al.* 2009). However, in a more recent study of MS, rs1800693 was found to be the primary functional variant in *TNFRSF1A* following statistical analyses and functional experiments (Gregory *et al.* 2012).

This intronic SNP (rs1800693, allele G, OR=1.12) leads to a truncated TNFR1 protein without exon 6 (6Δ-TNFR1). This mutant protein acts like soluble TNFR1, antagonising TNF signalling. It was concluded that this finding provides a plausible explanation why anti-TNF therapy is not effective in MS patients with this mutant protein. These observations in MS may mean that rs1800693 G allele may play a protective role in AS, as observed in the TASC GWAS (G allele, OR=0.85) (Reveille *et al.* 2010). However, it is necessary to carry out functional studies before drawing such a conclusion. Still, the primary *TNFRSF1A* SNP associated with AS may not be rs1800693 only, even it is the most strongly associated variant in the TASC GWAS. The low frequency non-synonymous coding variant rs4149584, which is in low LD with rs1800693 ($r^2=0.039$, HapMap release 22) may play a role in AS as well. The other AS associated SNPs in this study, rs4149577 and rs4149578, are in some degree of LD ($r^2=0.484$ and $r^2=0.106$, HapMap release 22) with rs1800693 and therefore, these associations may result from this LD structure.

In summary, there is evidence that dysregulation of signalling through the TNF α /TNFR1 axis as a result of mutations in *TNFRSF1A* can cause auto-inflammatory diseases, such as TRAPS, and that variation in this gene is associated with AS (Reveille *et al.* 2010; Davidson *et al.* 2011), IBD (Waschke *et al.* 2005; Lappalainen *et al.* 2008) and MS (De Jager *et al.* 2009; Gregory *et al.* 2012). A region on chromosome 16 containing a second member of the TNF signalling pathway, TRADD, is associated with AS (Pointon *et al.* 2010), suggesting that this pathway may be important in AS. To date the region of association on chromosome 16 is poorly defined and no clearly causal mutations

have yet been identified in *TRADD*. In view of the weak association between *TNFRSF1A* and AS in both the UK and Chinese (Davidson *et al.* 2011), further examination of the relationship of this locus with AS is clearly merited. Larger meta-analyses with more SNPs are required to confirm the association robustly. Understanding the functional consequences of *TNFRSF1A* variants could provide crucial information about its role in the pathogenesis of AS. With supporting evidence from this study, *TNFRSF1A* is now a candidate gene, identifying the causal SNP(s) will allow the refinement of therapies that target this pathway.

4.5. References

- Aggarwal, B. B. (2003). "Signalling pathways of the TNF superfamily: a double-edged sword." Nat Rev Immunol **3**(9): 745-756.
- Aksentijevich, I., Galon, J., *et al.* (2001). "The tumor-necrosis-factor receptor-associated periodic syndrome: new mutations in TNFRSF1A, ancestral origins, genotype-phenotype studies, and evidence for further genetic heterogeneity of periodic fevers." Am J Hum Genet **69**(2): 301-314.
- Armaka, M., Apostolaki, M., *et al.* (2008). "Mesenchymal cell targeting by TNF as a common pathogenic principle in chronic inflammatory joint and intestinal diseases." J Exp Med **205**(2): 331-337.
- Bazzoni, F., Gatto, L., *et al.* (2000). "Identification of novel polymorphisms in the human TNFR1 gene: distribution in acute leukemia patients and healthy individuals." Immunogenetics **51**(2): 159-163.
- Blonska, M., Shambharkar, P. B., *et al.* (2005). "TAK1 is recruited to the tumor necrosis factor-alpha (TNF-alpha) receptor 1 complex in a receptor-interacting protein (RIP)-dependent manner and cooperates with MEKK3 leading to NF-kappaB activation." J Biol Chem **280**(52): 43056-43063.
- Braun, J., Bollow, M., *et al.* (1995). "Use of immunohistologic and in situ hybridization techniques in the examination of sacroiliac joint biopsy specimens from patients with ankylosing spondylitis." Arthritis Rheum **38**(4): 499-505.
- Burton, P. R., Clayton, D. G., *et al.* (2007). "Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants." Nat Genet **39**(11): 1329-1337.
- Churchman, S. M., Church, L. D., *et al.* (2008). "A novel TNFRSF1A splice mutation associated with increased nuclear factor kappaB (NF-kappaB) transcription factor activation in patients with tumour necrosis factor receptor associated periodic syndrome (TRAPS)." Ann Rheum Dis **67**(11): 1589-1595.
- Davidson, S. I., Liu, Y., *et al.* (2011). "Association of STAT3 and TNFRSF1A with ankylosing spondylitis in Han Chinese." Ann Rheum Dis **70**(2): 289-292.
- Davis, J. C., Jr., van der Heijde, D. M., *et al.* (2008). "Efficacy and safety of up to 192 weeks of etanercept therapy in patients with ankylosing spondylitis." Ann Rheum Dis **67**(3): 346-352.
- De Jager, P. L., Jia, X., *et al.* (2009). "Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci." Nat Genet **41**(7): 776-782.
- Galon, J., Aksentijevich, I., *et al.* (2000). "TNFRSF1A mutations and autoinflammatory syndromes." Curr Opin Immunol **12**(4): 479-486.
- Gregory, A. P., Dendrou, C. A., *et al.* (2012). "TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis." Nature **488**(7412): 508-511.
- Jones, S. J., Ledgerwood, E. C., *et al.* (1999). "TNF recruits TRADD to the plasma membrane but not the trans-Golgi network, the principal subcellular location of TNF-R1." J Immunol **162**(2): 1042-1048.
- Kim, S., Moon, S. M., *et al.* (2008). "TNFR1 promoter -329G/T polymorphism results in allele-specific repression of TNFR1 expression." Biochem Biophys Res Commun **368**(2): 395-401.

- Kontoyiannis, D., Pasparakis, M., *et al.* (1999). "Impaired on/off regulation of TNF biosynthesis in mice lacking TNF AU-rich elements: implications for joint and gut-associated immunopathologies." Immunity **10**(3): 387-398.
- Kremer, J. M., Westhovens, R., *et al.* (2003). "Treatment of rheumatoid arthritis by selective inhibition of T-cell activation with fusion protein CTLA4Ig." N Engl J Med **349**(20): 1907-1915.
- Kumpfel, T., Hoffmann, L. A., *et al.* (2008). "Multiple sclerosis and the TNFRSF1A R92Q mutation: clinical characteristics of 21 cases." Neurology **71**(22): 1812-1820.
- Lange, U., Teichmann, J., *et al.* (2000). "Correlation between plasma TNF-alpha, IGF-1, biochemical markers of bone metabolism, markers of inflammation/disease activity, and clinical manifestations in ankylosing spondylitis." Eur J Med Res **5**(12): 507-511.
- Lappalainen, M., Halme, L., *et al.* (2008). "Association of IL23R, TNFRSF1A, and HLA-DRB1*0103 allele variants with inflammatory bowel disease phenotypes in the Finnish population." Inflamm Bowel Dis **14**(8): 1118-1124.
- Matsukura, H., Ikeda, S., *et al.* (2008). "Genetic polymorphisms of tumour necrosis factor receptor superfamily 1A and 1B affect responses to infliximab in Japanese patients with Crohn's disease." Aliment Pharmacol Ther **27**(9): 765-770.
- McDermott, M. F., Aksentijevich, I., *et al.* (1999). "Germline mutations in the extracellular domains of the 55 kDa TNF receptor, TNFR1, define a family of dominantly inherited autoinflammatory syndromes." Cell **97**(1): 133-144.
- Micheau, O., Lens, S., *et al.* (2001). "NF-kappaB signals induce the expression of c-FLIP." Mol Cell Biol **21**(16): 5299-5305.
- Pointon, J. J., Harvey, D., *et al.* (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**(6): 1243-1246.
- Poirier, O., Nicaud, V., *et al.* (2004). "Polymorphism R92Q of the tumour necrosis factor receptor 1 gene is associated with myocardial infarction and carotid intima-media thickness--the ECTIM, AXA, EVA and GENIC Studies." Eur J Hum Genet **12**(3): 213-219.
- Reveille, J. D., Sims, A. M., *et al.* (2010). "Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci." Nat Genet **42**(2): 123-127.
- Savic, S., Dickie, L. J., *et al.* (2012). "Familial Mediterranean fever and related periodic fever syndromes/autoinflammatory diseases." Curr Opin Rheumatol **24**(1): 103-112.
- Schall, T. J., Lewis, M., *et al.* (1990). "Molecular cloning and expression of a receptor for human tumor necrosis factor." Cell **61**(2): 361-370.
- Sica, A., Dorman, L., *et al.* (1997). "Interaction of NF-kappaB and NFAT with the interferon-gamma promoter." J Biol Chem **272**(48): 30412-30420.
- Son, Y. H., Jeong, Y. T., *et al.* (2008). "Roles of MAPK and NF-kappaB in interleukin-6 induction by lipopolysaccharide in vascular smooth muscle cells." J Cardiovasc Pharmacol **51**(1): 71-77.

- Stahl, E. A., Raychaudhuri, S., *et al.* (2010). "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci." Nat Genet **42**(6): 508-514.
- Stephens, M., Sloan, J. S., *et al.* (2006). "Automating sequence-based detection and genotyping of SNPs from diploid samples." Nat Genet **38**(3): 375-381.
- Swaminathan, B., Matesanz, F., *et al.* (2010). "Validation of the CD6 and TNFRSF1A loci as risk factors for multiple sclerosis in Spain." J Neuroimmunol **223**(1-2): 100-103.
- Waschke, K. A., Villani, A. C., *et al.* (2005). "Tumor necrosis factor receptor gene polymorphisms in Crohn's disease: association with clinical phenotypes." Am J Gastroenterol **100**(5): 1126-1133.
- Yang, J., Lin, Y., *et al.* (2001). "The essential role of MEKK3 in TNF-induced NF-kappaB activation." Nat Immunol **2**(7): 620-624.

CHAPTER 5

5. USING IMPUTATION TO FINE MAP KNOWN REGIONS & IDENTIFY LOW FREQUENCY VARIANTS ASSOCIATED WITH AS

5.1. Introduction

Genomewide association studies (GWAS) have proved remarkably successful in identifying genetic variants underlying complex diseases including ankylosing spondylitis (AS). There are now fourteen reliably confirmed regions including HLA-B27 that are known to contain genetic variants that affect risk of AS in individuals of European descent (Burton *et al.* 2007; Reveille *et al.* 2010; Evans *et al.* 2011). In the Wellcome Trust Case Control Consortium 2 (WTCCC2) genomewide association study (GWAS), associations with *ERAP1*, *IL23R*, *KIF21B*, *HLA-B*, *2p15*, *21q22*, *RUNX3*, *IL12B*, *LTBR-TNFRSF1A*, *ANTXR2*, *PTGER4*, *CARD9* and *TBKBP1* were observed along with the *HLA-B27-ERAP1* interaction in AS (Evans *et al.* 2011). But, the majority of strongly associated variants within these regions have been common (i.e. MAF >5%)- with the possible exception of non-synonymous coding SNP rs11209026 with a minor allele frequency (MAF) of 4% in the *IL23R* region.

However, like most complex diseases, the majority (>60%) of genetic variation underlying the condition remains unexplained, a phenomenon known as the

“missing heritability” problem (Maher 2008; Manolio *et al.* 2009). One explanation for this missing heritability is the existence of low frequency variants (MAF<5%), which predispose to disease. These lower frequency variants tend not to be present on genomewide SNP arrays, nor are they well tagged by the markers on these genotyping platforms. This is because of weak correlations between common tagging SNPs and rare variants due to their different minor allele frequencies (Bodmer & Bonilla 2008). Furthermore, larger sample sizes are required to detect association with lower frequency variants making their detection more difficult.

Genotype imputation is a statistical method of inferring genotypes of variants not directly genotyped in a GWAS, including low frequency variants. Imputation makes fine mapping faster and cheaper, as one can obtain imputed genotypes of many variants just by downloading publicly available haplotype data (e.g. HapMap or 1000 Genomes) from a reference population and employing one of the many popular imputation programmes, such as MACH. However, it is important to use an appropriate strategy for selection of the reference population (e.g. CEU population for Northern and Western Europeans; combination of different ethnicities for particularly imputing rarer variants) because of the different LD structure in different populations and to apply necessary technical measures to ensure imputation quality (Zhang *et al.* 2011).

The HapMap Project data are of relatively limited use for imputing lower frequency variants since the vast majority of SNPs in the HapMap panels are common (HapMap 2003; HapMap 2005; Frazer *et al.* 2007). The 1000 Genomes

Project aims to improve coverage of markers across the genome in particular low frequency and rare variants using high throughput sequencing technology. The pilot phase, intended to determine the best sequencing strategies, was completed in June 2009; data from low coverage sequencing (3.6X) of 60 individuals from the Centre de'Etude du Polymorphisme Humain (CEPH) families (Utah residents of Northern and Western European ancestry – CEU) were generated (on which the results of this chapter are based). In 2.4 gigabases of accessible genome, 14.4 million single nucleotide polymorphisms (SNPs) were discovered as a result of the low coverage sequencing of a total of 179 individuals from four populations (Han Chinese, Japanese, CEU and Yoruba) (Genomes 2010).

In this chapter, the 1000 Genomes Project pilot data (August 2009) were used as a reference set to impute variants in AS cases and controls from the Wellcome Trust Case-Control Consortium 2 (WTCCC2) study. The main aims of increasing the density of the case-control data by imputation include:

- 1) Facilitating the identification of novel common variants affecting the risk of AS by imputing variants that were not included in the HapMap 2 dataset.
- 2) Fine-mapping of known regions to facilitate the discovery of variants that are more strongly associated with AS, and that may be causal.
- 3) Investigating previously associated regions for independent secondary/tertiary signals via conditional analysis.

- 4) Identifying genetic associations in HLA-B27 positive and negative individuals separately by stratifying the study population using the HLA-B27 tagging SNP rs4349859 (Evans *et al.* 2011).
- 5) Identifying low frequency variants (MAF 1 - 5%) that affect the risk of AS by using single locus tests of association and variant clustering strategies (Lawrence *et al.* 2010; Morris & Zeggini 2010; Magi *et al.* 2011), which examine the combined effects of low frequency variants within gene regions.

5.2. Patients, Materials & Methods

5.2.1. Patients & Quality Control

The study population used in this chapter was a subset of the discovery sample (only WTCCC individuals, not individuals from the TASC) from the Wellcome Trust Case-Control Consortium 2 (WTCCC2) AS GWAS (Evans *et al.* 2011). Cases and controls were genotyped with the Illumina 660W-Quad microarray and custom Illumina Human 1.2M-Duo platforms, respectively. After quality control (See Chapter 2, Section 2.7.1), there were 1782 British and Australian cases of white European ancestry fulfilling modified New York Criteria for AS (van der Linden *et al.* 1984) and 5167 historical controls (Table 5.1). The WTCCC2 data were then filtered so that only SNPs common to both genotyping platforms were included in the imputation. In addition, SNPs with poor genotyping rates (>2% missingness) were excluded from analyses as well as individuals showing cryptic relatedness, different ancestry and low genotyping rates (Table 5.1; See Chapter

2, Section 2.7.1) (Evans *et al.* 2011). Minor allele frequency (MAF) filtering was not performed on the SNPs in order to retain more information for imputation of low frequency variants.

Study popuation before filtering				
	Female	Male	Non-specified	Total
ALL	3419	4238	15	7672
Cases				2005
Controls				5667
Study population after filtering				
ALL	3134	3815	0	6949
Cases				1782
Controls				5167
Number of SNPs before & after filtering				
Total before	Included	Excluded	Total after	
1,158,122	594,398	163,557	532,140	

Table 5.1. Features of the study population and the number of SNPs upon filtering. Cases and controls were genotyped on two different platforms, Illumina 660W-Quad microarray and custom Illumina Human 1.2M-Duo chips, respectively (Evans *et al.* 2011). Therefore, when only SNPs common to both platforms were included in this study, there was a large decrease in the number of SNPs.

5.2.2. Imputation & Statistical Analysis

Imputation was carried out using MACH 1.0 (Li *et al.* 2009; Li *et al.* 2010). Haplotypes (Utah residents of Northern and Western European ancestry – CEU, n=122) from the 1000 Genomes Project pilot study (August 2009) were used as the reference panel to infer untyped SNPs *in silico* in the WTCCC2 population sample (See Chapter 2, Section 2.7) (Genomes 2010). Mach2dat was used to perform all logistic regression analyses including HLA-B27-stratified and conditional analyses (See Chapter 2, Section 2.8.1) (Li *et al.* 2009; Li *et al.* 2010). Conditional analyses were performed in previously and newly identified regions of interest with strong evidence of association ($p < 5 \times 10^{-3}$) to detect possible independent secondary/tertiary association signals. Haplotype analysis of the *ERAP1-ERAP2* region was performed with imputed “best guess” genotypes of two SNPs, rs30187 (*ERAP1*) and rs4869313 (*ERAP2*), using PLINK (v1.07) in order to determine the haplotype structure in the region (See Chapter 2, Section 2.6.2) (Purcell *et al.* 2007). Imputed SNPs with imputation quality (RSQR) ≥ 0.3 were included in these analyses. However, all SNPs with RSQR < 0.5 were treated with caution as recommended in the MACH guidelines (http://genome.sph.umich.edu/wiki/MaCH:_1000_Genomes_Imputation_Cookbook). Manhattan and Q-Q plots were produced using R. Inflation factor (λ) calculations were also carried out with R (See Chapter 2, Section 2.8.1.2). Results of the analyses were annotated using PLINK (Purcell *et al.* 2007). Regional association plots were produced using Locuszoom (Genome build: hg18, LD population: 1000 Genomes August 2009 CEU) (Pruim *et al.* 2010). Statistical power was calculated using the Quanto software (Version 1.2.4, May 2009). The study had 17%–79% power to identify SNPs conferring an log-additive allelic

odds ratio of 1.2 with minor allele frequencies (MAFs) of 0.1–0.5 and a disease prevalence of 4:1000 at $\alpha=10^{-4}$ (i.e. p-value threshold used to identify regions of interest). Moreover, the study had 6%–88% power to identify SNPs conferring an log-additive allelic odds ratio of 1.5 with MAFs of 0.01–0.05 and a disease prevalence of 4:1000 at $\alpha=10^{-4}$. Thus, this study had the statistical power necessary to detect associations with low frequency variants ($1\% \leq \text{MAF} \leq 5\%$) with modest effect sizes ($\text{OR} \geq 1.5$).

Analyses combining the rare and low frequency SNPs were performed using GRANVIL (gene- or –region-based analysis of variants of intermediate and low frequency) with imputed genotypes and CCRaVAT (case-control rare variant analysis tool) using observed genotype data from the WTCCC2 study (See Chapter 2, Sections 2.8.2 & 2.6.4) (Lawrence *et al.* 2010; Morris & Zeggini 2010; Magi *et al.* 2011). Imputed SNPs of low frequency with $\text{RSQR} \geq 0.5$ (as recommended in http://genome.sph.umich.edu/wiki/MaCH:_1000_Genomes_Imputation_Cookbook) and $\text{MAF} < 5\%$ were included in the GRANVIL analysis. All genomic locations were obtained from genome build 36 (hg18). A genomewide significance threshold of 1.7×10^{-6} was used in these analyses after applying a bonferroni correction for 30,000 genes ($\text{p-value threshold} = 0.05/30,000$) (Magi *et al.* 2012).

5.3. Results

5.3.1. Logistic Regression Analysis

5.3.1.1. Q-Q Plots & Inflation Factors

Genomewide Q-Q plots and their corresponding inflation factor values showing the deviation from the expected chi-square (χ^2) distribution are presented in Figures 5.1 and 5.2. The genomewide inflation factor was 1.05 (Figure 5.1). The inflation factor was lower (1.04), when chromosome 6 containing the MHC region was excluded from the calculation (Figure 5.2). Inflation may possibly due to genetic association, population stratification, and cryptic relatedness. However, the genomic inflation factors and the Q-Q plots indicated that population stratification and cryptic relatedness were unlikely to be substantial concerns in this dataset.

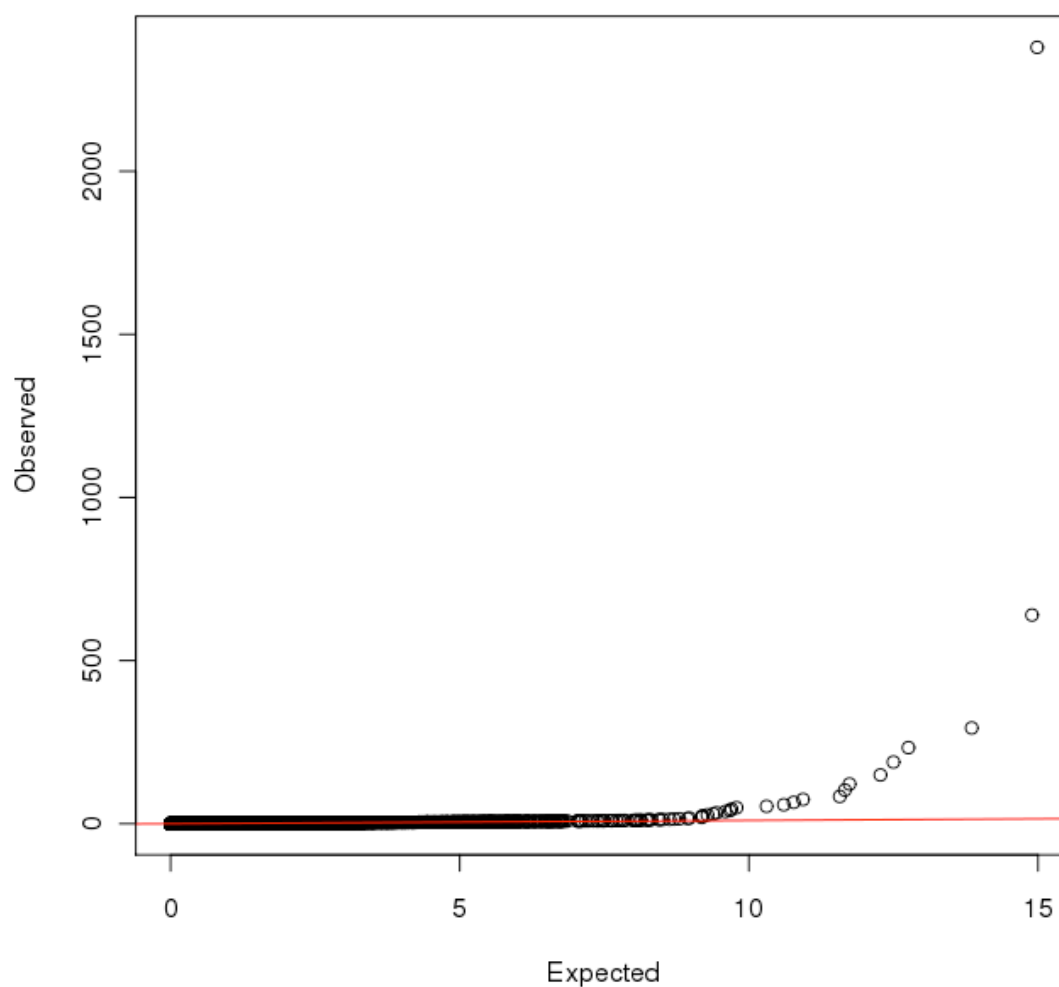


Figure 5.1. Genomewide Q-Q plot for SNPs with $RSQR \geq 0.3$. The red line and black circles mark the expected and the observed χ^2 statistics, respectively. The corresponding inflation factor is 1.05.

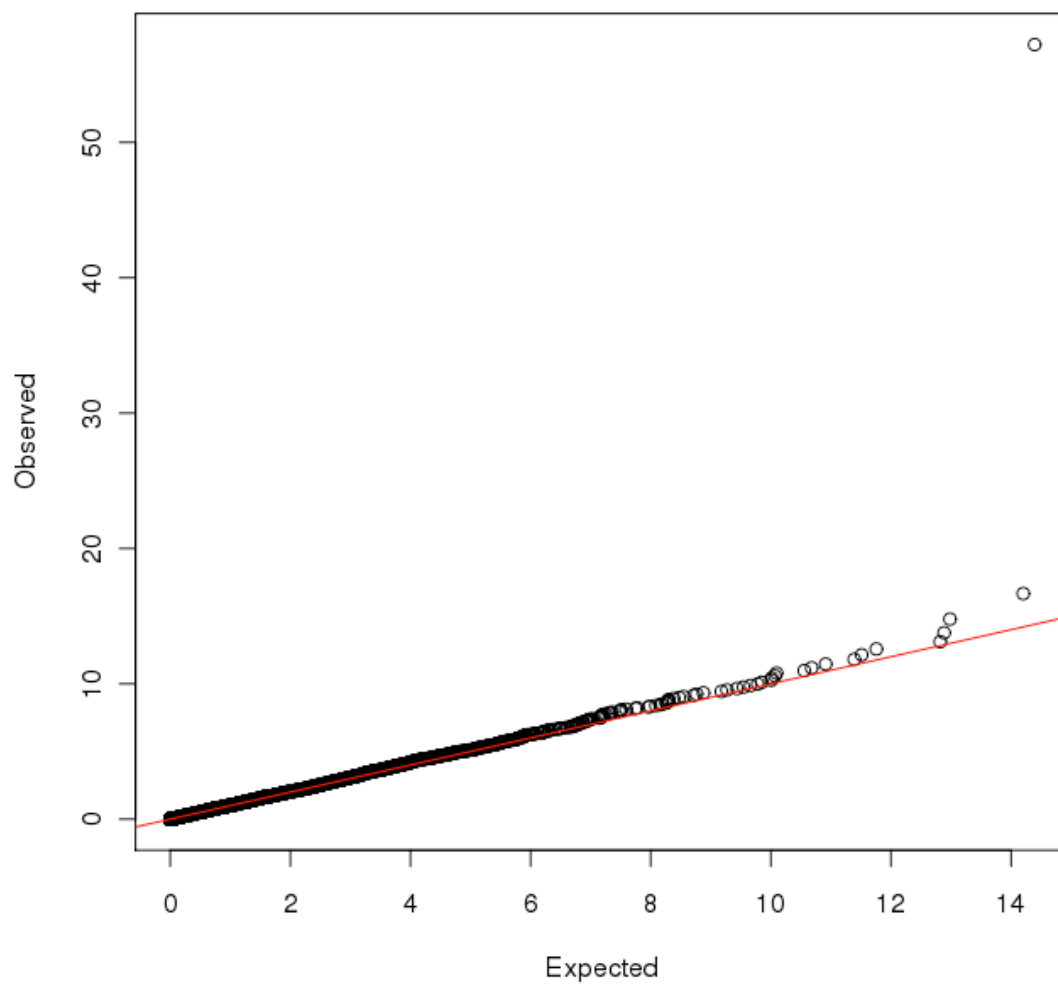


Figure 5.2. Q-Q plot for SNPs with $RSQR \geq 0.3$ excluding chromosome 6 that contains the MHC region. The red line and black circles mark the expected and the observed χ^2 statistics, respectively. The corresponding inflation factor is 1.04.

5.3.1.2. Genomewide Associations

Genomewide Manhattan plots are shown in Figure 5.3. Variants within five genes/regions achieved genome-wide significance ($p < 5 \times 10^{-8}$): *IL23R*, 2p15, *ERAP1*, *HLA-B* and 21q22 (Figure 5.3, Table 5.2). Other regions of interest containing variants with $p < 10^{-3}$ include *RUNX3*, *IL12B*, *UBE2U*, *IL6R*, *KIF21B*, *ESRRG*, *IL12RB2*, gene desert 2p16.3, *ZXDC*, *IL7R*, *CDKAL1*, *CARD9*, *LTBR*-*TNFRSF1A*, *TRADD* region, *MAP2K4*, *IL12B*, *ANTXR2*, *PTGER4*, *TBKBP1* and *UBE2L3* (Figure 5.3, Tables 5.2-5.4). A comparison of the results of this study and the same individuals from the WTCCC2 study (Evans *et al.* 2011) is provided in Tables 5.2 and 5.3. Stronger evidence of association was observed at all known genes/regions suggesting that imputation to the 1000 Genomes reference panel was successful in more precisely localising the underlying association signal, except for rs11616188 in the *TNFRSF1A*-*LTBR* region, compared to the results of the WTCCC2 study (This study: Allele G, OR=0.56, $p=0.0026$; WTCCC2: Allele A, OR=1.21, $p=1.2 \times 10^{-5}$). Further genes/regions of interest are discussed in more detail in Chapter 6.

Locuszoom plots for the genes/regions in Tables 5.2 and 5.3 are also presented below. These regional association plots show the SNP with the strongest evidence of association (i.e. top SNP) in purple and provide the LD pattern relative to this top SNP. Accuracy of imputation and validation of these results is addressed in Chapter 6..

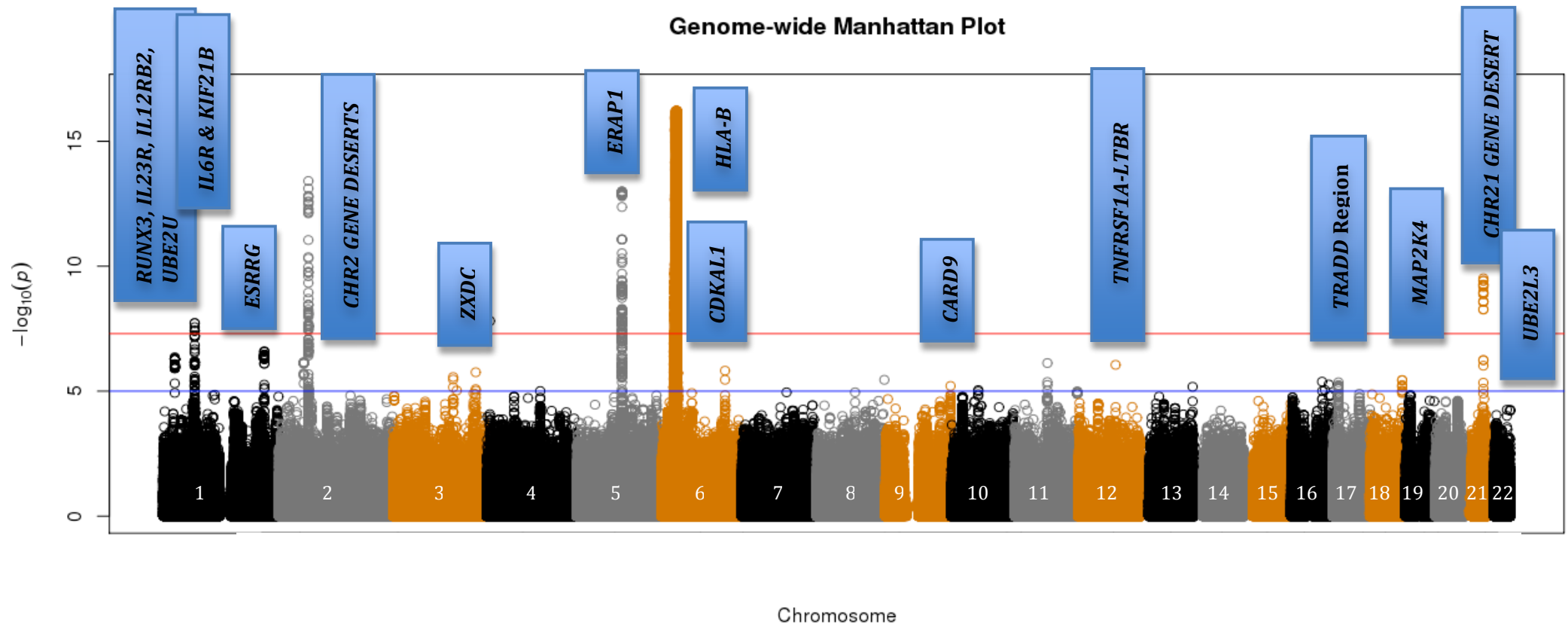
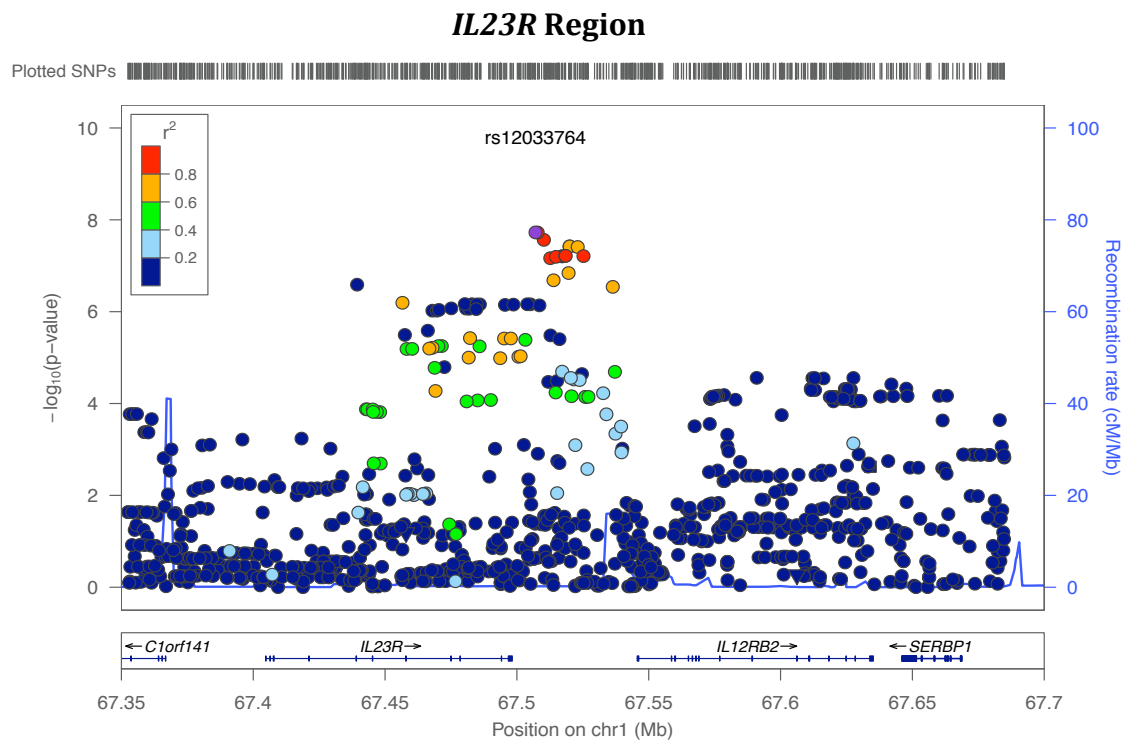


Figure 5.3. Genomewide Manhattan plot showing interesting association signals with $RSQR \geq 0.3$. The red and blue lines mark the genomewide (5×10^{-8}) and suggestive 10^{-5} p-value thresholds, respectively.

Imputed									WTCCC2				
CHR	SNP	Location	Gene/Region	Al.	Freq.	P	OR	RSQR	Al.	SNP	P	OR	r ²
1p31	rs12033764	67507070	<i>IL23R</i>	C	0.67	1.9×10^{-8}	0.79	0.91	A	rs11209032	4.9×10^{-8}	1.25	1.0
1q32	rs296548	199159466	<i>KIF21B</i>	C	0.63	4.4×10^{-5}	1.18	0.98	G	rs2297909	0.0017	1.14	0.6
2p15	rs13014154	62428947	-	A	0.38	3.9×10^{-14}	1.36	0.95	A	rs10865331	2.9×10^{-12}	1.32	0.8
5q15	rs30187	96150086	<i>ERAP1</i>	T	0.35	1.1×10^{-13}	1.35	1.00	T	rs30187	4.3×10^{-13}	1.35	-
6p21	rs4349859	31473766	<i>HLA-B</i>	G	0.85	$<10^{-300}$	0.02	1.00	A	rs4349859	$<10^{-200}$	48.9	-
21q22	rs2836878*	39387404	-	A	0.26	3.1×10^{-10}	0.75	1.00	G	rs378108	1.5×10^{-7}	1.23	0.4

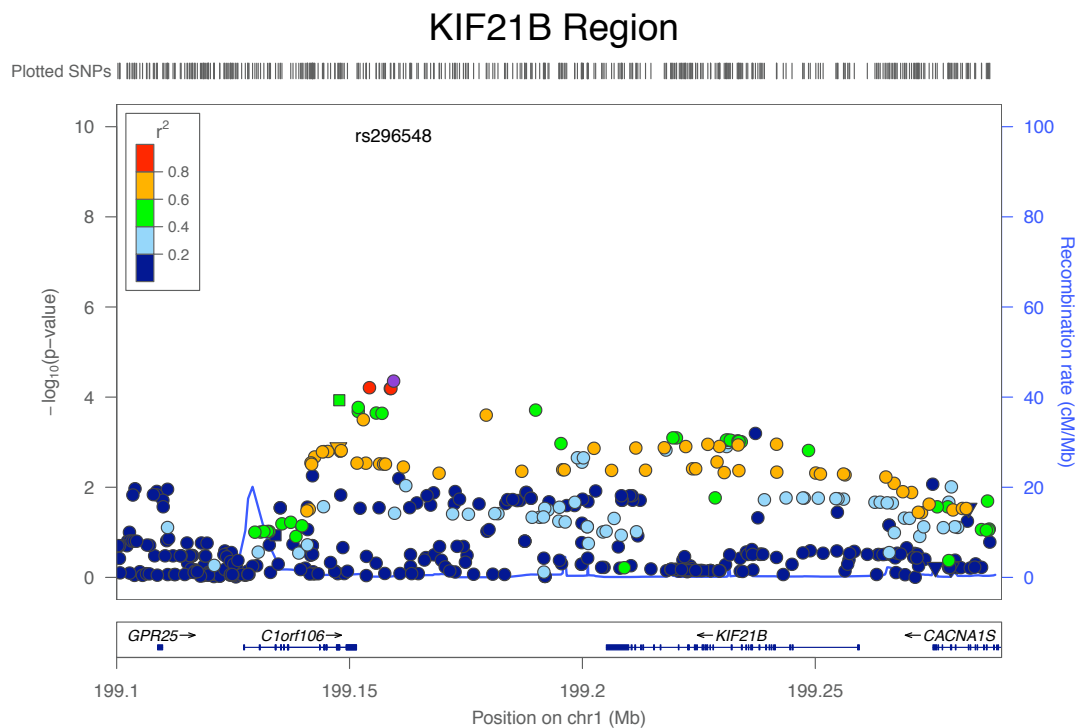
Table 5.2. Comparing most strongly associated imputed association signals in this study with those observed in the WTCCC2

discovery population in regions with previously confirmed associations (Evans *et al.* 2011). LD (r^2) values were obtained from the 1000 Genomes pilot data. LD values are listed where a different association signal (SNP) was observed compared to the WTCCC2 study. * SNP rs2836878 was associated with CD/UC (Anderson *et al.* 2011) and pediatric onset IBD (Allele A, OR=0.73, $p=6.01 \times 10^{-8}$) (Kugathasan *et al.* 2008). (Al.: Allele; Freq.: Frequency; CHR: Chromosome; RSQR: Imputation quality score)



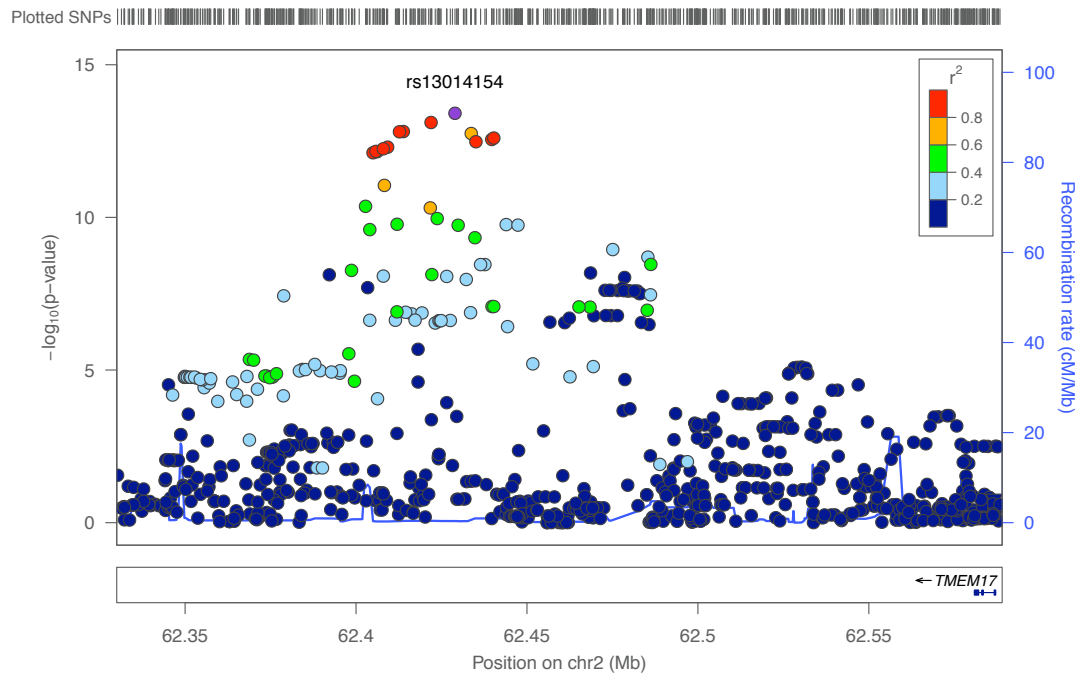
Plot 5.1. Regional association plot for the *IL23R* region on chromosome 1. The strongest association was observed with the SNP rs12033764 (Purple).

Conditional analysis revealed independent secondary and tertiary signals in this region (See Section 5.3.2). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)



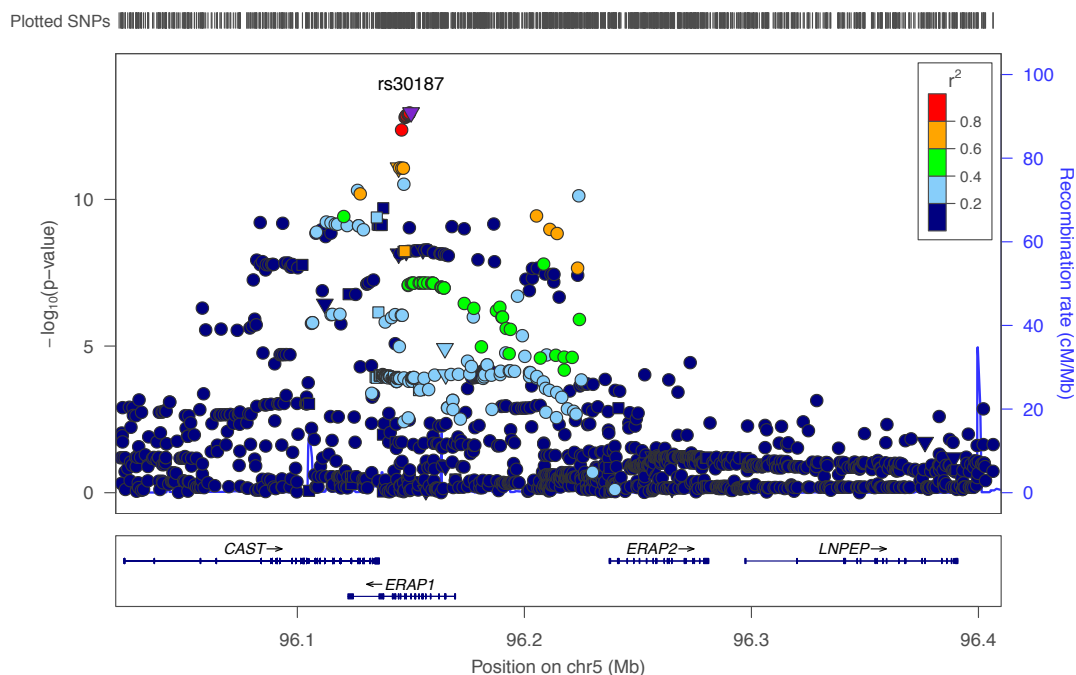
Plot 5.2. Regional association plot for the *KIF21B* region on chromosome 1. The strongest association was observed with the SNP rs296548 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

Chromosome 2 Gene Desert – 2p15

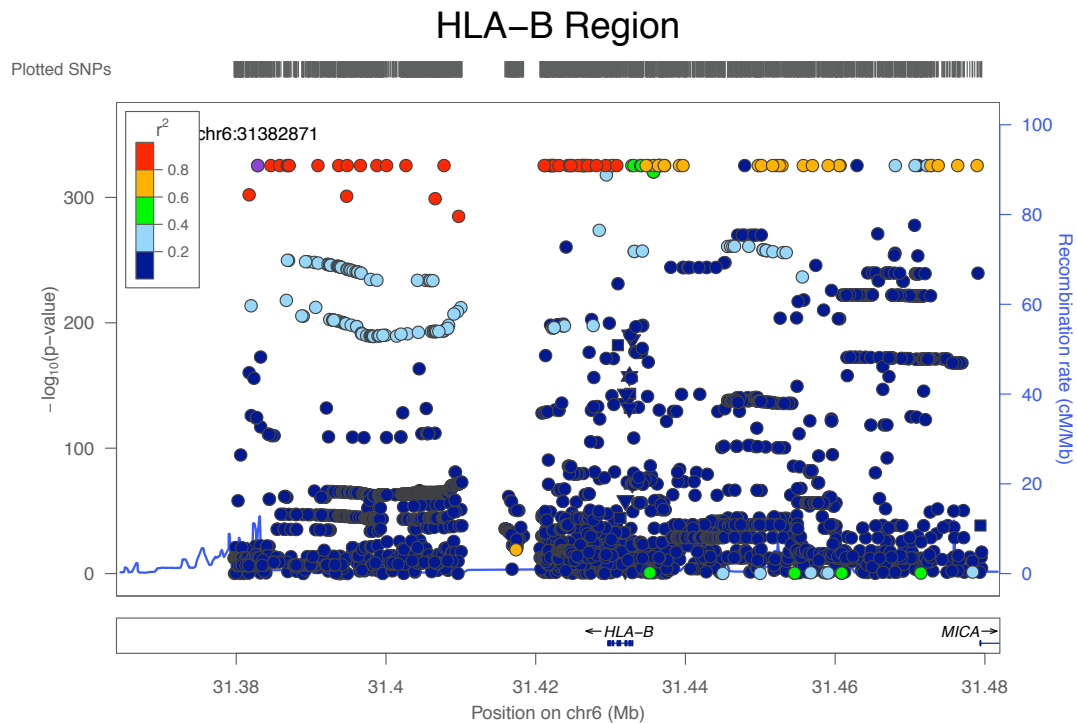


Plot 5.3. Regional association plot for the 2p15 region on chromosome 2. The strongest association was observed with the SNP rs13014154 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

ERAP1 Region

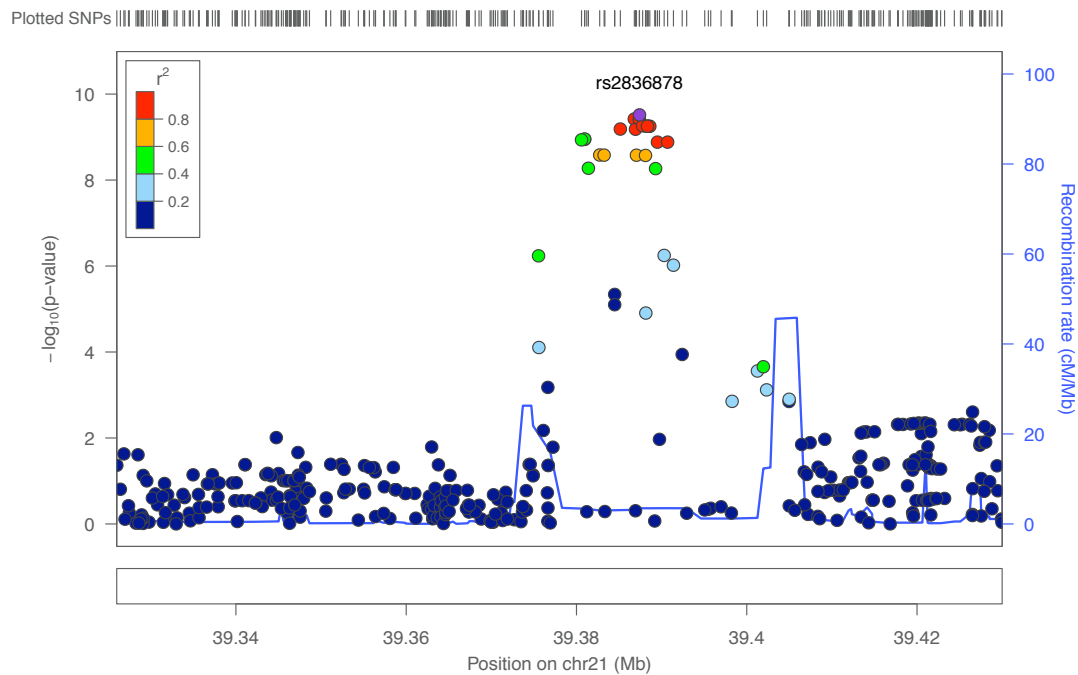


Plot 5.4. Regional association plot for the *ERAP1* region on chromosome 5. The strongest association was observed with the SNP rs30187 (Purple; Non-synonymous coding – Lysine → Arginine). Conditional analysis revealed independent secondary and tertiary signals in this region (See Section 5.3.2). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)



Plot 5.5. Regional association plot for the *HLA-B* region on chromosome 6, illustrating the high degree of linkage disequilibrium across the region. The strongest association was observed with the SNP rs4349859, but the top SNP was marked as chr6:31382871 (Purple) due to Locuszoom identifying the first SNP in the plot as a result of the ceiling effect. Conditional analysis revealed independent secondary and tertiary signals in this region (See Section 5.3.2). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

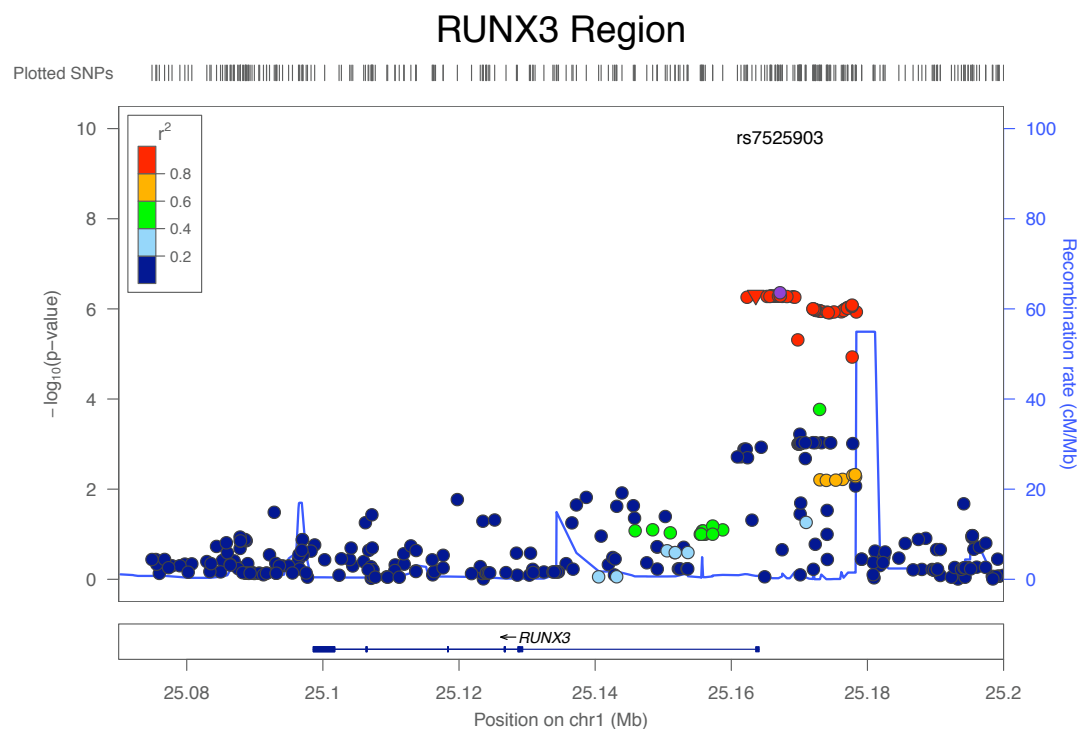
Chromosome 21 Gene Desert – 21q22



Plot 5.6. Regional association plot for the 21q22 region on chromosome 21. The strongest association was observed with the SNP rs2836878 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

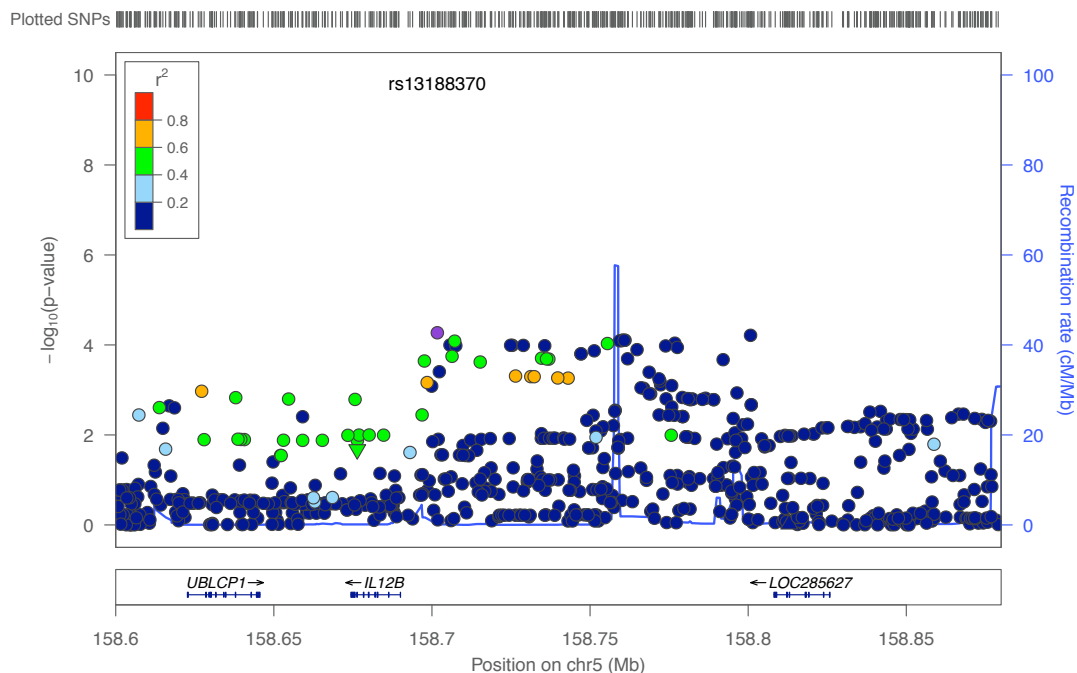
Imputed									WTCCC2				
CHR	SNP	Location	Gene/Region	Al.	Freq.	P	OR	RSQR	Al.	SNP	P	OR	r ²
1p36	rs7525903	25167160	<i>RUNX3</i>	A	0.52	4.4. x 10 ⁻⁷	1.22	0.98	A	rs11249215	2.1 x 10 ⁻⁵	1.19	0.8
5q33	rs6556416	158751323	<i>IL12B</i>	C	0.69	0.00014	1.18	0.95	C	rs6556416	1.0 x 10 ⁻⁴	1.18	-
12p13	rs11616188	6373003	<i>TNFRSF1A-</i> <i>LTBR</i>	G	0.56	0.0026	0.88	0.84	A	rs11616188	1.2 x 10 ⁻⁵	1.21	-
4q21	rs6821375	81182976	<i>ANTXR2</i>	C	0.25	0.00027	0.83	0.83	A	rs4389526	0.00054	1.15	0.6
5p13	rs9292771	40470409	<i>Ups. PTGER4</i>	T	0.63	0.0021	1.13	0.99	A	rs10440635	0.0027	1.13	0.4
9q34	rs3812565	138392323	<i>CARD9</i>	T	0.60	6.16 x 10 ⁻⁶	0.84	1.00	T	rs10781500	2.0 x 10 ⁻⁵	1.18	0.9
17q21	rs4794057	43141451	<i>TBKBP1</i>	C	0.48	0.0028	0.89	0.91	C	rs8070463	0.0025	1.13	0.9

Table 5.3. Comparing imputed association signals in this study with those observed in the WTCCC2 discovery population in regions with newly confirmed associations (Evans *et al.* 2011). LD (r²) values were obtained from the 1000 Genomes pilot data. LD values are listed where a different association signal (SNP) was observed compared to the WTCCC2 study. (Al.: Allele; Freq.: Frequency; CHR: Chromosome; RSQR: Imputation quality score; Ups.: Upstream)



Plot 5.7. Regional association plot for the *RUNX3* region on chromosome 1. The strongest association was observed with the SNP rs7525903 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

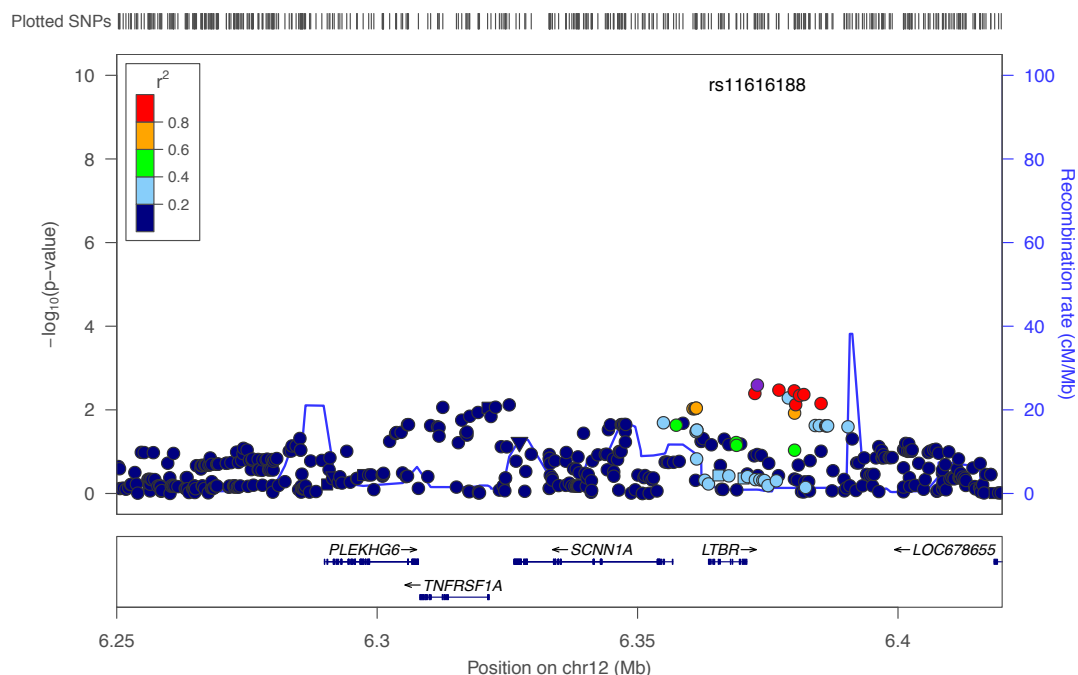
IL12B Region



Plot 5.8. Regional association plot for the *IL12B* region on chromosome 5. The strongest association was observed with the SNP rs13188370 (Purple).

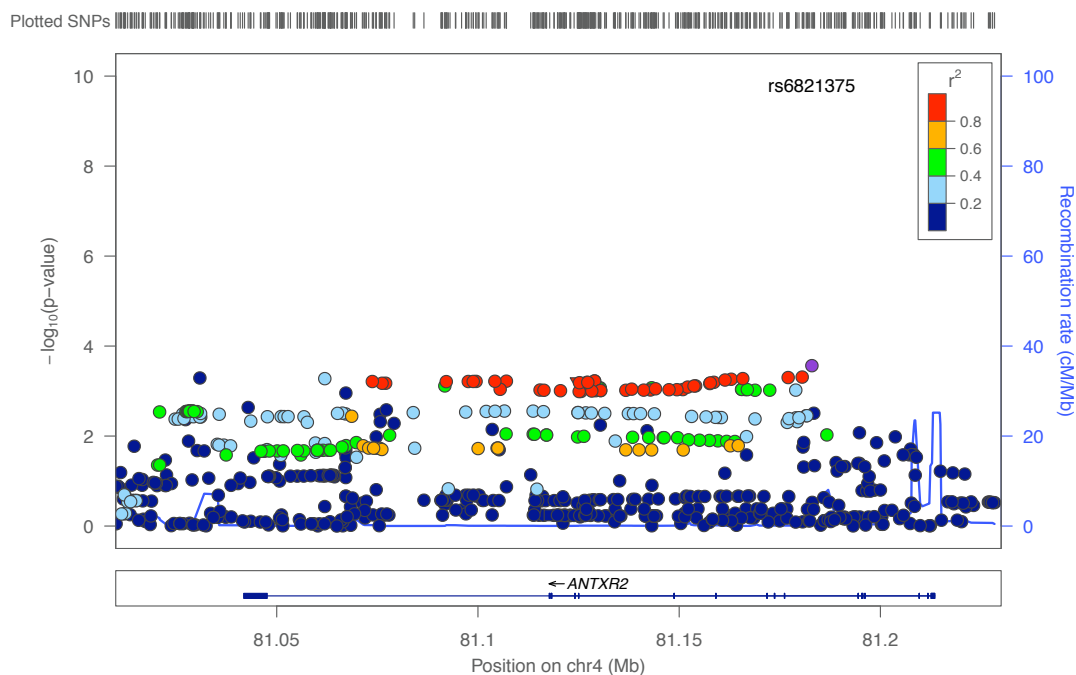
Conditional analysis revealed independent secondary and tertiary signals in this region (See Chapter 6, Section 6.2.2). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

TNFRSF1A-LTBR Region



Plot 5.9. Regional association plot for the *TNFRSF1A-LTBR* region on chromosome 12. The strongest association was observed with the SNP rs11616188 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

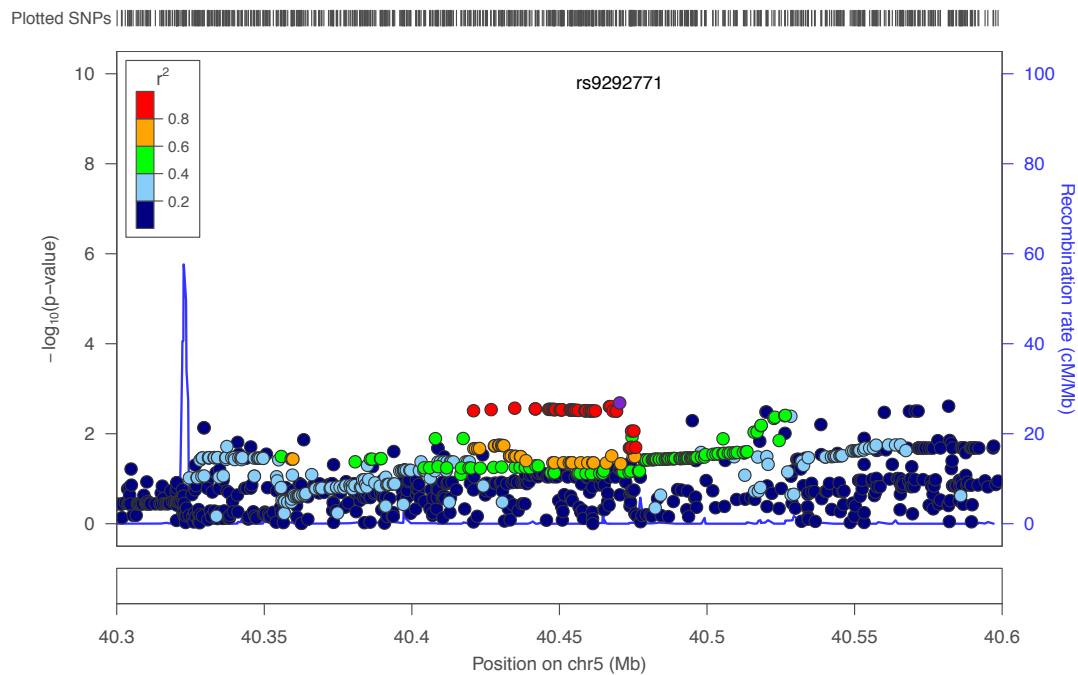
ANTXR2 Region



Plot 5.10. Regional association plot for the *ANTXR2* region on chromosome 4.

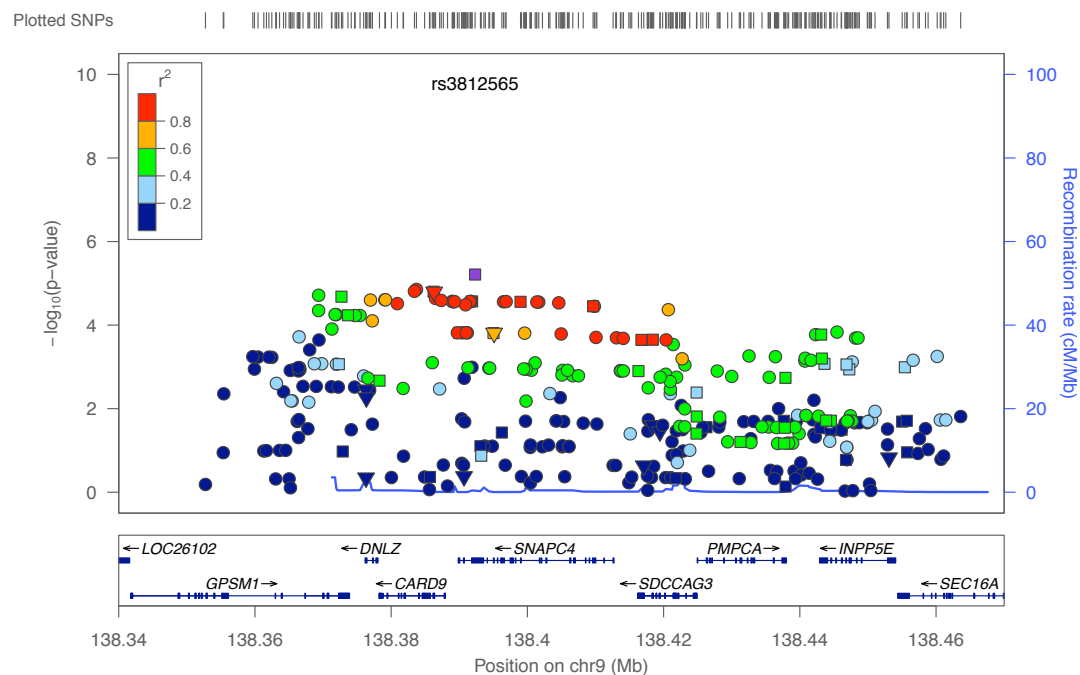
The strongest association was observed with the SNP rs6821375 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

Upstream *PTGER4*



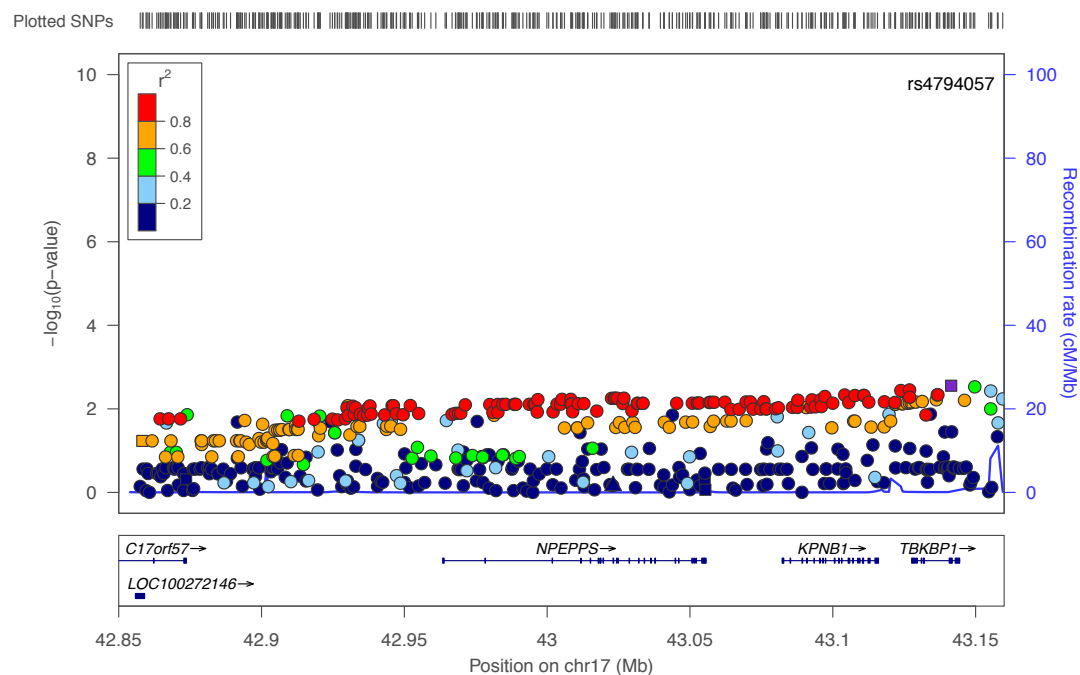
Plot 5.11. Regional association plot for the region 250 kb upstream of *PTGER4* on chromosome 5. The strongest association was observed with the SNP rs9292771 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

CARD9 Region



Plot 5.12. Regional association plot for the *CARD9* region on chromosome 9. The strongest association was observed with the SNP rs3812565 (Purple; Synonymous coding – Leucine → Leucine). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

TBKBP1 Region



Plot 5.13. Regional association plot for the *TBKBP1* region on chromosome 9.

The strongest association was observed with the SNP rs4794057 (Purple; Synonymous coding – Histidine → Histidine). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

Genes/regions of interest - Imputed								
CHR	SNP	Location	Gene	Al.	Freq.	P	OR	RSQR
1	rs825175	64452337	<i>UBE2U</i>	C	0.44	2.7×10^{-5}	0.85	1.0
1	rs6695045	152699581	<i>IL6R</i>	A	0.39	2.7×10^{-5}	1.18	0.99
1	rs6685432	67590966	<i>IL12RB2</i>	C	0.59	2.8×10^{-5}	1.19	0.97
1	rs11117811	215465206	<i>ESRRG</i>	G	0.72	2.5×10^{-7}	0.80	1.0
2	rs7591806	52700988	Gene desert 2p16.3	G	0.79	7.0×10^{-7}	0.79	0.97
3	rs12494818	127594195	<i>Downs. ZXDC</i>	T	0.88	2.7×10^{-6}	0.75	0.87
5	rs13156302	40641069	<i>Ups. PTGER4</i>	C	0.80	0.0016	0.86	0.99
5	rs1993879	35987897	<i>IL7R</i>	T	0.27	0.0005	0.86	1.0
6	rs6908425*	20836710	<i>CDKAL1</i>	T	0.22	3.6×10^{-5}	0.82	0.98
12	rs11064145**	6325359	<i>TNFRSF1A</i>	T	0.57	0.0076	1.12	0.90
16	rs9033*	65739500	<i>TRADD</i> region	G	0.44	4.2×10^{-6}	1.20	0.99
17	rs9894621	11941309	<i>MAP2K4</i>	G	0.54	4.4×10^{-6}	0.84	1.0
22	rs390408	20242272	<i>UBE2L3</i>	A	0.19	0.00019	1.20	0.99

Table 5.4. Genes/regions of interest identified in the logistic regression analysis of the imputed genotypes. * SNPs rs6908425 and rs9033 were associated with AS in previous candidate gene studies (rs6908425, OR=0.82, $p=1.1 \times 10^{-4}$; rs9033, OR=1.23, $p=2.0 \times 10^{-5}$) (Danoy *et al.* 2010; Pointon *et al.* 2010b). ** rs11064145 is in strong LD with rs1800693 ($r^2=0.754$, $D'=0.929$, HapMap release 22), which was associated with AS in TASC GWAS (Reveille *et al.* 2010)(Al.: Allele; Freq.: Frequency; CHR: Chromosome; RSQR: Imputation quality score; Ups.: Upstream; Downs.: Downstream.)

5.3.2. Conditional Analysis

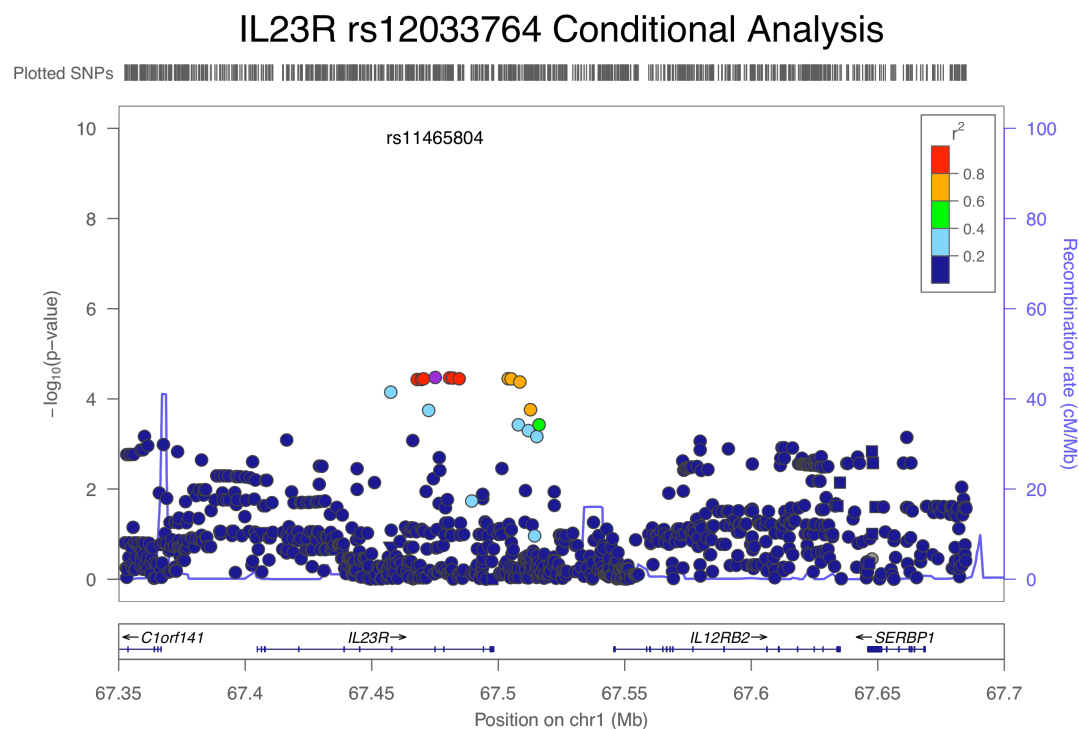
Conditional analyses were carried out in newly and previously identified regions of interest with strong evidence of association to detect possible independent signals (Tables 5.2-5.4). Independent signals ($p < 5 \times 10^{-3}$) were observed in 4 regions (Plots 5.14-5.18). Results of the conditional logistic regression analyses for 3 regions, *IL23R*, *ERAP1-ERAP2* and *HLA-B*, are listed in Table 5.5.

Independent association signals in the *IL12B* region are discussed in Chapter 6 (See Section 6.2.2). Independent secondary signals were also observed in these regions in the WTCCC2 study (Evans *et al.* 2011). In addition to these secondary signals, tertiary association signals were observed in the *IL23R* and *ERAP1-ERAP2* regions in this study.

Imputed								
CHR	SNP	Conditional on	Gene	Al.	Freq.	P	OR	RSQR
1p31	rs11465804	rs12033764	<i>IL23R</i>	T	0.94	3.3×10^{-5}	1.48	1.00
1p31	rs10889658	rs11465804-rs12033764	<i>IL23R</i>	G	0.34	0.00067	1.16	1.00
5q15	rs4869313	rs30187	<i>ERAP1-ERAP2</i>	G	0.55	7.3×10^{-8}	0.79	0.98
5q15	rs7716677	rs30187-rs4869313	<i>ERAP1</i>	T	0.73	0.00021	1.21	0.99
6p21	chr6:31435013	rs4349859	<i>HLA-B</i>	C	0.85	4.6×10^{-12}	0.05	0.96

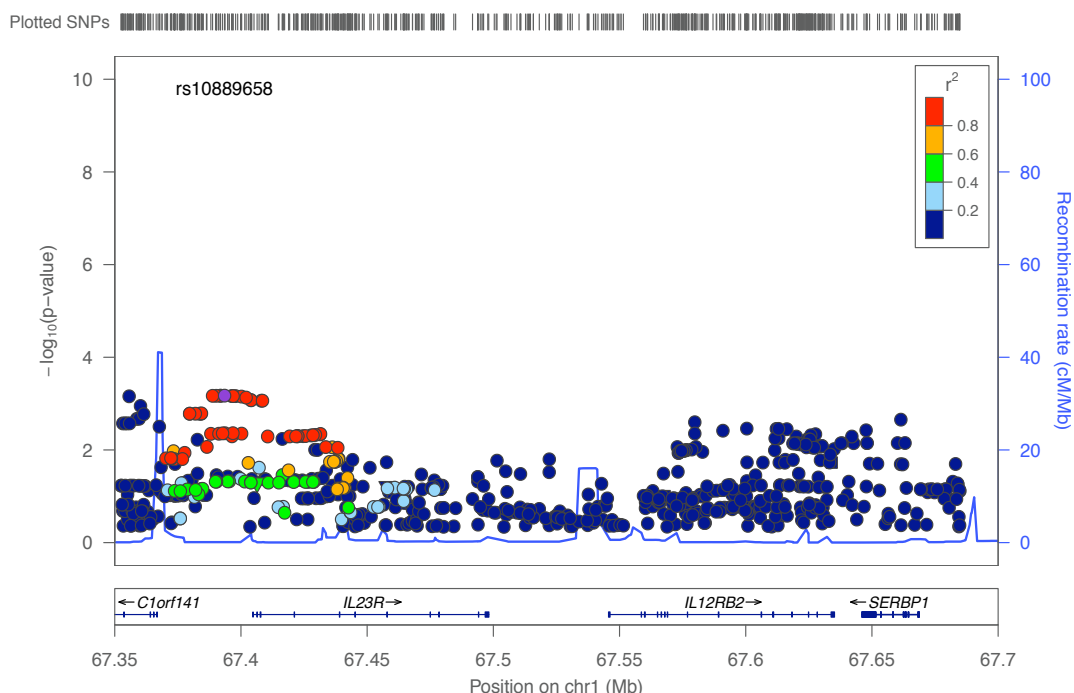
Table 5.5. Results of the conditional analyses showing residual independent associations in 3 regions, *IL23R*, *ERAP1-ERAP2* and *HLA-B*.

(Al.: Allele; Freq.: Frequency; CHR: Chromosome; RSQR: Imputation quality score.



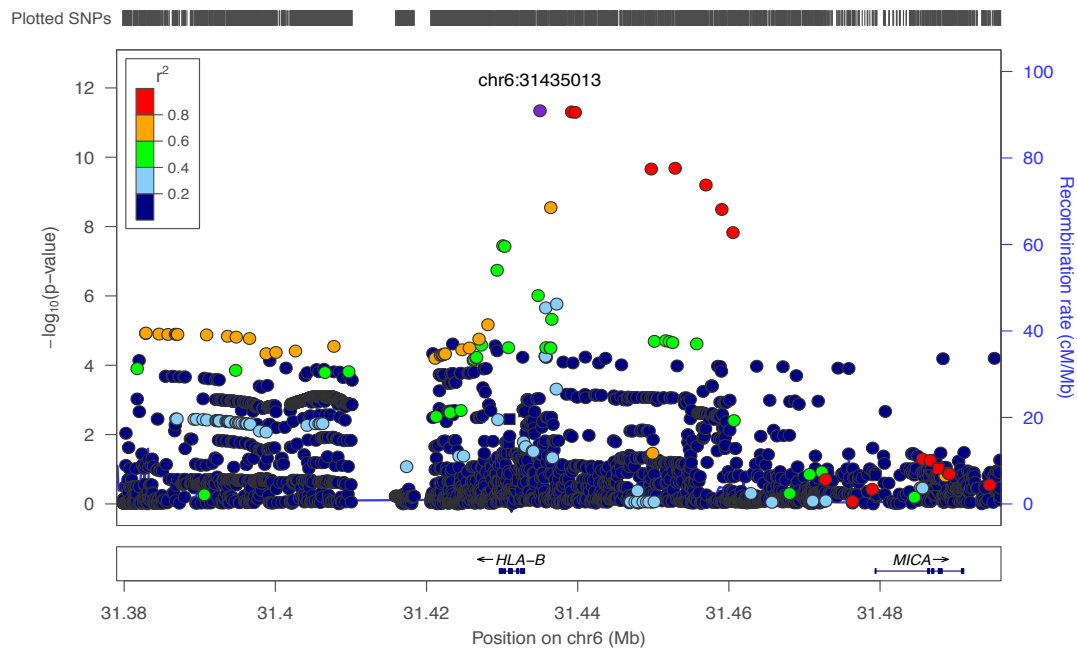
Plot 5.14. *IL23R* regional association plot after conditioning on rs12033764 (Plot 5.1) to look for additional independent association signals. The strongest association was observed with the SNP rs11465804 (Purple). Conditional analysis revealed independent secondary and tertiary signals in this region. The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

IL23R rs12033764 rs11465804 Conditional Analysis

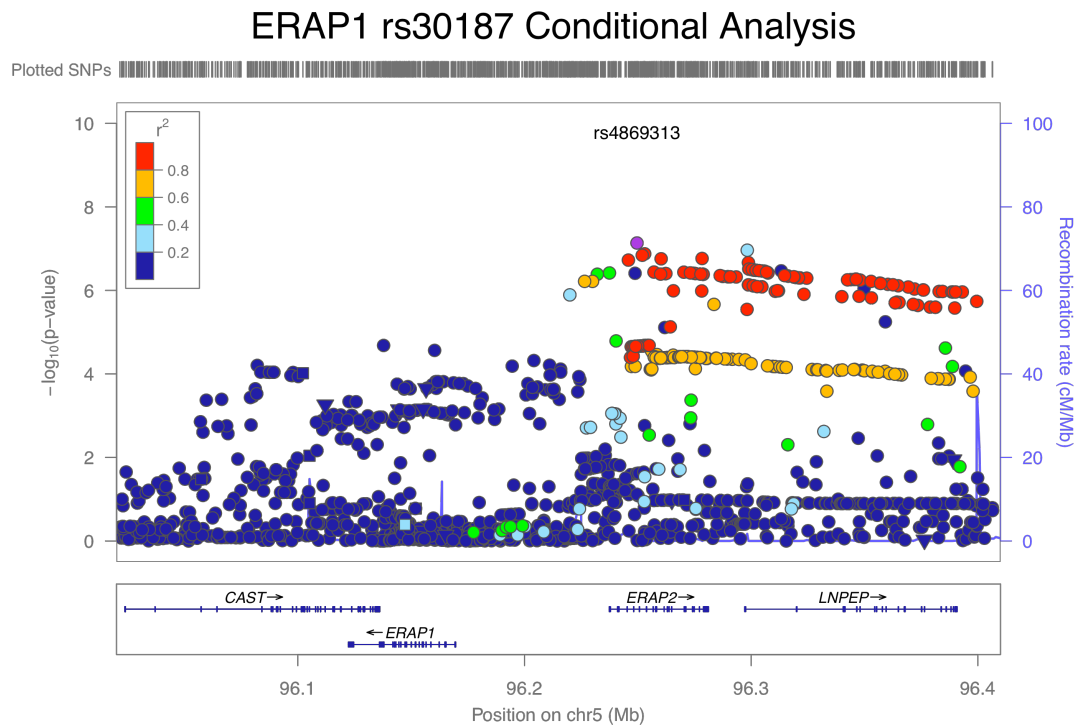


Plot 5.15. *IL23R* regional association plot after conditioning on both rs12033764 and rs11465804 to look for residual association signals. The strongest association was observed with the SNP rs10889658 (Purple). Conditional analysis revealed independent secondary and tertiary signals in this region. The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

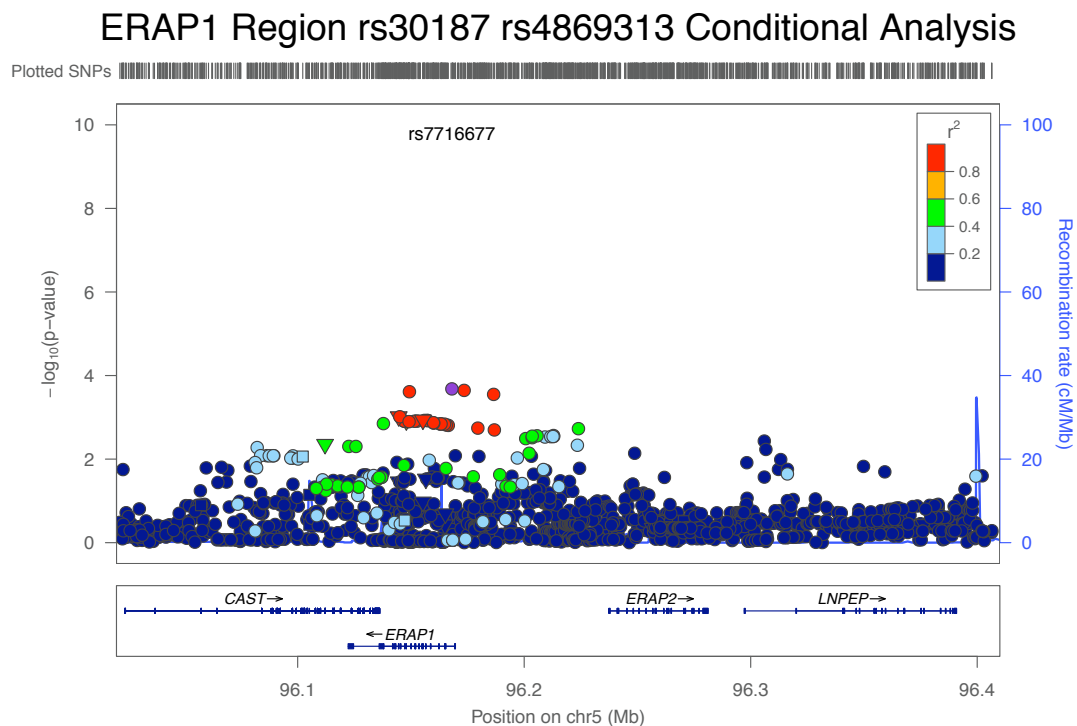
HLA-B rs4349859 Conditional Analysis



Plot 5.16. *HLA-B* regional association plot after conditioning on rs4349859 (*HLA-B*27 tagging SNP, which shows the strongest association with AS; Plot 5.5) to look for additional independent association signals. The strongest association was observed with the SNP chr6:31435013 (Purple), which may either tag *HLA-B*27 subtypes that are not well tagged by rs4349859, or perhaps other *HLA-B* alleles associated with AS (e.g. *HLA-B*60). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)



Plot 5.17. *ERAP1* regional association plot after conditioning on rs30187 (Plot 5.4) to look for additional independent association signals. The strongest association was observed with the SNP rs4869313 (Purple) in *ERAP2*. Conditional analysis revealed an independent association signal in *ERAP2*, which was masked by the effect of rs30187 (*ERAP1*) due to the haplotype structure in the region (See further analyses and explanation in Section 5.3.3). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)



Plot 5.18. *ERAP1* regional association plot after conditioning on both rs30187 (*ERAP1*) and rs4869313 (*ERAP2*) to look for residual association signals. The strongest association was observed with the SNP rs7716677 (Purple). Conditional analysis revealed independent secondary and tertiary signals in this region. The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

5.3.3. Further Analyses of the *ERAP1-ERAP2* Region

An association signal in *ERAP2* was not identified in previous AS GWAS (Burton *et al.* 2007; Reveille *et al.* 2010; Evans *et al.* 2011), but was apparent after conditioning on rs30187 (*ERAP1*) in this study (See Section 5.3.2, Plot 5.17). The intuitive explanation for why this conditional analysis increases evidence for association at *ERAP2* comes from looking at the haplotype frequencies across *ERAP1* and *ERAP2* (Table 5.6). The LD across these loci means that the risk allele at rs30187 (*ERAP1*) (T) is most often paired with the protective allele at rs4869313 (*ERAP2*) (G) and likewise the protective allele at rs30187 (*ERAP1*) (C) is most often paired with the risk allele at rs4869313 (*ERAP2*) (T). The two alleles at the different loci therefore tend to cancel each other out in an unconditioned analysis. However, after conditioning on rs30187 (*ERAP1*), an effect at *ERAP2* becomes apparent. This effect can be seen clearly in Table 5.7 where the odds ratio at rs4869313 (*ERAP2*) has been calculated across each of the rs30187 (*ERAP1*) genotypes. While the odds ratio associated with rs4869313 (*ERAP2*) was not particularly impressive in the entire sample, when individuals were stratified according to genotype at rs30187 (*ERAP1*), the odds ratio within each group was much more striking (OR>1.19). A further complication is brought about as a consequence of the interaction between HLA-B27 and *ERAP1* (See Sections 5.3.4.1 and 5.4.3). It initially appeared as if *ERAP2* might only have had an effect in HLA-B27 negative individuals (Table 5.8), and not in HLA-B27 positive individuals. However, conditioning on rs30187 (*ERAP1*) clearly showed that rs4869313 (*ERAP2*) was indeed a risk locus, even in HLA-B27 positive individuals (rs4869313 OR=1.20 conditioning on rs30187, Table 5.8).

Haplotype (rs30187- rs4869313)	Frequency in cases	Frequency in controls	Chi- square	Degree of freedom	P-value
TT	0.07	0.05	23.4	1	1.3×10^{-6}
CT	0.39	0.395	0.3	1	0.6
TG	0.33	0.28	27.5	1	1.6×10^{-7}
CG	0.21	0.275	54.8	1	1.3×10^{-13}

Table 5.6. Haplotype analysis with two SNPs rs30187 (*ERAP1*) and rs4869313 (*ERAP2*). This analysis shows that the rs30187 (*ERAP1*) risk allele T and rs4869313 (*ERAP2*) protective allele G tend to occur together (TG) more often than the double-protective haplotype (CG) and the double-risk haplotype (TT) in the study population. This haplotype (TG) also occurs more frequently in cases compared to controls (33% vs. 28%, $p=1.6 \times 10^{-7}$). Due to this haplotype structure, the effect of rs4869313 (*ERAP2*) was not apparent without conditioning on rs30187 (*ERAP2*).

All individuals			
SNP (ERAP2)	Allele	Frequency Cases	Frequency Controls
rs4869313	T	0.46	0.45
rs4869313	G	0.54	0.55
OR = 1.07			
rs30187 (ERAP1) T/T individuals			
SNP (ERAP2)	Allele	Frequency Cases	Frequency Controls
rs4869313	T	0.19	0.16
rs4869313	G	0.81	0.84
OR = 1.19			
rs30187 (ERAP1) T/C individuals			
SNP (ERAP2)	Allele	Frequency Cases	Frequency Controls
rs4869313	T	0.42	0.37
rs4869313	G	0.58	0.63
OR = 1.20			
rs30187 (ERAP1) C/C individuals			
SNP (ERAP2)	Allele	Frequency Cases	Frequency Controls
rs4869313	T	0.65	0.59
rs4869313	G	0.35	0.41
OR = 1.30			

Table 5.7. A crude stratified analysis of the frequency of rs4869313 (*ERAP2*) alleles in three different rs30187 (*ERAP1*) genotypes. rs4869313 (*ERAP2*) protective allele (G) tends to occur more often with the risk allele (T) of rs30187 (*ERAP1*), which in turn masks this protective effect of rs4869313 G allele. This pattern was also observed in HLA-B27 positive and negative populations in this study.

SNP	OR (Allele)	Study population
rs4869313 (<i>ERAP2</i>)	1.07 (T)	All
rs4869313 (<i>ERAP2</i>)	1.26 (T)	All, Conditioning on rs30187 (<i>ERAP1</i>)
rs4869313 (<i>ERAP2</i>)	1.01 (T)	HLA-B27 positives
rs4869313 (<i>ERAP2</i>)	1.20 (T)	HLA-B27 positives, Conditioning on rs30187 (<i>ERAP1</i>)
rs4869313 (<i>ERAP2</i>)	1.14 (T)	HLA-B27 negatives

Table 5.8. HLA-B27-stratified OR calculations. These ORs show that the effect of rs30187 (*ERAP1*) is masking that of rs4819313 (*ERAP2*). When this effect is eliminated in HLA-B27 positives by conditioning on rs30187 (*ERAP1*), the effect of rs4869313 (*ERAP2*) becomes apparent. Thus, the *ERAP2* association is present in both HLA-B27 negative and positive populations (HLA-B27 negative cases and controls: 256 vs. 4699; HLA-B27 positive cases and controls: 1526 vs. 445).

5.3.4. HLA-B27 Stratified Analysis

5.3.4.1. HLA-B27 Positive Associations with AS

The Q-Q plot and the corresponding inflation factor are shown in Figure 5.4 for genomewide association testing in HLA-B27 positive cases and controls only. The genomewide inflation factor was $\lambda = 1.02$ (Figure 5.4) suggesting that the effect of population stratification, cryptic relatedness and genotyping/imputation error was minimal. Since the inflation observed in the Q-Q plot is towards the end of the distribution, it is likely that this inflation is due to genetic associations.

The genomewide Manhattan plot for HLA-B27 positives is shown in Figure 5.5. There were no signals that reached the genomewide significance level ($p < 5 \times 10^{-8}$). Regions of interest which contained association signals with $p < 10^{-4}$ included upstream of *ZNF281*, *HECW2*, *ERAP1*, *PEBP4*, *CARD9* and the region around *ITGA11* (Figure 5.5, Table 5.9, Plot 5.19). These genes/regions were selected based on their functions and associations with AS and other related inflammatory disease such as IBD and psoriasis.

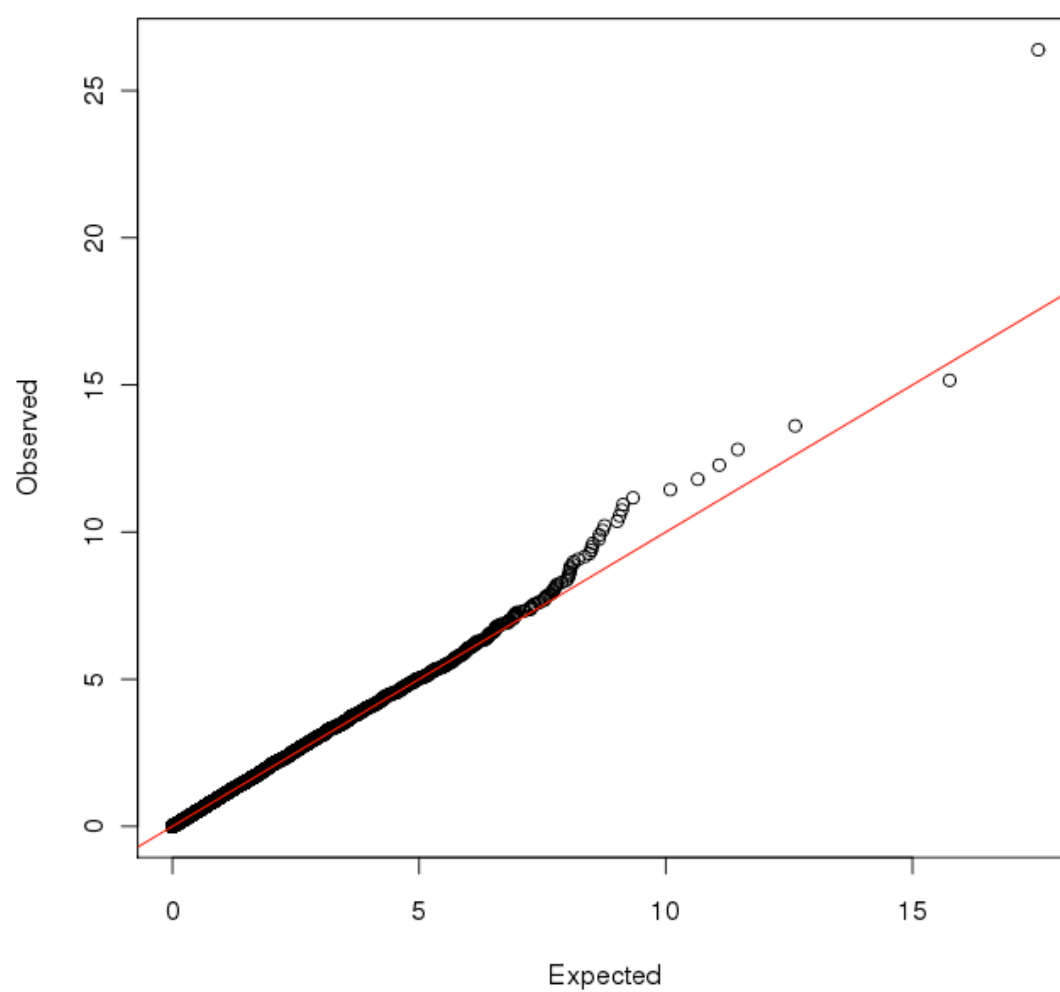


Figure 5.4. Q-Q plot for SNPs with $RSQR \geq 0.3$ in HLA-B27 positive individuals. The red line and black circles mark the expected and the observed values for the genetic tests of association, respectively. The corresponding inflation factor is 1.02.

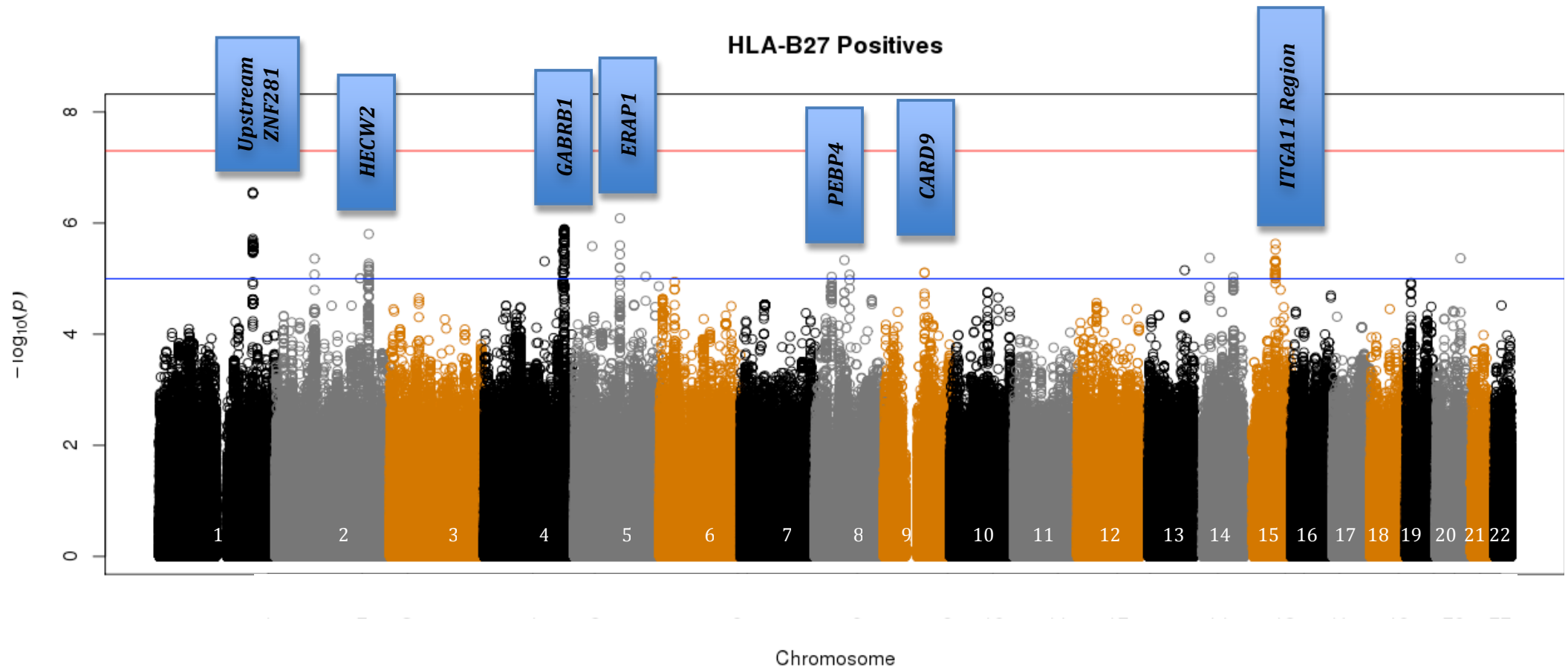
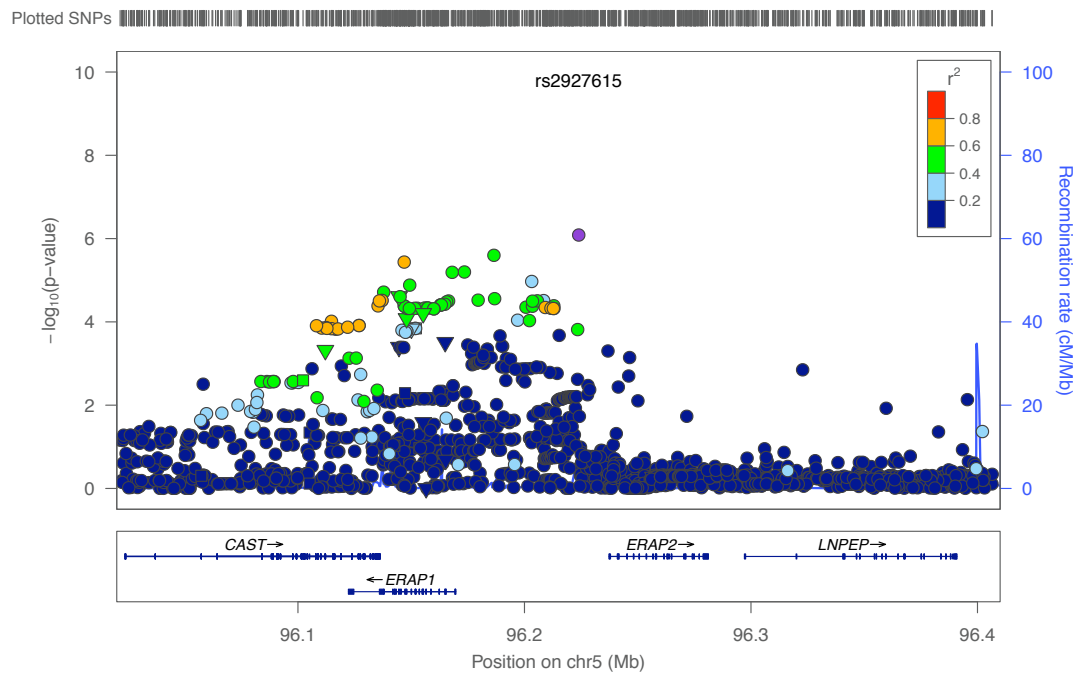


Figure 5.5. Genomewide Manhattan plot showing interesting association signals with $RSQR > 0.3$ in HLA-B27 positive individuals. The red and blue lines mark the genomewide (5×10^{-8}) and suggestive (10^{-5}) p-value thresholds, respectively.

Genes/regions of interest in HLA-B27 positives - Imputed								
CHR	SNP	Location	Gene	Al.	Freq.	P	OR	RSQR
1q32	rs16846759	198713508	<i>Upstream ZNF281</i>	C	0.86	2.8×10^{-7}	1.67	1.0
2q32	chr2:197119803	197119803	<i>HECW2</i>	A	0.89	1.6×10^{-6}	1.77	0.90
4p12	chr4:46933108	46933108	<i>GABRB1</i>	A	0.88	6.4×10^{-6}	1.63	0.99
5q15	rs2927615	96223958	<i>ERAP1</i>	G	0.74	8.2×10^{-7}	1.56	0.95
8p21	rs2457432	22807169	<i>PEBP4</i>	A	0.58	0.00013	1.34	1.0
9q34	rs4077515	138386317	<i>CARD9</i>	C	0.57	0.00028	0.75	1.0
15q23	chr15:66531585	66531585	<i>ITGA11 region</i>	C	0.58	2.4×10^{-6}	0.68	0.95

Table 5.9. Genes/regions of interest identified in the logistic regression analysis of the imputed genotypes in HLA-B27 positive individuals. (Al.: Allele; Freq.: Frequency; CHR: Chromosome; RSQR: Imputation quality score.)

***ERAP1* in HLA-B27 Positives**



Plot 5.19. *ERAP1* association in HLA-B27 positive individuals (as defined by rs4349859). This association was absent in HLA-B27 negatives (See Section 5.3.4.2, Plot 5.23). These observations show that there may be a functional interaction between *ERAP1* and HLA-B27 affecting the disease risk (Evans *et al.* 2011). The strongest association was observed with the SNP rs2927615 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

5.3.4.2. HLA-B27 Negative Associations with AS

The genomewide Q-Q plot and corresponding inflation factor for tests of association in HLA-B27 negative individuals are shown in Figure 5.6. The genomewide inflation factor was $\lambda = 0.99$ (Figure 5.6) suggesting that the effect of population stratification, cryptic relatedness and genotyping/imputation error was minimal.

The genomewide Manhattan plot for HLA-B27 negatives is shown in Figure 5.7. There were significant associations within the MHC (*HLA-B*), *CHN2* and *FOXC2* meeting the threshold for genomewide significance ($p < 5 \times 10^{-8}$) (Plots 5.20-5.22). Other regions of interest with $p < 10^{-4}$ include *ZFYVE9*, *DDX60L-PALLD*, *TRPC7* and *ERAP2* (Figure 5.7, Table 5.10, Plot 5.23). None of these associations were observed in HLA-B27 positive individuals apart from the MHC region.

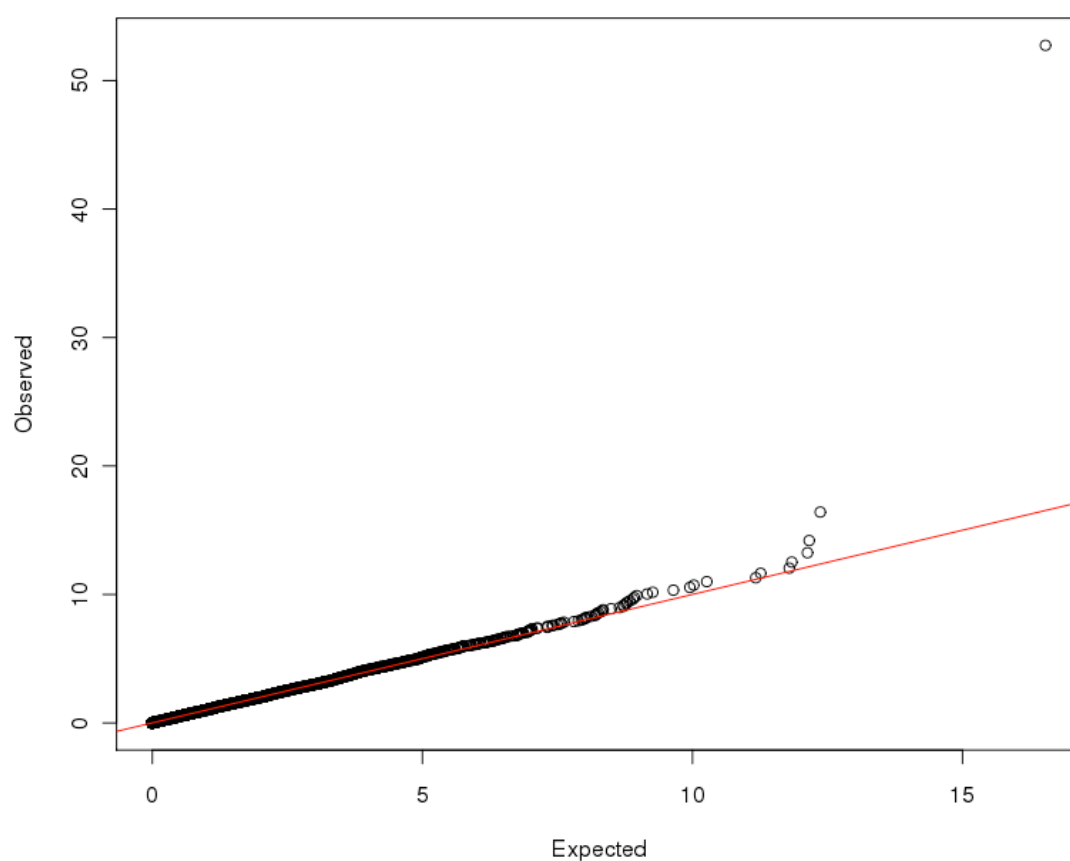


Figure 5.6. Q-Q plot for SNPs with $RSQR \geq 0.3$ in HLA-B27 negative individuals.

The red line and black circles mark the expected and the observed values respectively. The corresponding inflation factor is 0.99.

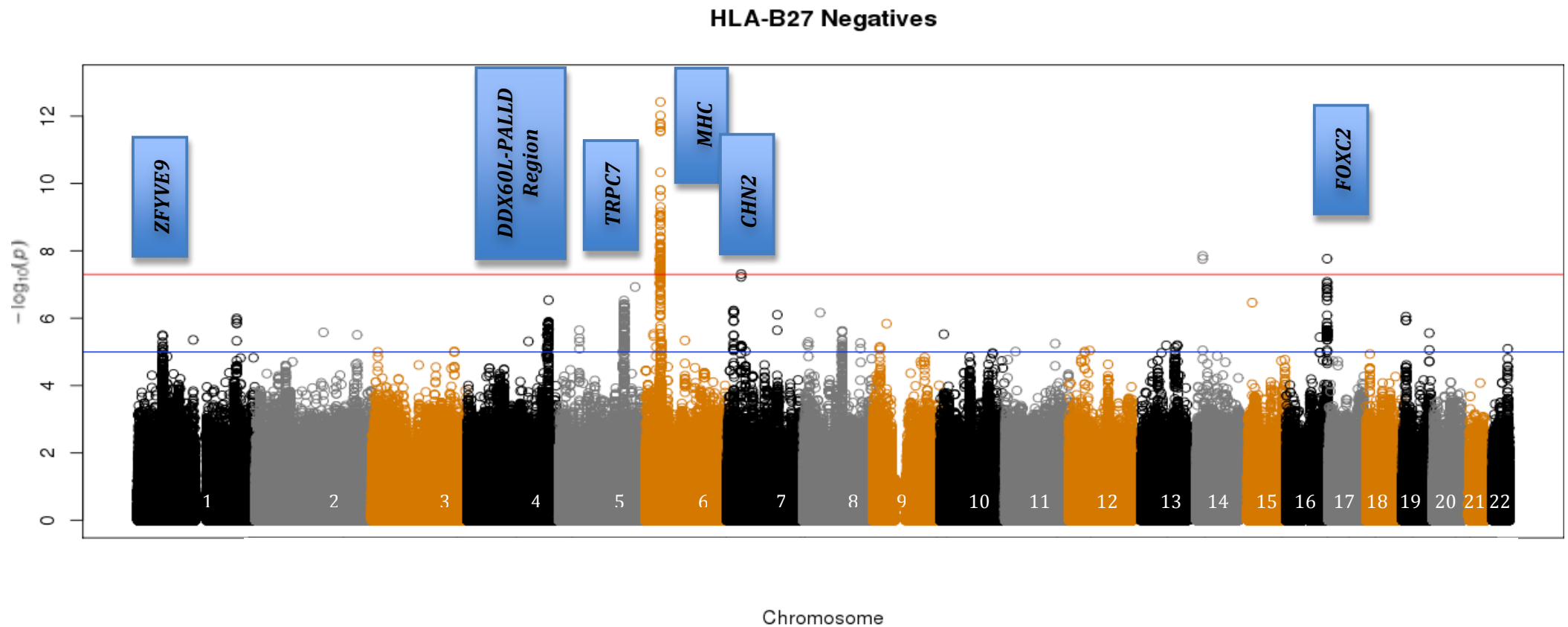
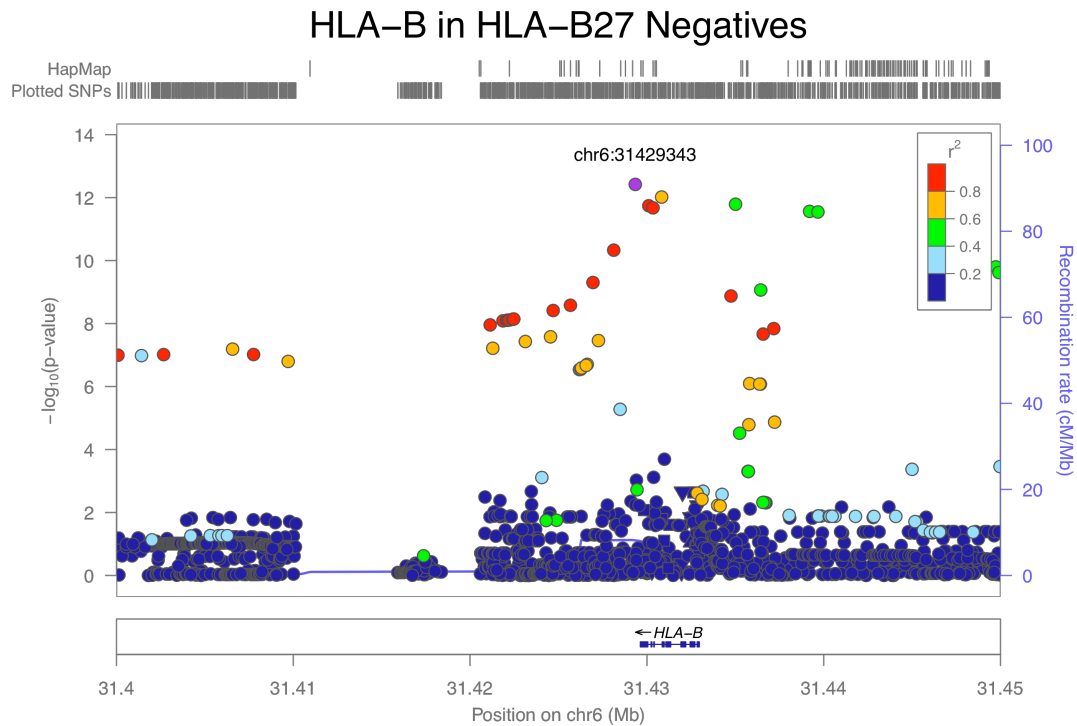


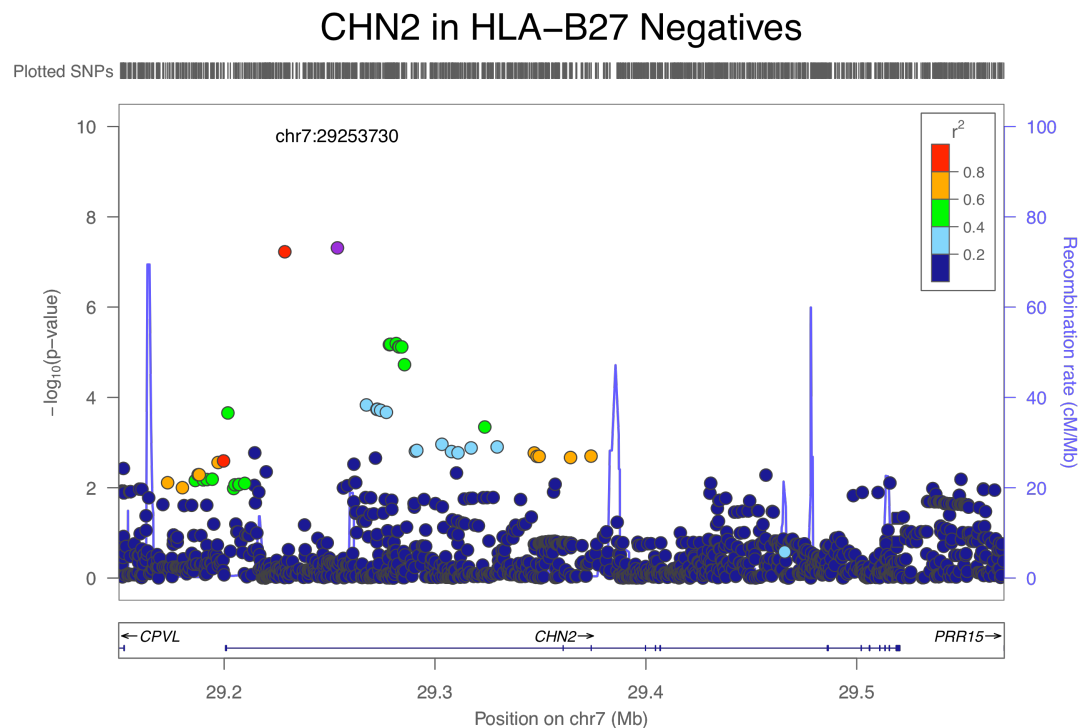
Figure 5.7. Genomewide Manhattan plot showing interesting association signals with $RSQR > 0.3$ in HLA-B27 negative individuals. The red and blue lines mark the genomewide (5×10^{-8}) and suggestive (10^{-5}) p-value thresholds, respectively.

Genes/regions of interest in HLA-B27 negatives - Imputed								
CHR	SNP	Location	Gene	Al.	Freq.	P	OR	RSQR
1p32	rs346556	52459118	<i>ZFYVE9</i>	G	0.66	3.2×10^{-6}	0.64	0.91
4q32	chr4:169512527	169512527	<i>DDX60L-PALLD</i> region	C	0.85	2.9×10^{-7}	0.57	0.98
5q15	chr5:96252997	96252997	<i>ERAP2</i>	C	0.87	0.0015	1.66	0.97
5q31	chr5:136148558	136148558	<i>TRPC7</i> region	C	0.83	3.0×10^{-7}	0.58	0.98
6p21	chr6:31429343	31429343	MHC: <i>HLA-B</i>	G	0.83	3.8×10^{-13}	0.21	0.95
7p15	chr7:29253730	29253730	<i>CHN2</i>	G	0.97	5.0×10^{-8}	0.33	0.89
16q24	rs13332319	85153895	<i>FOXC2</i>	G	0.65	1.7×10^{-8}	1.89	0.91

Table 5.10. Genes/regions of interest identified in the logistic regression analysis of the imputed genotypes in HLA-B27 negative individuals. (Al.: Allele; Freq.: Frequency; CHR: Chromosome; RSQR: Imputation quality score.)

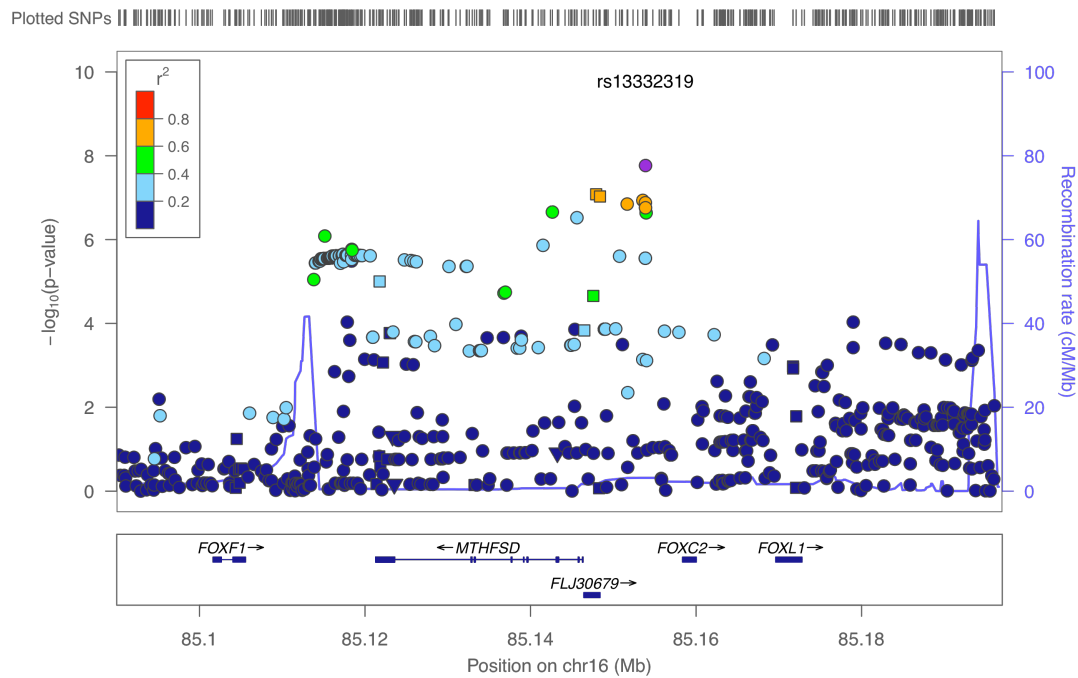


Plot 5.20. Residual association in the *HLA-B* region in HLA-B27 negative individuals (as defined by rs4349859). This association was absent in HLA-B27 positive individuals (See Appendix 2, Plot 5.24 for the regional association plot). The strongest association was observed with the SNP chr6:31429343 (Purple) (See Section 5.4.2.2. for a detailed discussion of this residual association). Conditional analysis also revealed an independent secondary signal in this region (See Section 5.3.2, Plot 5.16). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)



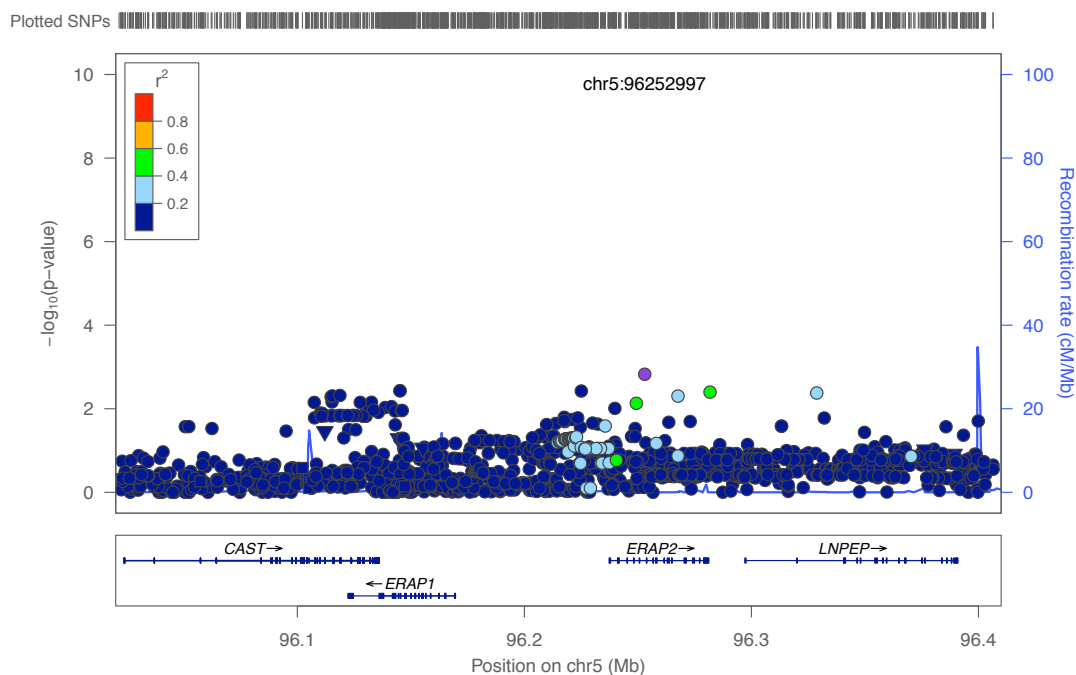
Plot 5.21. *CHN2* association in HLA-B27 negative individuals (as defined by rs4349859). The strongest association was observed with the SNP chr7:29253730 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

FOXC2 in HLA-B27 Negatives



Plot 5.22. *FOXC2* association in HLA-B27 negative individuals (as defined by rs4349859). The strongest association was observed with the SNP rs13332319 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

ERAP1 in HLA-B27 Negatives



Plot 5.23. *ERAP1* association in HLA-B27 negative individuals (as defined by rs4349859). A strong association signal was present in the *ERAP1* region in HLA-B27 positives (See Section 5.3.4.1, Plot 5.19). However, this association was absent in this analysis with HLA-B27 negatives. The strongest association was observed with the SNP chr5:96252997 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

5.3.5. Rare/Low-Frequency Variant Analysis

5.3.5.1. CCRaVAT Analysis

Rare/low frequency variant (MAF<5%) analysis with the WTCCC2 genotype data revealed two associations at the genomewide significance level ($p < 1.7 \times 10^{-6}$) in the MHC region. There were no statistically significant associations in the rest of the genome. The Manhattan plot for chromosome 6 in Figure 5.8 shows strong association within the MHC class I genes HLA-L ($p = 9.6 \times 10^{-16}$) and HLA-B ($p = 2.1 \times 10^{-6}$).

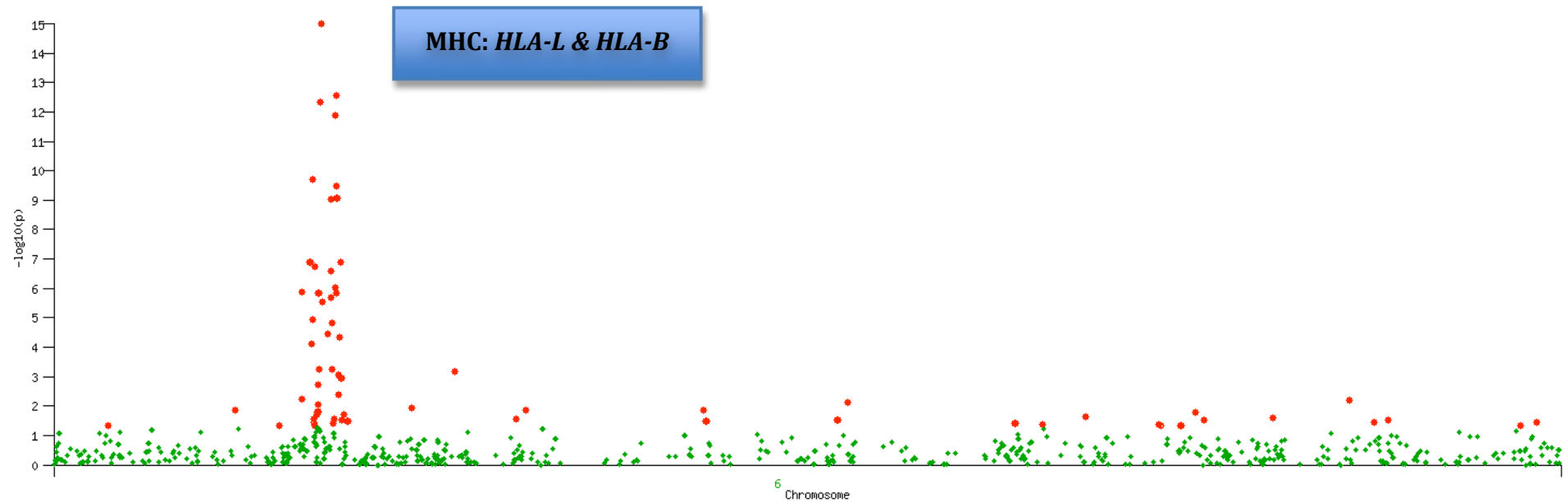


Figure 5.8. Manhattan plot for CCRAVaT analysis of the whole of chromosome 6 showing association signals within gene regions containing low frequency SNPs ($MAF < 0.05$) that were genotyped in the WTCCC2. The red dots mark the $p < 0.05$ threshold.

5.3.5.2. GRANVIL Analysis

Rare/low frequency variant (MAF<5%) analysis with the imputed genotype data revealed only one association at the genomewide significance level $p < 1.7 \times 10^{-6}$ in *HLA-B*, which is located in the MHC region. Selected results of the GRANVIL analysis of the imputed rare/low frequency (MAF<5%) variants are listed in Table 5.11 showing possible candidate genes, as well as *HLA-B* ($p = 4.4 \times 10^{-14}$). These genes were selected on the basis of their functions, all playing roles in the immune system. When conditioned on the HLA-B27 status (rs4349859 genotype), there were two residual associations detected in the MHC region (*DDR1* (6p21.3), $p = 0.00012$; *GLO1* (6p21.3), $p = 0.00019$) (See Appendix 2, Table 5.16 for a more complete list of the results).

CHR	Gene	Marker Count	Sample Count	Rare Variant Count	Effect (Beta)	P
1	<i>IL12RB2</i>	39	6949	12707	-0.85	0.0045
6	<i>HLA-B</i>	13	6949	4015	-2.59	4.4×10^{-14}
7	<i>IL6</i>	2	6949	1229	-0.27	0.0081
13	<i>TNFSF13B</i>	5	6949	656	-0.84	0.0068
16	<i>NOD3</i>	14	6949	3025	0.81	0.024
18	<i>TNFRSF11A</i>	23	6949	7444	-0.86	0.031

Table 5.11. Selected results of the GRANVIL analysis with rare/low frequency variants (MAF<5%) within gene regions. (CHR: Chromosome; Marker count: Number of rare markers in gene region; Sample count: Number of samples in analysis; Rare variant count: Count of rare alleles found in individuals; Beta: Effect size)

5.4. Discussion

5.4.1. Logistic Regression Analysis

Strong associations with the genes/regions *IL23R*, 2p15, *ERAP1*, 21q22 and *HLA-B* were detected at the genomewide significance level ($p < 5 \times 10^{-8}$). Results of the analyses were generally consistent with the results from the WTCCC2 study, although imputing to the 1000 Genomes data potentially refined the association signal by identifying variants that had slightly stronger levels of significance than in the WTCCC2 study (Tables 5.2-5.4; Figures 5.3 & 5.9; Plots 5.1-5.13). For instance, the top SNP at 21q22 (rs2836878) achieved substantially stronger association with AS (Table 5.2; $p = 3.1 \times 10^{-10}$, OR=0.75), compared to the top SNP, rs378108 reported in the WTCCC2 study (Table 5.2; $p = 1.5 \times 10^{-7}$, OR=1.23, $r^2 = 0.4$). Interestingly, rs2836878 has also been associated with CD and UC in a GWAS meta-analysis (Anderson *et al.* 2011), as well as pediatric onset inflammatory bowel disease in another association study (Allele A, OR=0.73, $p = 6.01 \times 10^{-8}$) (Kugathasan *et al.* 2008). In *KIF21B*, the top SNP rs296548 had a p-value of 4.4×10^{-5} and an odd ratio of 1.18, compared to rs2297909 in the WTCCC2 study (Table 5.2; $p = 0.0017$, OR=1.14, $r^2 = 0.6$). The top SNPs in *IL23R* (rs12033764) and the gene desert 2p15 (rs13014154) were in strong LD ($r^2 > 0.8$) with relatively similar ORs and p-values compared to the WTCCC2 top SNPs (Table 5.2). In *ERAP1* and *HLA-B*, top associated SNPs (rs30187 (a non-synonymous SNP) and rs4349859) were the same in both studies with relatively similar ORs and p-values (Table 5.2).

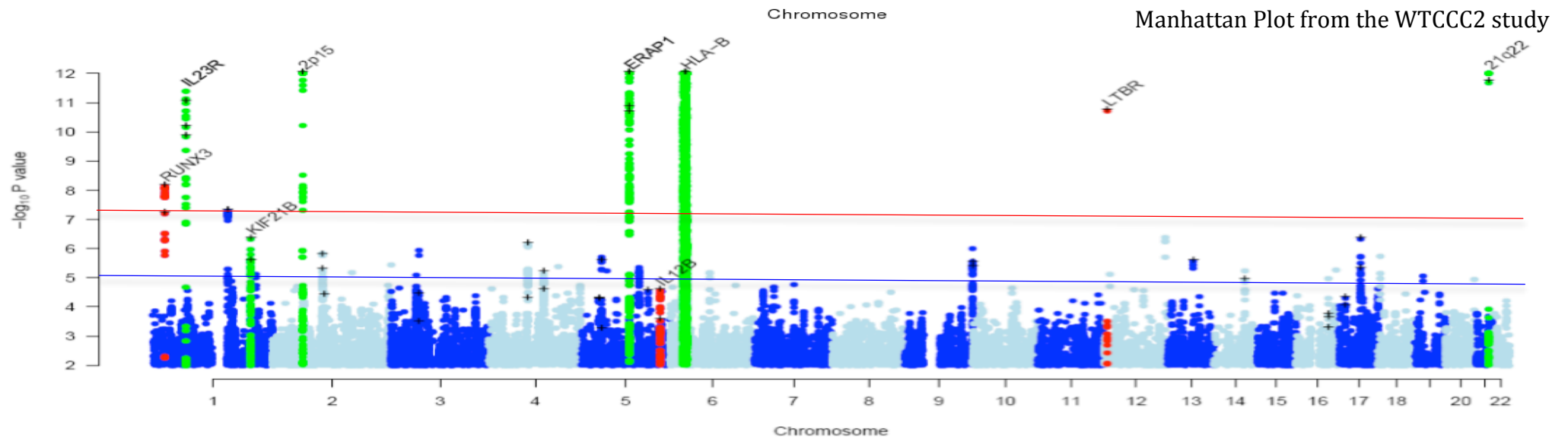
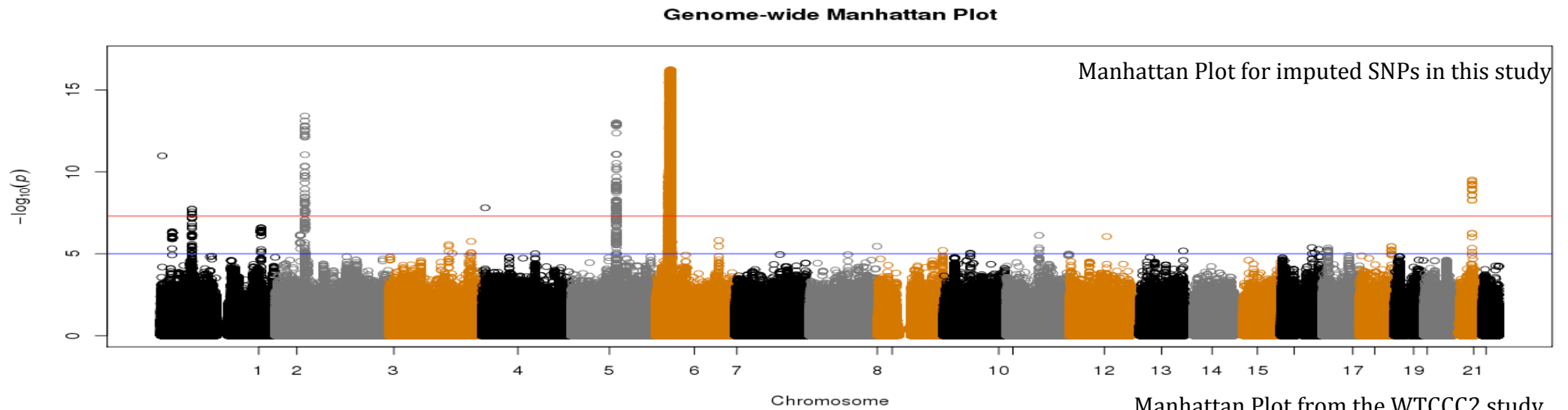


Figure 5.9. Comparison of association signals from the WTCCC2 (all individuals including those in TASC) and this study. The WTCCC2 Manhattan plot (modified from Evans *et al.*, 2011) shows the association signals observed in the discovery set (Green: previously confirmed associations; red: newly confirmed associations) (Evans *et al.* 2011). The Manhattan plot for the present study shows all association signals observed after imputing using the 1000 Genomes data.

Additional genetic associations were observed in *UBE2U*, *IL12RB2*, *IL6R*, *ESRRG*, a second gene desert on chromosome 2 (2p16.3), *ZXDC*, *IL7R*, *CDKAL1*, *TRADD* region, *MAP2K4* and *UBE2L3* (Table 5.4). These regions were selected on the basis of the strength of association ($p < 10^{-4}$), as well as biological/functional relevance to AS etiology, and previous associations with AS and related inflammatory diseases such as CD, UC and psoriasis. Relevant information on these genes is summarised in Table 5.12. SNPs rs6908425 (*CDKAL1*) and rs9033 (*TRADD*) were associated with AS in previous candidate gene studies (rs6908425, OR=0.82, $p=1.1 \times 10^{-4}$ (Danoy *et al.* 2010); rs9033, OR=1.23, $p=2.0 \times 10^{-5}$ (Pointon *et al.* 2010b)). Moreover, additional association signals in *TNFRSF1A* and upstream of *PTGER4* were detected (Table 5.4; rs11064145, $p=0.0076$; rs13156302, $p=0.0016$, respectively). *TNFRSF1A* SNP rs11064145 is in strong LD with rs1800693 ($r^2=0.754$, $D'=0.929$, HapMap release 22), which was associated with AS in TASC GWAS (See Chapter 4, Section 4.4.3) (Reveille *et al.* 2010).

CHR	Gene	Encoding	Association with other diseases/Functional importance
1q31	<i>UBE2U</i>	Ubiquitin-conjugating enzyme	Targeting peptides for degradation; potential role in antigen processing (NCBI; refseq).
1p31	<i>IL12RB2</i>	Interleukin 12 receptor subunit	Behcet's disease (Remmers <i>et al.</i> 2010), CD (Kugathasan <i>et al.</i> 2007; McGovern <i>et al.</i> 2009; D'Addabbo <i>et al.</i> 2011). The expression of this gene is up-regulated by interferon gamma in T _H 1 cells, playing a role in T _H 1 cell differentiation.
1q21	<i>IL6R</i>	Interleukin 6 receptor subunit	T _H 17 pathway (Volpe <i>et al.</i> 2008).
1q41	<i>ESRRG</i>	Estrogen receptor-related receptor	Negatively regulates bone morphogenetic protein 2-induced osteoblast differentiation and bone formation via RUNX2 (Jeong <i>et al.</i> 2009).
2p16.3	Gene desert	-	A second gene desert near 2p15. Possible presence of non-coding RNA that may be important for controlling gene expression.
3q21	<i>Downstream ZXDC</i>	Zinc finger protein	Over-expression of ZXDC in human cell lines resulted in super-activation of MHC class I and class II promoters by class II trans-activator (CIITA). Conversely, silencing of ZXDC expression reduced the ability of CIITA to activate transcription of MHC class II genes (Al-Kandari <i>et al.</i> 2007).
5p13	<i>IL7R</i>	Interleukin 7 receptor subunit	MS (De Jager <i>et al.</i> 2009), UC (Anderson <i>et al.</i> 2011). Survival of naïve and memory CD8 T cells (Schluns <i>et al.</i> 2000).
6p22	<i>CDKAL1</i>	Member of the methylthiotransferase family	AS (Danoy <i>et al.</i> 2010), CD (Franke <i>et al.</i> 2010), Psoriasis, UC (Anderson <i>et al.</i> 2011). Function unknown.
16q22	<i>TRADD region</i>	Death domain containing adaptor protein	AS. Apoptosis and NF-κB activation via TNFR1 (Pointon <i>et al.</i> 2010b).
17p12	<i>MAP2K4</i>	Protein kinase	Activates c-Jun NH2-terminal kinases (JNK) in response to pro-inflammatory cytokines. T cells deficient in MAP2K4 showed impaired IL2 production, suggesting a key role of the MAP2K4/JNK pathway in inflammation (Cuenda 2000).
22q11	<i>UBE2L3</i>	Ubiquitin-conjugating enzyme	CD, celiac disease, RA, SLE. Ubiquitination of NF-κB precursor (Fransen <i>et al.</i> 2010).

Table 5.12. Selected genes/regions of interest from the analysis of imputed data. (CHR: Chromosome; CD: Crohn's disease; MS: Multiple sclerosis; UC: Ulcerative colitis; RA: Rheumatoid arthritis; SLE: Systemic lupus erythematosus).

5.4.2. Conditional Analysis for Independent Association Signals

Logistic regression analysis conditioning on the top SNP in each region of interest with strong evidence of association revealed independent associations within the *IL12B*, *IL23R*, *ERAP1-ERAP2* and MHC regions (Table 5.5). The observation of independent association signals in *IL12B* is discussed in Chapter 6.

5.4.2.1. *IL23R* Region

Association of *IL23R* variants with AS has already been discussed in detail in Chapter 3. In addition to the findings in Chapter 3, two strong and independent associations with SNPs rs12033764 and rs11465804 (conditional $p=3.3 \times 10^{-5}$, OR=1.48) were detected as well as another relatively less strong association with rs10889658 (conditional $p=0.00067$, OR=1.16) in the *IL23R* region (Table 5.5, Plots 5.1, 5.14, 5.15). Two independent associations within *IL23R* were also detected in the WTCCC2 study (rs11209032 and rs11209026) (Evans *et al.* 2011). These SNPs are in strong LD with the SNPs detected in this study (rs11209026 – rs11465804 $r^2=0.88$, HapMap release 22; rs11209032 – rs12033764 $r^2=0.96$, 1000 Genomes pilot project). Minor allele (T) of rs12033764 and minor allele (G) of rs10889658 increase the risk of disease (OR=1.27 and OR=1.16, respectively), whereas minor allele (G) of rs11465804 decreases the risk (OR=0.68). Individuals may have varying risk of disease depending on the genotypes at these three SNPs in *IL23R*. Furthermore, the IBD-associated SNP rs10889677, which affects the expression of *IL23R* via influencing microRNA binding (Already discussed in Chapter 3, Section 3.4.5), is in LD with rs12033764 ($r^2=0.56$, $D'=0.78$, 1000 Genomes Project Pilot data) (Zwiers *et al.* 2012). This SNP (rs10889677) was also not associated with AS in

the rs12033764 conditional analysis ($p=0.99$, Plot 5.4). These findings may indicate that the haplotype tagged by rs12033764 affects the risk of disease via this possible functional SNP rs10889677. However, such a conclusion requires further functional analyses in AS cases.

5.4.2.2. *HLA-B & the MHC Region*

Residual association was identified in the major histocompatibility complex (MHC) region upon conditioning on the HLA-B27 tagging SNP rs4349859. The strongest association was with chr6:31435013 upstream of *HLA-B* (Table 5.5, $p=4.6 \times 10^{-12}$, OR=0.05; Plots 5.5 & 5.16). This residual association may due to other non-B27 MHC associations identified in AS such as MHC class I HLA-B60 and MHC class II HLA-DRB1 (Brown *et al.* 1996; Brown *et al.* 1997; Brown *et al.* 1998; Sims *et al.* 2007), although it is likely that classical HLA typing or statistical imputation with HLA*IMP at these loci would be necessary to confirm this hypothesis (Leslie *et al.* 2008; Dilthey *et al.* 2011).

HLA-B60 has also been identified as a risk allele for AS in Northern and Western Europeans in several genetic studies (Brown *et al.* 1996; Reveille 2011). The first reports of association with HLA-B60 suggested that the association was restricted to patients who were also HLA-B27 positive (Robinson *et al.* 1989). However, subsequent studies in larger samples indicate that the association is also present in HLA-B27 negative patients (Brown *et al.* 1996). In studies looking at HLA-B27 negatives and positives separately, it has been shown that HLA-B60 is independently associated with AS. Association with HLA-B60 in HLA-B27 negative cases has also been demonstrated in other ethnic groups, such as the

Taiwanese (Wei *et al.* 2004). Moreover, an association between HLA-DRB1 and AS has also been observed in European Caucasians and non-Caucasians such as Spanish (Brown *et al.* 1997; Brown *et al.* 1998; Perez-Guijo *et al.* 2002).

The HLA-B27 tagging SNP rs4349859 is 41-kb centromeric of *HLA-B*. In the WTCCC2 study, it was confirmed that rs4349859 tags all the major European-ancestry ankylosing spondylitis-associated subtypes, HLA-B*2702, HLA-B*2705 and HLA-B*2708. However, it had a sensitivity of 98% and a specificity of 99% for tagging HLA-B27 in the WTCCC2 study (Evans *et al.* 2011). Therefore, it is also possible that the observed residual association is partly due to rs4349859 not perfectly (100%) tagging HLA-B27.

5.4.2.3. *ERAP1-ERAP2* Region

Two strong and independent associations within the *ERAP1-ERAP2* region were detected, rs30187 (*ERAP1*) and rs4869313 (*ERAP2*) (conditional $p=7.3 \times 10^{-8}$, OR=0.79), as well as another relatively less strong association with rs7716677 (*ERAP1*) (conditional $p=0.00021$, OR=1.21) (Table 5.5, Plots 5.4, 5.17 & 5.18). Two independent associations were also detected in the WTCCC2 study with rs30187 and rs10050860 in *ERAP1* (Evans *et al.* 2011). However, the association with rs4869313 (*ERAP2*) was not apparent in the WTCCC2 study, possibly due to having less dense genotype data (rs4869313 is not preset in HapMap release 22). Similar effects have been detected in other studies (Spencer *et al.* 2011), where association with one SNP is enhanced when the analysis is conditioned on another SNP. Spencer *et al.* have shown that this effect typically occurs where the two alleles associated with increased disease risk occur on opposite haplotypes.

The conditioning in a multivariable analysis adjusts for the cancelling out effect of the SNP on the other haplotype (Spencer *et al.* 2011). This “haplotype effect” was observed in the *ERAP1-ERAP2* region, where the risk allele (T) of rs30187 (*ERAP1*) tends to occur more often with the protective allele (G) of rs4869313 (*ERAP2*) (Tables 5.6 & 5.7). It is also possible to observe the effect of this haplotype on the risk of disease (OR). When the effect of rs30187 (*ERAP1*) was eliminated in the conditional analysis, the effect size (OR) of rs4869313 increased (Table 5.8; Allele T, OR=1.07, conditional OR=1.26). Furthermore, the effect size (OR) of rs4869313 (*ERAP2*) in HLA-B27 positives was 1.01 (Allele T). But, upon conditioning on rs30187 (*ERAP1*) in HLA-B27 positives, this effect size (OR) increased to 1.20. Thus, rs4869313 (*ERAP2*) is associated with AS in both HLA-B27 positives and negatives, but rs30187 (*ERAP1*) masks this association due to the haplotype structure in the region.

Previously, a 3-SNP haplotype was identified in this region in a familial AS study (rs27044 – rs30187 – rs2549782; GTT, $p=0.008$, frequency=0.07; CCG, $p=0.043$, frequency=0.24) (Tsui *et al.* 2010). However, the LD between *ERAP2* SNPs rs2549782 and rs4869313 is sufficiently low ($r^2=0.6$) that there could be another haplotype in the region. Moreover, *ERAP2* SNPs rs4869313 and rs2549794 are in strong LD ($r^2=0.9$ and $D'=1$). rs2549794 was associated with CD in a genomewide meta-analysis study ($p=4.47 \times 10^{-11}$, OR=1.05) (Franke *et al.* 2010). However, CD is not associated with MHC class I genes, and therefore, observing an association with strongly linked SNPs (rs4869313 and rs2549794) both in AS and CD provides support for a potential role of *ERAP2* in AS independent of HLA-B27, as expected from the observed association in both

HLA-B27 positives and negatives. Furthermore, there is good evidence for functional synergy between ERAP1 and ERAP2 in trimming peptides for antigen presentation, making *ERAP2* a suitable functional candidate (Nguyen *et al.* 2011; Birtley *et al.* 2012). The genetic effect at this locus is likely to be larger than the single SNP analysis results, requiring further haplotype analysis, replication and confirmation with genotyped data in larger study populations.

5.4.3. HLA-B27-Stratified Analysis

Evidence of association at *ERAP1* was detected in HLA-B27 positive individuals (Figure 5.5, Table 5.6, Plot 5.19), but not in HLA-B27 negative individuals (Figure 5.7, Plot 5.23). This interaction between ERAP1 and HLA-B27 was also observed in the WTCCC2 study (Evans *et al.* 2011), and was further supported by the functional role of ERAP1 in peptide trimming for presentation by HLA-B27 in immune response by CD8+ cytotoxic T cells. ERAP1 acts as a “ruler” trimming antigenic peptides down to an optimal length of 9 amino acids, which then can be presented by HLA class I molecules (Saric *et al.* 2002; Chang *et al.* 2005). In functional studies of ERAP1 comparing rs30187 protective allele (mutant, C) with the wild-type ERAP1 allele (T), it was observed that the mutant ERAP1 has a slower rate of peptide trimming (40%) (Evans *et al.* 2011).

A residual association in the MHC region (*HLA-B*) was also observed in the analysis of HLA-B27 negative individuals (Figure 5.7, Table 5.10, Plot 5.20; chr6:31429343, $p=3.8 \times 10^{-13}$, OR=0.21), similar to the residual association observed in the rs4349859 (HLA-B27 tagging) conditional analysis (See Section

5.4.2.2; Table 5.5, Plot 5.16). This association is likely to be due to an association within the MHC independent of HLA-B27 (See Section 5.4.2.2).

Other differential genetic associations were also observed in HLA-B27 positive and negative individuals (Figures 5.5 & 5.7, Tables 5.9 & 5.10, Plots 5.21 & 5.22). Possible candidate genes selected on the basis of strength of association and biological/functional relevance are summarized in Tables 5.13 and 5.14 for HLA-B27 positive and negatives, respectively. These observations require further replication and confirmation in independent populations.

CHR	Gene	Encoding	Association with other diseases/Functional importance
1q32	<i>Upstream ZNF281</i>	Zinc finger protein	Transcription factor binding GC-rich promoters (Law <i>et al.</i> 1999).
2q32	<i>HECW2</i>	E3 ubiquitin ligase	Ubiquitination by HECW2 regulates protein stability (Miyazaki <i>et al.</i> 2003).
4p12	<i>GABRB1</i>	Gamma-aminobutyric acid A receptor, beta 1 subunit	Gamma-aminobutyric acid A receptor is a multi-subunit chloride channel that facilitates inhibitory synaptic transmission in the central nervous system (Whiting <i>et al.</i> 1999).
5q15	<i>ERAP1</i>	Endoplasmic reticulum aminopeptidase 1	AS (Evans <i>et al.</i> 2011), PS (Strange <i>et al.</i> 2010). Trimming HLA class I-binding precursors for presentation on MHC class I molecules; acts as a monomer or as a heterodimer with ERAP2 (Birtley <i>et al.</i> , 2012) (Harvey <i>et al.</i> 2009; Birtley <i>et al.</i> 2012).
8p21	<i>PEBP4</i>	Phosphatidylethanolamine (PE)-binding protein	Promotes cellular resistance to TNF-induced apoptosis (Wang <i>et al.</i> 2004).
9q34	<i>CARD9</i>	Caspase-associated recruitment domain (CARD)-containing protein; member of the CARD protein family.	AS (Pointon <i>et al.</i> 2010a), UC (Zhernakova <i>et al.</i> 2008; Rivas <i>et al.</i> 2011), CD (Franke <i>et al.</i> 2010). Selectively associated with the CARD domain of BCL10, a positive regulator of apoptosis and pro-inflammatory NF- κ B activation; acts as an activator of innate immune response against intracellular pathogens; non-synonymous SNP, rs4077515 and rs3812571, associated with AS are also associated with CARD9 expression levels (Pointon <i>et al.</i> 2010a).
15q23	<i>ITGA11</i> region	Alpha integrin	Integrins are heterodimeric cell surface receptors; ITGA11 is a collagen-binding integrin (Lu <i>et al.</i> 2010).

Table 5.13. Highlighted associations in the HLA-B27 positive sub-population. (CHR: Chromosome)

CHR	Gene	Encoding	Association with other diseases/Functional importance
1p32	<i>ZFYVE9</i>	Zinc finger protein	Acts as a mediator of TGF-beta signaling; interacts with TGF-beta receptor, cytoplasmic SMAD2 and SMAD3 proteins (Runyan <i>et al.</i> 2005).
4q32	<i>DDX60L-PALLD</i> <i>region</i>	Dead box polypeptide 60-like protein; cytoskeleton-associated protein	<i>PALLD</i> encodes a cytoskeletal protein required for organising the actin cytoskeleton; involved in the control of cell shape, adhesion, and contraction (NCBI; refseq).
5q31	<i>TRPC7</i> region	Transient receptor potential cation channel	Involved in intracellular ion (Mg^{+2} and Ca^{+2}) homeostasis and proliferation of osteoblasts (Abed & Moreau 2007).
6p21	<i>MHC</i> region	Major histocompatibility complex region	Possible <i>HLA-B60</i> and <i>HLA-DRB1</i> association in HLA-B27 negative AS cases (Brown <i>et al.</i> 1996; Brown <i>et al.</i> 1997; Brown <i>et al.</i> 1998).
7p15	<i>CHN2</i>	Beta-chimerin	GTPase activating protein; may also play a role in inhibiting/regulating T cell receptor signalling required for T cell development and function (Caloca <i>et al.</i> 2008); also provides a diacylglycerol (DAG)-dependent mechanism for regulation of adhesion and chemotaxis of T cells (Siliceo <i>et al.</i> 2006).
16q24	<i>FOXC2</i>	Forkhead box protein; member of forkhead family of transcription factors	Promotes osteoblast proliferation, survival and differentiation via up-regulation of integrin- $\beta 1$ in response to stimuli inducing bone formation (Park <i>et al.</i> 2011).

Table 5.14. Highlighted associations in the HLA-B27 negative sub-population. (CHR: Chromosome)

5.4.4. Rare/Low-frequency Variant Analysis

Two different methods for analysing rare/low-frequency (MAF<5%) variant associations were employed in this study. CCRaVAT and GRANVIL both use a similar strategy for detecting associations with rare/low frequency variants, in which rare variants within gene regions (i.e. accumulation of minor alleles within a gene region) are assumed to increase the risk of disease (Lawrence *et al.* 2010; Morris & Zeggini 2010; Magi *et al.* 2011). This “collapsing strategy” increases statistical power to detect associations with rare/low frequency variants by combining their effects across a gene region. CCRaVAT was used in the analysis of genotype data from the WTCCC2 study (Evans *et al.* 2011). A strong association with *HLA-B* was identified in this analysis ($p=2.1 \times 10^{-6}$). The strongest association was with an MHC class I pseudogene *HLA-L* ($p=9.6 \times 10^{-16}$) (Figure 5.8). However, because of the strong LD across the MHC region on chromosome 6, it is difficult to identify the actual causal gene(s) within this region other than *HLA-B*. The *HLA-L* association may be due to the strong *HLA-B* association in the MHC region, across which strong LD exists.

The GRANVIL analysis of the imputed SNPs revealed a number of interesting associations. Imputation of SNPs using the 1000 Genomes data provided much denser genotype data for the GRANVIL analysis of low frequency variants. Genes selected on the basis of strength of association and biological/functional relevance are listed in Table 5.11. In addition to the strong *HLA-B* association at genomewide significance ($p=4.4 \times 10^{-14}$), 5 other immune-related genes (*IL12RB2*, *IL6*, *TNFSF13B*, *NOD3* and *TNFRSF11A*) were also noteworthy. Relevant information on these genes is summarised in Table 5.15. Chromosome

6 analyses conditioning on the HLA-B27 tagging SNP rs4349859 detected residual associations with genes linked to the MHC region (AS-associated *HLA-B* is located at 6p21.3) (*DDR1* (6p21.3), $p=0.00012$; *GLO1* (6p21.3), $p=0.00019$). However, this result may reflect associations with non-HLA-B27 loci within the MHC, which are not tagged by rs4349859, such as HLA-B60 (6p21.3) and *HLA-DRB1* (6p21.3) (Brown *et al.* 1996; Brown *et al.* 1998).

It has been shown that imputing with the most recent 1000 Genomes Project data, which provide haplotypes from larger multiple ancestry groups (Phase I, June 2011 interim release, 1094 individuals), and the genotype data available from the previous GWAS (e.g. WTCCC1, 2007) is a cost-effective and powerful way of analysing associations with rare/low frequency variants compared to expensive re-sequencing of large study populations (Magi *et al.* 2012). Similar to the analysis in this chapter, GRANVIL was used for the analysis of rare/low frequency variants in this study. However, it is necessary to replicate the associations observed in this chapter and other similar studies in independent populations, and employ other methods for the analysis of rare/low frequency variants as new and better methods are still being developed. For instance, CCRaVAT and GRANVIL assume that all rare/low frequency variants increase disease risk. There are other methods available for detecting associations with rare/low frequency variants such as C-alpha that do not make such an assumption, allowing for low frequency risk alleles as well as low frequency protective alleles (Neale *et al.* 2011).

CHR	Gene	Encoding	Association with other diseases/Functional importance
1p31	<i>IL12RB2</i>	Interleukin 12 receptor subunit	Behcet's disease (Remmers <i>et al.</i> 2010), CD (Kugathasan <i>et al.</i> 2007; McGovern <i>et al.</i> 2009; D'Addabbo <i>et al.</i> 2011). The expression of this gene is up-regulated by interferon gamma in T _H 1 cells, playing a role in T _H 1 cell differentiation.
6p21	<i>HLA-B</i>	HLA class I molecules	AS (Evans <i>et al.</i> 2011), MS (De Jager <i>et al.</i> 2009). Antigen presentation; HLA-B27 is strongly associated with AS. HLA-B60 may also play a role in AS susceptibility (Robinson <i>et al.</i> 1989; Brown <i>et al.</i> 1996; Brown <i>et al.</i> 1997).
7p21	<i>IL6</i>	Cytokine; interleukin 6	Intervertebral disc degeneration in Finnish (Kelempisioti <i>et al.</i> 2011). Acts via IL6 receptor to stimulate pro-inflammatory responses in T _H 17 pathway (Volpe <i>et al.</i> 2008).
13q32	<i>TNFSF13B</i>	Cytokine; member of tumor necrosis factor ligand family	Playing a role in survival of B cells and may also activate macrophages leading to activation of NF- κ B (Lee <i>et al.</i> 2011).
16p13	<i>NOD3</i>	Nucleotide-binding oligomerisation domain (NOD) containing protein	NOD proteins function as pattern recognition receptors (PRRs) stimulating response against bacteria, as well as regulating apoptosis via caspases (Inohara & Nunez 2003). NOD2 encoding a similar protein was associated with CD (Franke <i>et al.</i> 2010).
18q22	<i>TNFRSF11A</i>	Member of TNF receptor superfamily; also known as <i>RANK</i>	Paget's disease of bone (Albagha <i>et al.</i> 2010). Activates NF- κ B and MAPK8/JNK; important regulator of the interaction between T cells and dendritic cells; studies in mice suggest that this receptor directly mediates the osteoprotegerin ligand (OPGL)-induced osteoclastogenesis in osteoclast precursor cells (NCBI;Refseq).

Table 5.15. Highlighted associations from the GRANVIL analysis with rare/low frequency (MAF<5%) variants within gene regions. (CHR: Chromosome; CD: Crohn's disease; MS: Multiple sclerosis)

5.4.5. Technical & Statistical Considerations

The 1000 Genomes pilot project sequenced only 122 chromosomes from the CEU population (individuals of Northern and Western European ancestry in Utah) (Genomes 2010). Therefore, it is important to note that the expected imputation accuracy when using the haplotypes from the 1000 Genomes Project pilot data was lower than that expected when using the most recent 1000 Genomes haplotypes from a larger reference population. Furthermore, imputation accuracy was also expected to be lower for rare/low frequency variants, requiring use of higher accuracy measures (imputation accuracy $RSQR > 0.5$ for variants with $MAF < 5\%$). This is discussed further in Chapter 6.

It is probable that associations with private, rare and some low frequency risk variants specific to cases were missed, as these variants are not likely to be present in the reference panel. In a study comparing the HapMap data with the 1000 Genomes Project pilot data, it was estimated that 72% of the HapMap (merged phases 2 and 3) SNPs are present in the 1000 Genomes Project pilot data. This number increased to 99% when HapMap SNPs with $MAF < 5\%$ were excluded from the comparison (Buchanan et al., 2012). Even though almost all of the common SNPs ($MAF > 5\%$) are found in both HapMap and 1000 Genomes Project pilot data, there is a significant number of rare/low frequency variants that are not present in both datasets. It is important to note such limitations of the 1000 Genomes Project pilot reference panel, and to be cautious when drawing conclusions and making comparisons between different studies.

A study by Huang *et al.* used a more recent reference panel from the 1000 Genomes Project (EUR 20100804, 566 haplotypes) to impute SNPs in the WTCCC1

population (WTCCC1 2007). The study successfully identified two missed associations in type 1 diabetes and type 2 diabetes, as well as refining the *IL23R* association in CD. They concluded that imputation using the 1000 Genomes Project data can identify novel associations, and refine previously observed associations to facilitate determination of plausible explanations for the detected associations (Huang *et al.* 2012).

In addition to identifying refined and similar associations to the WTCCC2 study, novel associations were also observed in this chapter. A noteworthy example is the association with *ERAP2* (See Section 5.4.2.3). This association was probably missed in the WTCCC2 analysis due to having a less dense genotype data. However, this *ERAP2* association and the other novel associations detected in this study still require replication and confirmation in independent studies with high statistical power.

5.5. References

- Abed, E. & Moreau, R. (2007). "Importance of melastatin-like transient receptor potential 7 and cations (magnesium, calcium) in human osteoblast-like cell proliferation." *Cell Prolif* **40**(6): 849-865.
- Al-Kandari, W., Jambunathan, S., *et al.* (2007). "ZXDC, a novel zinc finger protein that binds CIITA and activates MHC gene transcription." *Mol Immunol* **44**(4): 311-321.
- Albagha, O. M., Visconti, M. R., *et al.* (2010). "Genome-wide association study identifies variants at CSF1, OPTN and TNFRSF11A as genetic risk factors for Paget's disease of bone." *Nat Genet* **42**(6): 520-524.
- Anderson, C. A., Boucher, G., *et al.* (2011). "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47." *Nat Genet* **43**(3): 246-252.
- Birtley, J. R., Saridakis, E., *et al.* (2012). "The crystal structure of human endoplasmic reticulum aminopeptidase 2 reveals the atomic basis for distinct roles in antigen processing." *Biochemistry* **51**(1): 286-295.
- Bodmer, W. & Bonilla, C. (2008). "Common and rare variants in multifactorial susceptibility to common diseases." *Nat Genet* **40**(6): 695-701.
- Brown, M. A., Kennedy, L. G., *et al.* (1998). "The effect of HLA-DR genes on susceptibility to and severity of ankylosing spondylitis." *Arthritis Rheum* **41**(3): 460-465.
- Brown, M. A., Kennedy, L. G., *et al.* (1997). "Susceptibility to ankylosing spondylitis in twins: the role of genes, HLA, and the environment." *Arthritis Rheum* **40**(10): 1823-1828.
- Brown, M. A., Pile, K. D., *et al.* (1996). "HLA class I associations of ankylosing spondylitis in the white population in the United Kingdom." *Ann Rheum Dis* **55**(4): 268-270.
- Burton, P. R., Clayton, D. G., *et al.* (2007). "Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants." *Nat Genet* **39**(11): 1329-1337.
- Caloca, M. J., Delgado, P., *et al.* (2008). "Role of chimaerins, a group of Rac-specific GTPase activating proteins, in T-cell receptor signaling." *Cell Signal* **20**(4): 758-770.
- Chang, S. C., Momburg, F., *et al.* (2005). "The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism." *Proc Natl Acad Sci U S A* **102**(47): 17107-17112.
- Cuenda, A. (2000). "Mitogen-activated protein kinase kinase 4 (MKK4)." *Int J Biochem Cell Biol* **32**(6): 581-587.
- D'Addabbo, A., Palmieri, O., *et al.* (2011). "Discovering genetic variants in Crohn's disease by exploring genomic regions enriched of weak association signals." *Dig Liver Dis* **43**(8): 623-631.
- Danoy, P., Pryce, K., *et al.* (2010). "Association of variants at 1q32 and STAT3 with ankylosing spondylitis suggests genetic overlap with Crohn's disease." *PLoS Genet* **6**(12): e1001195.
- De Jager, P. L., Jia, X., *et al.* (2009). "Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci." *Nat Genet* **41**(7): 776-782.

- Dilthey, A. T., Moutsianas, L., *et al.* (2011). "HLA*IMP--an integrated framework for imputing classical HLA alleles from SNP genotypes." *Bioinformatics* **27**(7): 968-972.
- Evans, D. M., Spencer, C. C., *et al.* (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**(8): 761-767.
- Franke, A., McGovern, D. P., *et al.* (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." *Nat Genet* **42**(12): 1118-1125.
- Fransen, K., Visschedijk, M. C., *et al.* (2010). "Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease." *Hum Mol Genet* **19**(17): 3482-3488.
- Frazer, K. A., Ballinger, D. G., *et al.* (2007). "A second generation human haplotype map of over 3.1 million SNPs." *Nature* **449**(7164): 851-861.
- Genomes (2010). "A map of human genome variation from population-scale sequencing." *Nature* **467**(7319): 1061-1073.
- HapMap (2003). "The International HapMap Project." *Nature* **426**(6968): 789-796.
- HapMap (2005). "A haplotype map of the human genome." *Nature* **437**(7063): 1299-1320.
- Harvey, D., Pointon, J. J., *et al.* (2009). "Investigating the genetic association between ERAP1 and ankylosing spondylitis." *Hum Mol Genet* **18**(21): 4204-4212.
- Huang, J., *et al.* (2012). "1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data." *Eur J Hum Genet*.
- Inohara, N. & Nunez, G. (2003). "NODs: intracellular proteins involved in inflammation and apoptosis." *Nat Rev Immunol* **3**(5): 371-382.
- Jeong, B. C., Lee, Y. S., *et al.* (2009). "The orphan nuclear receptor estrogen receptor-related receptor gamma negatively regulates BMP2-induced osteoblast differentiation and bone formation." *J Biol Chem* **284**(21): 14211-14218.
- Kelempisioti, A., Eskola, P. J., *et al.* (2011). "Genetic susceptibility of intervertebral disc degeneration among young Finnish adults." *BMC Med Genet* **12**: 153.
- Kugathasan, S., Baldassano, R. N., *et al.* (2008). "Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease." *Nat Genet* **40**(10): 1211-1215.
- Kugathasan, S., Saubermann, L. J., *et al.* (2007). "Mucosal T-cell immunoregulation varies in early and late inflammatory bowel disease." *Gut* **56**(12): 1696-1705.
- Law, D. J., Du, M., *et al.* (1999). "ZBP-99 defines a conserved family of transcription factors and regulates ornithine decarboxylase gene expression." *Biochem Biophys Res Commun* **262**(1): 113-120.
- Lawrence, R., Day-Williams, A. G., *et al.* (2010). "CCRaVAT and QuTie-enabling analysis of rare variants in large-scale case control and quantitative trait association studies." *BMC Bioinformatics* **11**: 527.
- Lee, S. M., Kim, E. J., *et al.* (2011). "BAFF and APRIL induce inflammatory activation of THP-1 cells through interaction with their conventional receptors and activation of MAPK and NF-kappaB." *Inflamm Res* **60**(9): 807-815.
- Leslie, S., Donnelly, P., *et al.* (2008). "A statistical method for predicting classical HLA alleles from SNP data." *Am J Hum Genet* **82**(1): 48-56.

- Li, Y., Willer, C., *et al.* (2009). "Genotype imputation." Annu Rev Genomics Hum Genet **10**: 387-406.
- Li, Y., Willer, C. J., *et al.* (2010). "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." Genet Epidemiol **34**(8): 816-834.
- Lu, N., Carracedo, S., *et al.* (2010). "The human alpha11 integrin promoter drives fibroblast-restricted expression in vivo and is regulated by TGF-beta1 in a Smad- and Sp1-dependent manner." Matrix Biol **29**(3): 166-176.
- Magi, R., Asimit, J. L., *et al.* (2012). "Genome-Wide Association Analysis of Imputed Rare Variants: Application to Seven Common Complex Diseases." Genet Epidemiol.
- Magi, R., Kumar, A., *et al.* (2011). "Assessing the impact of missing genotype data in rare variant association analysis." BMC Proc **5 Suppl 9**: S107.
- Maher, B. (2008). "Personal genomes: The case of the missing heritability." Nature **456**(7218): 18-21.
- Manolio, T. A., Collins, F. S., *et al.* (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.
- McGovern, D. P., Rotter, J. I., *et al.* (2009). "Genetic epistasis of IL23/IL17 pathway genes in Crohn's disease." Inflamm Bowel Dis **15**(6): 883-889.
- Miyazaki, K., Ozaki, T., *et al.* (2003). "A novel HECT-type E3 ubiquitin ligase, NEDL2, stabilizes p73 and enhances its transcriptional activity." Biochem Biophys Res Commun **308**(1): 106-113.
- Morris, A. P. & Zeggini, E. (2010). "An evaluation of statistical approaches to rare variant analysis in genetic association studies." Genet Epidemiol **34**(2): 188-193.
- Neale, B. M., Rivas, M. A., *et al.* (2011). "Testing for an unusual distribution of rare variants." PLoS Genet **7**(3): e1001322.
- Nguyen, T. T., Chang, S. C., *et al.* (2011). "Structural basis for antigenic peptide precursor processing by the endoplasmic reticulum aminopeptidase ERAP1." Nat Struct Mol Biol **18**(5): 604-613.
- Park, S. J., Gadi, J., *et al.* (2011). "The forkhead transcription factor Foxc2 promotes osteoblastogenesis via up-regulation of integrin beta1 expression." Bone **49**(3): 428-438.
- Perez-Guijo, V., Munoz, E., *et al.* (2002). "Distribution of HLA-DRB1 genes in patients with sporadic ankylosing spondylitis in the south of Spain." Joint Bone Spine **69**(5): 458-462.
- Pointon, J. J., Harvey, D., *et al.* (2010a). "Elucidating the chromosome 9 association with AS; CARD9 is a candidate gene." Genes Immun **11**(6): 490-496.
- Pointon, J. J., Harvey, D., *et al.* (2010b). "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**(6): 1243-1246.
- Pruim, R. J., Welch, R. P., *et al.* (2010). "LocusZoom: regional visualization of genome-wide association scan results." Bioinformatics **26**(18): 2336-2337.
- Purcell, S., Neale, B., *et al.* (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.
- Remmers, E. F., Cosan, F., *et al.* (2010). "Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behcet's disease." Nat Genet **42**(8): 698-702.

- Reveille, J. D. (2011). "The genetic basis of spondyloarthritis." Ann Rheum Dis **70 Suppl 1**: i44-50.
- Reveille, J. D., Sims, A. M., *et al.* (2010). "Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci." Nat Genet **42**(2): 123-127.
- Rivas, M. A., Beaudoin, M., *et al.* (2011). "Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease." Nat Genet **43**(11): 1066-1073.
- Robinson, W. P., van der Linden, S. M., *et al.* (1989). "HLA-Bw60 increases susceptibility to ankylosing spondylitis in HLA-B27+ patients." Arthritis Rheum **32**(9): 1135-1141.
- Runyan, C. E., Schnaper, H. W., *et al.* (2005). "The role of internalization in transforming growth factor beta1-induced Smad2 association with Smad anchor for receptor activation (SARA) and Smad2-dependent signaling in human mesangial cells." J Biol Chem **280**(9): 8300-8308.
- Saric, T., Chang, S. C., *et al.* (2002). "An IFN-gamma-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides." Nat Immunol **3**(12): 1169-1176.
- Schluns, K. S., Kieper, W. C., *et al.* (2000). "Interleukin-7 mediates the homeostasis of naive and memory CD8 T cells in vivo." Nat Immunol **1**(5): 426-432.
- Siliceo, M., Garcia-Bernal, D., *et al.* (2006). "Beta2-chimaerin provides a diacylglycerol-dependent mechanism for regulation of adhesion and chemotaxis of T cells." J Cell Sci **119**(Pt 1): 141-152.
- Sims, A. M., Barnardo, M., *et al.* (2007). "Non-B27 MHC associations of ankylosing spondylitis." Genes Immun **8**(2): 115-123.
- Spencer, C. C., Plagnol, V., *et al.* (2011). "Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21." Hum Mol Genet **20**(2): 345-353.
- Strange, A., Capon, F., *et al.* (2010). "A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1." Nat Genet **42**(11): 985-990.
- Tsui, F. W., Haroon, N., *et al.* (2010). "Association of an ERAP1 ERAP2 haplotype with familial ankylosing spondylitis." Ann Rheum Dis **69**(4): 733-736.
- van der Linden, S., Valkenburg, H. A., *et al.* (1984). "Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria." Arthritis Rheum **27**(4): 361-368.
- Volpe, E., Servant, N., *et al.* (2008). "A critical function for transforming growth factor-beta, interleukin 23 and proinflammatory cytokines in driving and modulating human T(H)-17 responses." Nat Immunol **9**(6): 650-657.
- Wang, X., Li, N., *et al.* (2004). "A novel human phosphatidylethanolamine-binding protein resists tumor necrosis factor alpha-induced apoptosis by inhibiting mitogen-activated protein kinase pathway activation and phosphatidylethanolamine externalization." J Biol Chem **279**(44): 45855-45864.
- Wei, J. C., Tsai, W. C., *et al.* (2004). "HLA-B60 and B61 are strongly associated with ankylosing spondylitis in HLA-B27-negative Taiwan Chinese patients." Rheumatology (Oxford) **43**(7): 839-842.
- Whiting, P. J., Bonnert, T. P., *et al.* (1999). "Molecular and functional diversity of the expanding GABA-A receptor gene family." Ann N Y Acad Sci **868**: 645-653.

- WTCCC1 (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* **447**(7145): 661-678.
- Zhang, B., Zhi, D., *et al.* (2011). "Practical Consideration of Genotype Imputation: Sample Size, Window Size, Reference Choice, and Untyped Rate." *Stat Interface* **4**(3): 339-352.
- Zhernakova, A., Festen, E. M., *et al.* (2008). "Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP." *Am J Hum Genet* **82**(5): 1202-1210.
- Zwiers, A., Kraal, L., *et al.* (2012). "Cutting edge: a variant of the IL-23R gene associated with inflammatory bowel disease induces loss of microRNA regulation and enhanced protein production." *J Immunol* **188**(4): 1573-1577.

CHAPTER 6

6. CONFIRMATION OF IMPUTATION RESULTS BY DIRECT GENOTYPING

6.1. Introduction

There are now numerous publicly available software packages for the imputation of untyped genetic polymorphisms and the statistical analysis of these data (Li *et al.* 2010; Marchini & Howie 2010). The accuracy of imputation for each of these methodologies has typically been evaluated by comparing imputed genotypes against a wet lab gold standard. I used a similar approach to check the validity of the imputed results from Chapter 5 by directly genotyping the most promising SNPs associated with risk of AS, as defined by a combination of putative functional relevance, strength of statistical association, and prior evidence from previously published studies (See Table 6.1 and Section 6.2 below for a more detailed explanation of the SNP selection procedure). Specifically, I identified 40 novel regions of interest from the logistic regression analyses of 1782 cases and 5167 controls (See Chapter 5). After considering the selection criteria, SNPs from 10 distinct genomic regions were chosen for direct genotyping to confirm the imputation and association results from the previous chapter. These SNPs were subsequently genotyped in a set of 347 cases common to this study and the WTCCC2 study to determine the imputation accuracy. The same set of 10 SNPs were then genotyped in 1804 cases and 1848 controls that were independent from

those individuals in the WTCCC2 (Chapter 5) in an attempt to replicate the putative associations at these loci.

Recently, a genetic interaction between variants within the *ERAP1* gene and HLA-B27 was identified in modifying susceptibility to AS, (Evans *et al.* 2011). The authors concluded that the interaction was likely to reflect ERAP1's serial trimming of potential peptide antigens for binding to HLA-B27 in preparation for presentation for immune recognition by cytotoxic T cells. In order to facilitate the search for statistical interactions between HLA-B27 and variants at other loci, we genotyped the HLA-B27 tag SNP, rs4349859 to determine HLA-B27 status of all cases and controls in the replication study in addition to the 10 putatively associated variants under study. The study population was stratified by HLA-B27 status, and the selected 10 loci were tested for association in HLA-B27 positives and negatives separately.

6.2. Patients, Materials & Methods

6.2.1. Selection of Genes/Regions of Interest

I used four criteria for identifying regions/genes for direct genotyping. All of the selected genes/regions met the first three of these criteria:

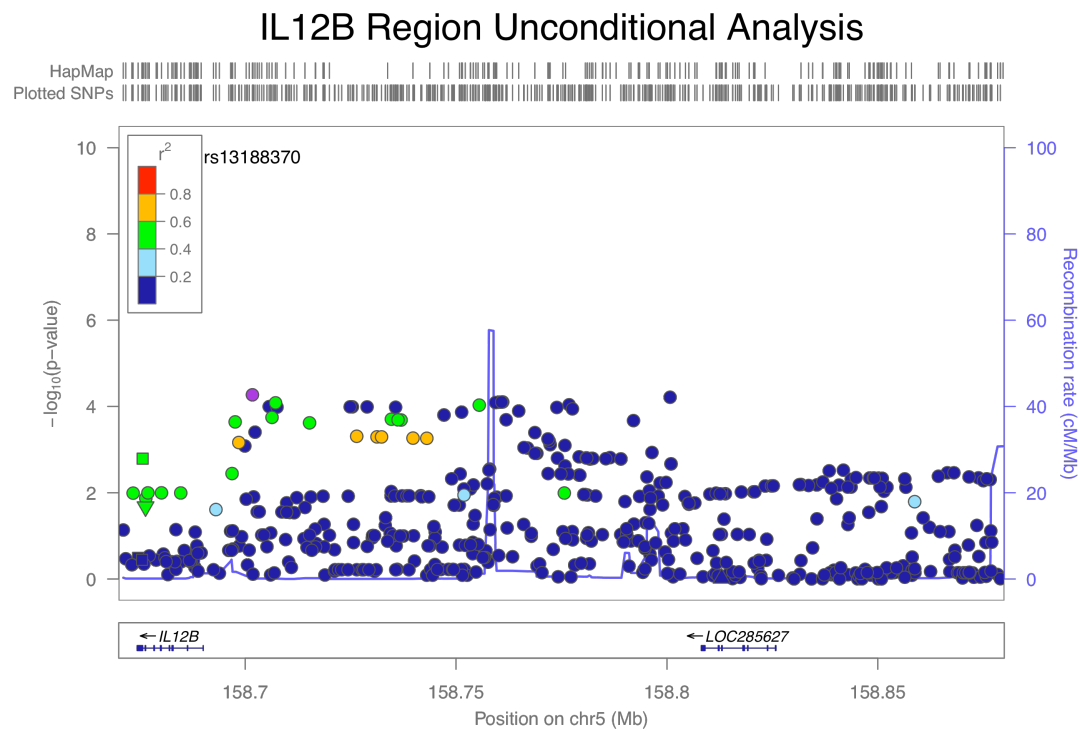
6. They were among the top associated regions from Chapter 5 (p-values between 10^{-4} and 10^{-5}) that were not followed up in the published WTCCC2-TASC study (Evans et al. 2011),
7. The presence of multiple associated SNPs in the region with p-values less than 10^{-4} ,
8. Plausible functional relevance of genes in the region to the aetiology of AS (i.e. biological plausibility in the pathogenesis of AS and/or association with other inflammatory diseases (eg. Crohn's disease (CD), ulcerative colitis (UC), psoriasis (PS), rheumatoid arthritis (RA) and multiple sclerosis (MS)),
9. Appropriate tissue expression of genes of interest in the associated region (i.e. bone, joint, connective tissue, eye, skin, lymphoid tissues).

6.2.2. Selection of SNPs

For each region selected for follow-up, the imputed SNP with the lowest p value was chosen for direct confirmatory genotyping and replication. None of the SNPs had been genotyped in either the WTCCC2 discovery or replication populations (Evans *et al.* 2011). Conditional analyses were performed using the mach2dat software conditioning on the top variant in the region to check for the existence of secondary signals independent from the lead SNP (Li *et al.* 2009). This process and the pattern of association in each region is illustrated by regional association

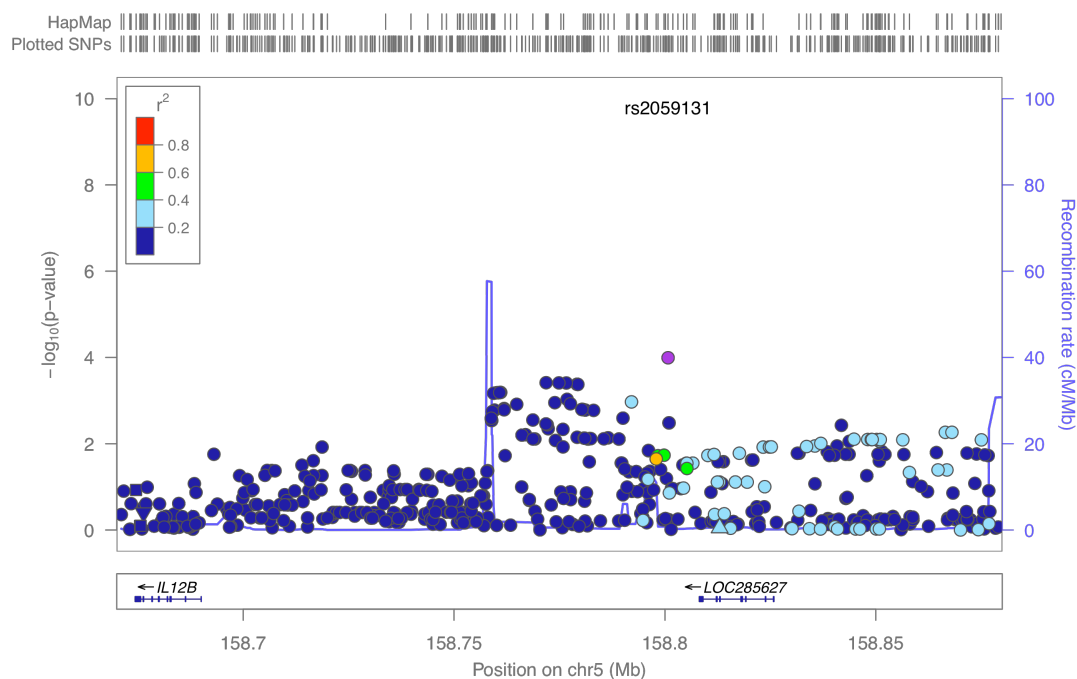
plots made using the Locuszoom software package (Pruim *et al.* 2010). The Locuszoom software identifies the most strongly associated SNP in a region (as assessed by p-value) and then displays the linkage disequilibrium between this marker and every other SNP in the region via a colour-coding scheme. The plot therefore makes it easy to see the existence of possible secondary signals.

In the *IL12B* region on chromosome 5, there were secondary and tertiary independent associations with risk of AS after conditional analyses (rs13188370, rs2059131 and rs2013718) (Plots 6.1-6.3). All three SNPs were therefore selected for confirmatory genotyping and replication. None of the other regions examined showed strong evidence for the existence of secondary signals ($p < 5 \times 10^{-3}$) (See Appendix 3 for all regional association plots and detailed information on selected genes/regions).

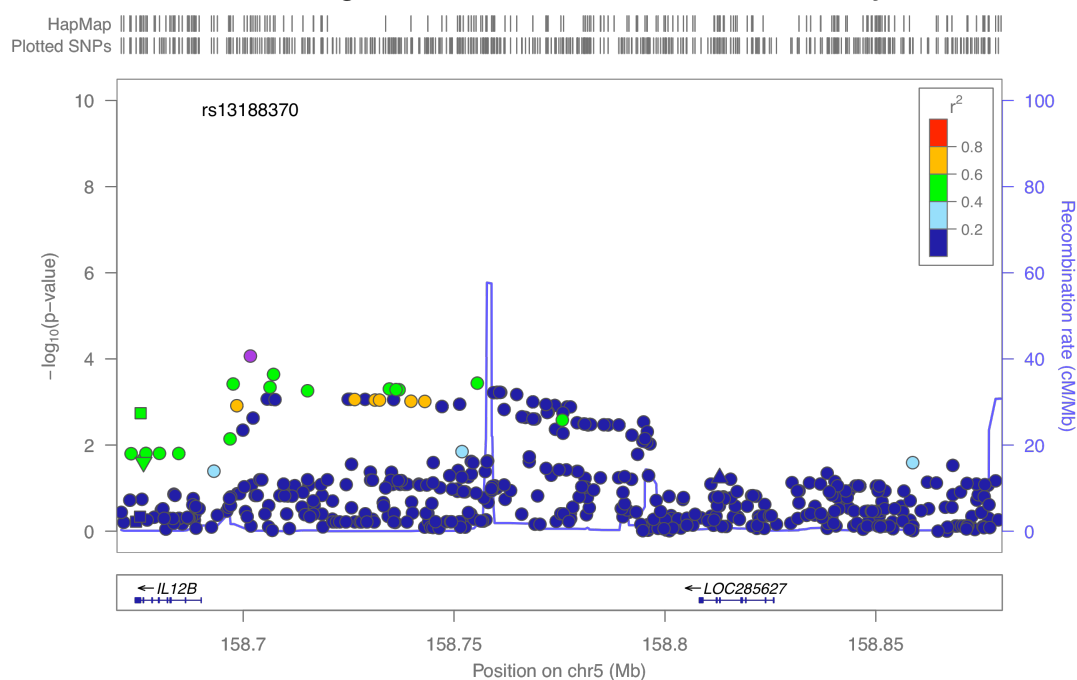


Plot 6.1. *IL12B* region unconditional analysis – showing the top hit SNP (rs13188370, purple) in the analysis of imputed genotypes, and the LD pattern relative to the top SNP ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant).

(A) IL12B Region rs13188370 Conditional Analysis

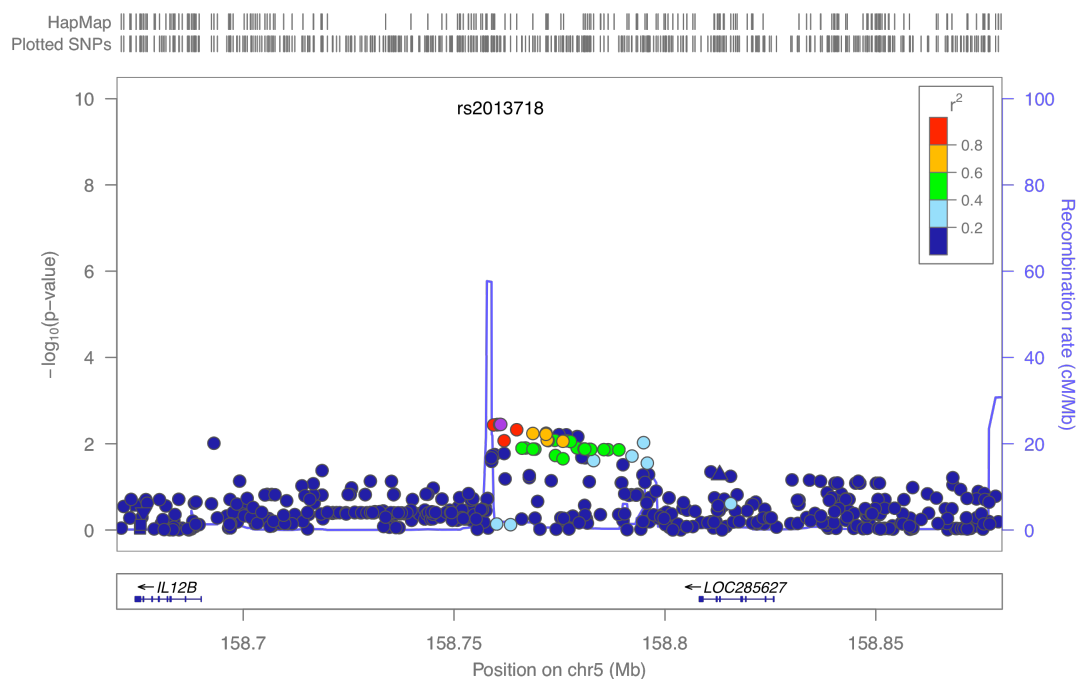


(B) IL12B Region rs2059131 Conditional Analysis

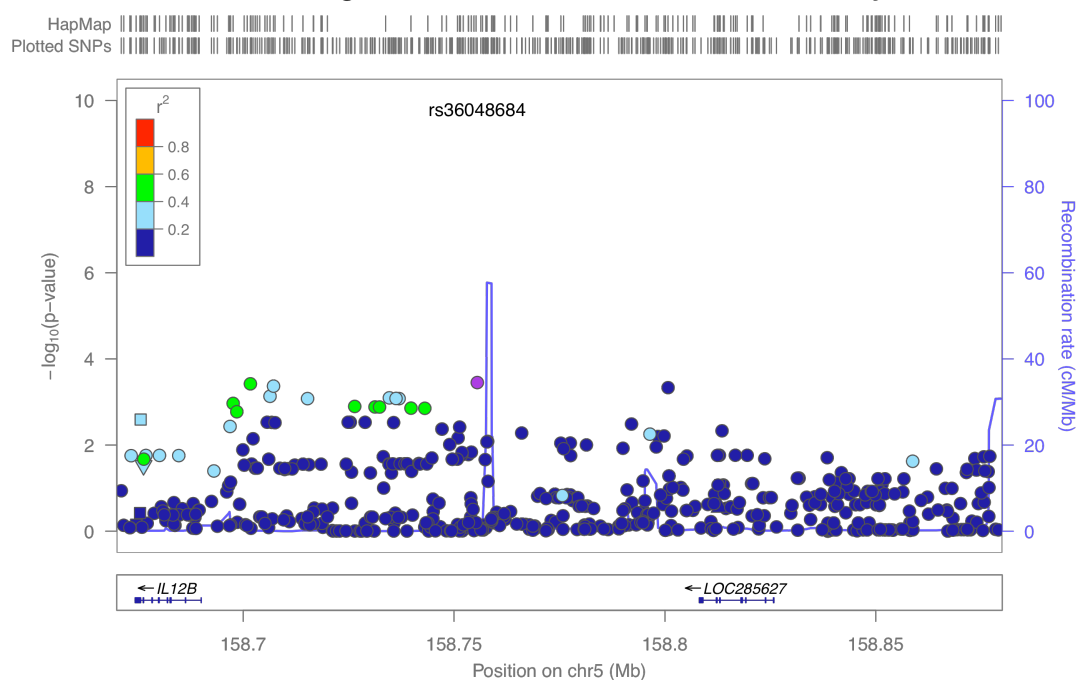


Plot 6.2. (A) *IL12B* region rs13188370 conditional analysis (rs2059131 conditional $p=1.0 \times 10^{-4}$), (B) *IL12B* region rs2059131 conditional analysis using imputed data only – both plots showing that rs2059131 is also independently associated with AS.

(A) *IL12B* Region rs13188370 rs2059131 Conditional Analysis



(B) *IL12B* Region rs2013718 Conditional Analysis



Plot 6.3. (A) *IL12B* region rs13188370 and rs2059131 conditional analysis (rs2013718 conditional $p=3.5 \times 10^{-3}$), (B) *IL12B* region rs2013718 conditional analysis using imputed data only – both plots showing that there is a third independent association in the region with rs2013718.

Sequences for SNPs selected for genotyping were obtained from the NCBI SNP database (<http://www.ncbi.nlm.nih.gov/snp/>, dbSNP build 132). These sequences were controlled for the presence of repeat sequences, which might have interfered with genotyping procedures, using the Ensembl database (http://www.ensembl.org/Homo_sapiens/Info/Index, GRCh37). SNPs within repeat sequences were replaced where possible. Two tagging SNPs within repeat sequences, rs175715 (*BATF*) and rs58532255 (*GAN*), were consequently replaced with SNPs in LD with them, rs175721 ($r^2 > 0.8$) and rs12447469 ($r^2 > 0.6$), respectively. Conditional association analyses showed that there were no residual associations within the regions not tagged by rs175721 and rs12447469 and therefore, these SNPs reflect the association at the original SNPs (See Appendix 3, Plots 6.9 and 6.11). The SNPs selected for genotyping are listed in Table 6.1

SNP	GENE	LOCATION GB 37 (ENSEMBL, GRCh37)
rs13188370	Upstream <i>IL12B</i>	5:158769129
rs2013718	Upstream <i>IL12B</i>	5:158828480
rs2059131	Upstream <i>IL12B</i>	5:158868175
rs2294851	<i>HHAT</i> - exon 6 NONSYN coding	1:210577884
rs77143393 (Chr12:42311603)	Upstream <i>IRAK4</i>	12:44025336
rs175721	Upstream <i>BATF</i>	14:75988131
rs12447469	<i>GAN</i> - intronic	16:81354858
rs77587461 (Chr19:10496951)	Upstream <i>ATG4D</i>	19:10635951
rs11666993	Downstream <i>SMARCA4</i>	19:11176075
rs60951857 (Chr19:4969039)	<i>JMJD2B</i> - intronic	19:5018039
rs4349859	HLA-B27 tagging, downstream HLA-B	6:31365787

Table 6.1. List of the SNPs selected for direct genotyping and their genomic locations.

6.2.3. Sample Population, Genotyping & Statistical Analysis

In order to replicate the associations at the 10 regions of interest, I compared 1804 AS cases from the UK with 1848 controls, comprising 720 healthy local blood donors and 1128 individuals with osteoarthritis (OA) as no genetic overlap has been identified between OA and AS. All cases in the study fulfilled the modified New York criteria for AS (See Chapter 2, Section 2.1) (van der Linden *et al.* 1984).

Genotyping of SNPs was undertaken by KBiosciences (Herts, UK), using the KASPar technology (See Chapter 2, Section 2.4.2.1). Genotypes with significant deviation from Hardy-Weinberg equilibrium ($p < 0.05$) were excluded from further analyses. A chi-square test and Cochran-Armitage test for trend were used as allelic and additive tests, respectively (See Chapter 2, Section 2.6.2). Cases of AS common to the present study and the original WTCCC2 study were excluded from association analyses ($n=347$), but were used to check the accuracy of imputation in the original study (See below). There were no control individuals common to this study and the WTCCC2 paper. Statistical analyses were performed using PLINK (v1.07, August 2009). Furthermore, allele frequencies of OA cases and healthy blood donors were compared using chi-square tests to ensure that these populations had similar allele frequencies at the variants of interest and it was appropriate to combine them. P-values (2-sided) less than 0.05 were considered statistically significant.

Statistical power to detect an association in the replication study was estimated using the software Quanto (Version 1.2.4, May 2009), assuming that the allelic odds ratio (OR) estimates from the imputed data analysis (Chapter 5) were an accurate reflection of the true underlying effect sizes (Table 6.2). A type I error level of $\alpha = 0.05$ (2-sided), a disease prevalence of 4:1000 and a log-additive disease model were used in these statistical power calculations. The study had 47%–98% power to identify SNPs conferring an OR of 1.5 with minor allele frequencies (MAFs) of 0.01–0.05, and 67%–97% power to identify SNPs with an OR of 1.2 with MAFs of 0.1–0.5.

6.2.4. Imputation Accuracy: Concordance between Imputed & Directly Typed Genotypes

There were 347 AS cases common to this study and the WTCCC2 study. Wet lab genotypes for these individuals were compared with imputed genotypes (See Chapter 5) in order to determine the accuracy of imputation. The concordance values for each SNP were reported as percentages (%). Genotypes were also stratified into 3 groups as major allele homozygotes, minor allele homozygotes and heterozygotes. Genotype concordances (%) were calculated separately for these groups. This stratification step was specifically important for low frequency variants (MAF<5%), which may have high genotype concordances just by correct imputation of major allele homozygotes.

6.3. Results

6.3.1. Concordance between Imputation & Direct Genotyping

Overall genotype concordance (%) between imputed and directly typed SNPs in cases (n=347) was >97% for 6 out of 9 SNPs, 93% for one SNPs. Two other SNPs, rs13188370 and rs2013718, had overall concordances of 87% and 14%, respectively (See Section 6.4.1, Table 6.2). When these genotype concordances were stratified into 3 groups as major allele homozygotes, minor allele homozygotes and heterozygotes, the observed heterozygote genotype concordances were lower for all low frequency variants ($MAF \leq 5\%$) compared to the overall concordance (Table 6.3). Low frequency variants, rs77143393 and rs175721, also had lower concordances for minor allele homozygotes (0% for both SNP). Two common SNPs, rs12447469 and rs 11666993, also had slightly lower concordance rates for minor allele homozygotes (91% and 97%, respectively). Two other common SNPs, rs2059131 and rs4349859, had lower concordances for major allele homozygotes (89% and 98%, respectively). Observed MAFs for all the SNPs followed the trend observed in the genotype concordances. MAFs of the low frequency variants (rs77143393 and rs175721) with low heterozygote concordances (89% and 80%, respectively) decreased, whereas MAF of the low frequency variant, rs77587461, slightly increased (minor allele count 22 vs 23, MAF 0.035 vs. 0.036) due to low heterozygote concordance (86%) and slightly lower major allele homozygote concordance (99%) compared to the other low frequency variants. MAFs of two SNPs, rs2294851 and rs4349859, with high overall and stratified genotype concordance rates stayed the same in both studies providing confidence in the

associations observed with these two SNPs in the analysis of imputed genotypes (Table 6.3).

SNP		rs13188370	rs2013718	rs2059131	rs2294851	rs77143393	rs175721	rs12447469	rs77587461	rs11666993	rs4349859
CHR BP (build 36)		5 158701707	5 158761058	5 158800753	1 208644507	12 42311603	14 75057884	16 79912359	19 10496951	19 11037075	6 31473766
Gene		Upstream <i>IL12B</i>	Upstream <i>IL12B</i>	Upstream <i>IL12B</i>	<i>HHAT</i> (NSYN)	Upstream <i>IRAK4</i>	Upstream <i>BATF</i>	<i>GAN-CMIP</i> (Intronic)	Upstream <i>ATG4D</i>	Downstream <i>SMARCA4</i>	HLA-B27 tagging SNP
Imputation Study (Chapter 5)	Cases/ Controls	1782/5167	1782/5167	1782/5167	1782/5167	1782/5167	1782/5167	1782/5167	1782/5167	1782/5167	1782/5167
	Minor Allele	A	A	C	T	T	G	G	T	G	A
	MAF - All (Cases/ Controls)	0.10 (0.082/ 0.10)	0.34 (0.37/ 0.33)	0.34 (0.38/ 0.34)	0.15 (0.13/ 0.16)	0.0355 (0.049/ 0.033)	0.07 (0.055/ 0.073)	0.46 (0.43/ 0.47)	0.058 (0.039/ 0.056)	0.36 (0.33/ 0.37)	0.15 (0.44/ 0.044)
	Allele, OR	A, 0.73	G, 0.83	T, 0.84	C, 1.25	T, 1.53	C, 1.37	G, 0.85	T, 0.62	G, 0.84	A, 50.00
	Chi-square	16.32	15.59	16.07	15.70	17.32	14.38	16.74	18.84	17.27	2379.74
	P-value	5.36 x 10 ⁻⁵	7.85 x 10 ⁻⁵	6.11 x 10 ⁻⁵	7.41 x 10 ⁻⁵	3.16 x 10 ⁻⁵	0.00015	4.29 x 10 ⁻⁵	1.42 x 10 ⁻⁵	3.24 x 10 ⁻⁵	<0.00
	RSQR	0.82	0.71	0.88	0.99	0.88	0.94	0.98	0.70	0.99	1.00
	Genotype*** Concordance	276/316 (87%)	45/312 (14%)	294/317 (93%)	313/314 (>99%)	311/316 (98%)	302/311 (97%)	305/315 (97%)	311/318 (97%)	313/315 (>99%)	314/317 (>99%)

Table 6.2. Analysis of imputed genotypes (See Chapter 5), and overall concordance (%; in bold) between directly typed and imputed genotypes. (***)Concordance between genotypes of cases common to the WTCCC2 imputation (Chapter 5) and this study.)

		Direct Genotyping				Imputation Concordance			
SNP	Gene	MAF in cases (Allele)	# of individuals genotyped successfully	All homozygotes Major, Minor Allele	Heterozygotes	Correctly imputed homozygotes; Major, minor (#, %)	Correctly imputed heterozygotes (#, %)	MAF in cases (Allele)	Overall genotype concordance
rs13188370	Ups. <i>IL12B</i>	0.036 (A)	316	293 293, --	23	255, 87% 255, 87%, --	21, 91%	0.10 (A)	87%
rs2013718	Ups. <i>IL12B</i>	--	312	312 (monomorphic)	--	45, 14%	--	--	14%
rs2059131	Ups. <i>IL12B</i>	0.38 (C)	317	164 121, 43	153	149, 91% 108, 89%, 41, 95%	145, 95%	0.40 (C)	93%
rs2294851	<i>HHAT</i>	0.11 (T)	314	251 247, 4	63	250, 99.6% 246, 99.6%, 4, 100%	63, 100%	0.11 (T)	>99%
rs77143393	Ups. <i>IRAK4</i>	0.06 (T)	316	281 280, 1	35	280, 99.6% 280, 100%, 0, 0%	31, 89%	0.05 (T)	98%
rs175721	Ups. <i>BATF</i>	0.06 (G)	311	276 275, 1	35	274, 99% 274, 99.6%, 0, 0%	28, 80%	0.05 (G)	97%
rs12447469	<i>GAN-CMIP</i>	0.42 (G)	315	163 107, 56	152	157, 96% 106, 99%, 51, 91%	148, 97%	0.41 (G)	97%
rs77587461	Ups. <i>ATG4D</i>	0.035 (T)	318	296 296, --	22	292, 99% 292, 99%, --	19, 86%	0.036 (T)	97%
rs11666993	Downs. <i>SMARCA4</i>	0.32 (G)	315	180 148, 32	135	179, 99% 148, 100%, 31, 97%	134, 99%	0.31 (G)	>99%
rs4349859	B27-tagging	0.42 (A)	317	63 58, 5	254	62, 98% 57, 98%, 5, 100%	252, 99%	0.42 (A)	>99%

Table 6.3. Concordance (%) between directly typed and imputed genotypes divided into three genotype groups. Results for lower

frequency SNPs (MAF<5%) are in bold. (Ups.: Upstream; Downs.: Downstream; #: Number)

6.3.2. Case-Control Study & HLA-B27-Stratified Analysis

Only 9 of the 11 SNPs selected for investigation were successfully genotyped. The assay for rs60951857 (*JMJD2B*) failed completely and rs2013718 proved to be monomorphic in the study population after two further attempts to type the variant using re-designed primers. OA cases and healthy blood donors did not differ in allele frequency across any of the SNPs selected for genotyping ($p > 0.05$ for all SNPs), and were therefore combined to form a larger statistically more powerful control population. Results for the allelic chi-square tests and the Cochran-Armitage test of trend for 9 SNPs that were successfully genotyped are shown in Tables 6.4 and 6.5, respectively. Other than the expected HLA-B27 association with AS (Table 6.4 & 6.5, $p < 0.00$, OR=15.6), there were no significant ($p < 0.05$) associations detected. There were also no statistically significant associations after stratifying cases and controls according to the HLA-B27 status (all $p > 0.05$, results not shown). All genotypes were in Hardy-Weinberg equilibrium as shown in Tables 6.6 and 6.7.

SNP	rs13188370	rs2013718*	rs2059131	rs2294851	rs77143393	rs175721	rs12447469	rs77587461	rs11666993	rs4349859
CHR BP (build 36)	5 158701707	5 158761058	5 158800753	1 208644507	12 42311603	14 75057884	16 79912359	19 10496951	19 11037075	6 31473766
Gene	Upstream <i>IL12B</i>	Upstream <i>IL12B</i>	Upstream <i>IL12B</i>	<i>HHAT</i> (NSYN, Exon 6)	Upstream <i>IRAK4</i>	Upstream <i>BATF</i>	<i>GAN</i> -CMIP (Intronic)	Upstream <i>ATG4D</i>	Downstream <i>SMARCA4</i>	HLA-B27 tagging SNP
Minor Allele, Major Allele	A, T	All A/A	C, T	T, C	T, C	G, C	G, A	T, C	G, A	A, G
MAF Cases/ Controls	0.035/ 0.030	-	0.31/ 0.30	0.14/ 0.15	0.040/ 0.038	0.070/ 0.075	0.47/ 0.48	0.047/ 0.048	0.35/ 0.35	0.44/ 0.048
Total Number of Cases/Controls	1804/1848	-	1804/1848	1804/1848	1804/1848	1804/1848	1804/1848	1804/1848	1804/1848	1804/1848
Allele, OR, (95% CI)	A, 1.16, (0.90-1.51)	-	T, 0.95, (0.86-1.05)	C, 1.06, (0.93-1.20)	T, 1.05, (0.83-1.33)	C, 1.09 (0.91-1.30)	G, 0.97 (0.88-1.06)	T, 0.98 (0.79-1.21)	G, 1.012 (0.92-1.11)	A, 15.6, (13.2-18.4)
Chi-square	1.30	-	0.87	0.80	0.15	0.86	0.52	0.04	0.06	1513
P-value	0.25	-	0.35	0.37	0.70	0.35	0.47	0.84	0.81	0.00
Power	20%	-	16%	15%	7%	16%	10%	5%	5%	99%

Table 6.4. Results of the statistical analysis for typed genotypes of the selected SNPs in the replication study. Power calculations

presented in this table were performed with the OR estimates in this replication study. *All cases and controls homozygous A/A. (CHR:

Chromosome, BP: Base pair, MAF: Minor allele frequency, OR: Odds ratio, NSYN: Non-synonymous coding SNP, CI: Confidence interval.)

SNP	rs13188370	rs2013718*	rs2059131	rs2294851	rs77143393	rs175721	rs12447469	rs77587461	rs11666993	rs4349859
CHR BP (build 36)	5 158701707	5 158761058	5 158800753	1 208644507	12 42311603	14 75057884	16 79912359	19 10496951	19 11037075	6 31473766
Gene	Upstream <i>IL12B</i>	Upstream <i>IL12B</i>	Upstream <i>IL12B</i>	<i>HHAT</i> (NSYN)	Upstream <i>IRAK4</i>	Upstream <i>BATF</i>	<i>GAN-CMIP</i> (Intronic)	Upstream <i>ATG4D</i>	Downstream <i>SMARCA4</i>	HLA-B27 tagging SNP
Minor Allele	A	All A/A	C	T	T	G	G	T	G	A
MAF Cases/ Controls	0.035/ 0.030	-	0.31/ 0.30	0.14/ 0.15	0.040/ 0.038	0.070/ 0.075	0.47/ 0.48	0.047/ 0.048	0.35/ 0.35	0.44/ 0.048
Total Number of Cases/Controls	1804/1848	-	1804/1848	1804/1848	1804/1848	1804/1848	1804/1848	1804/1848	1804/1848	1804/1848
Chi-square	1.28	-	0.86	0.79	0.16	0.88	0.51	0.04	0.06	1951
P-value	0.26	-	0.35	0.37	0.69	0.35	0.47	0.83	0.81	0.00

Table 6.5. Results of the Cochran-Armitage test of trend for typed genotypes of the selected SNPs in the replication study. (*All cases and controls homozygous A/A.) (CHR: Chromosome, BP: Base pair, MAF: Minor allele frequency, OR: Odds ratio)

SNP	rs13188370	rs2013718*	rs2059131	rs2294851	rs77143393	rs175721	rs12447469	rs77587461	rs11666993	rs4349859
Number of cases	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151
Chi-square	0.21	-	0.023	0.19	2.35	0.39	2.69	0.13	0.49	913.03
P-value	0.64	-	0.88	0.66	0.12	0.53	0.10	0.71	0.48	1.44 x 10 ⁻²⁰⁰

Table 6.6. Hardy-Weinberg equilibrium test for all case genotypes as a quality control. (*All A/A)

SNP	rs13188370	rs2013718*	rs2059131	rs2294851	rs77143393	rs175721	rs12447469	rs77587461	rs11666993	rs4349859
Number of controls	1848	-	1848	1848	1848	1848	1848	1848	1848	1848
Chi-square	3.34	-	0.45	0.50	0.048	1.21	0.37	0.39	0.47	1.99
P-value	0.068	-	0.50	0.48	0.83	0.27	0.54	0.53	0.49	0.16

Table 6.7. Hardy-Weinberg equilibrium test for all control genotypes as quality control. (*All A/A)

6.4. Discussion

6.4.1. Imputed Genotypes vs. Directly Typed Genotypes

The high overall genotype concordances (97%) between typed and imputed genotypes for at least six of the SNPs identified in Chapter 5 and the HLA-B27 tag SNP rs4349859, suggest that imputation was likely to be accurate for these SNPs, and consequently, that the association test statistics which were based on them were unlikely to reflect imputation error. Even though the genotype concordances were assessed using only on a limited number of cases (n=347), this number is comparable to other studies, which have investigated imputation accuracy, and should be sufficient to assess concordance, at least in the case of common SNPs (Pei *et al.* 2008; Hao *et al.* 2009; Li *et al.* 2009; Nothnagel *et al.* 2009; Sung *et al.* 2012). Additionally, the fact that directly genotyped and imputed common (MAF>5%) SNPs had similar MAFs is consistent with the imputation being accurate for these SNPs (Tables 6.2 - 6.4). However, common SNPs, rs2059131 and rs12447469, had low overall concordance (93%) and low minor allele homozygote concordance (91%), respectively. Therefore, the associations observed in the analysis of imputed genotypes for these SNPs may be due to imputation error resulting in type I error (i.e. false positives) (Tables 6.2 & 6.3).

Imputation of low frequency SNPs (MAF<5%) are likely to be more prone to error, specifically because of errors in imputing heterozygotes and minor allele homozygotes. In this study, such imputation errors were observed when genotype concordances were stratified into major allele homozygotes, minor

allele homozygotes and heterozygotes (Table 6.3). Low frequency SNPs, rs77143393, rs175721 and rs77587461, had heterozygote concordances of 89%, 80% and 86%, respectively. Minor allele homozygote concordance was 0% for two of these SNPs (rs77143393 and rs175721). These errors were reflected in the MAFs of the low frequency SNPs. Even such slight changes in the MAFs of low frequency variants may drastically affect the results of association analysis. These changes may either influence the statistical power (i.e. the lower the MAF, the lower the statistical power of a study to detect an association), or lead to false positives (type I error), depending on the direction of the change in the MAFs in both cases and controls. For example, when imputing a SNP that has a MAF of 1%, an imputation routine might always predict the common homozygote, in which case the overall genotype concordance would be high (~99%), but the power to detect association would be zero because the other genotype classes are never called correctly. Therefore, it is important to examine whether the imputation program is calling heterozygotes and rare homozygotes correctly for low frequency variants, otherwise power can be severely affected.

R^2 (RSQR) values measuring imputation quality varied between 0.70 and 1.00 for the SNPs of interest (Table 6.2). An R^2 threshold of 0.3 is generally recommended for filtering imputed SNPs in genomewide association studies (Li *et al.* 2010).

However, when using the 1000 Genomes pilot project dataset as a reference panel for imputation, a threshold of 0.5 is recommended as the genotypes are of lower quality compared to the HapMap genotypes (Abecasis 1000 Genomes Imputation Cookbook;

http://genome.sph.umich.edu/wiki/MaCH:_1000_Genomes_Imputation_Cookbo

ok). In the present study, the imputation quality values were consistent with the genotype concordances (Tables 6.2 & 6.3). Additionally, the mach2dat program used for analysis of imputed genotypes takes imputation uncertainty into account in construction of its association test statistics by regressing affection status on expected allelic dosage (rather than best guess genotype) and therefore, the association test statistics are often robust to small degrees of error/uncertainty in prediction.

The *IL12B* region contained a variant with a very low concordance (rs2013718), and two other SNPs, which had more moderate genotype concordances of 87% (rs13188370) and 93% (rs2059131) (Table 6.2 & 6.3). The low concordances may be due to genotyping errors in the wet lab sample, or errors in the imputation. According to the 1000 Genomes dataset (August 2010, 629 individuals of mixed origin), rs13188370 has a MAF of 0.014 (A, allele count 17/1258) (NCBI SNP database). Moreover, rs13188370 has a MAF of 0.033 (allele A) according to the HapMap CEU population dataset (n=60), and a MAF of 0.050 according to the 1000 Genomes data on the same CEU population (n=54). Direct genotyping results for rs13188370 are in agreement with the MAFs from populations of the European descent (MAF~0.03) (Figure 6.2 for rs13188370 genotyping cluster plot), whereas the MAF calculated from the imputed genotypes is not. Therefore, it is more likely that there was inaccuracy in the imputation of this SNP and we should be cautious about the association signals in this region.

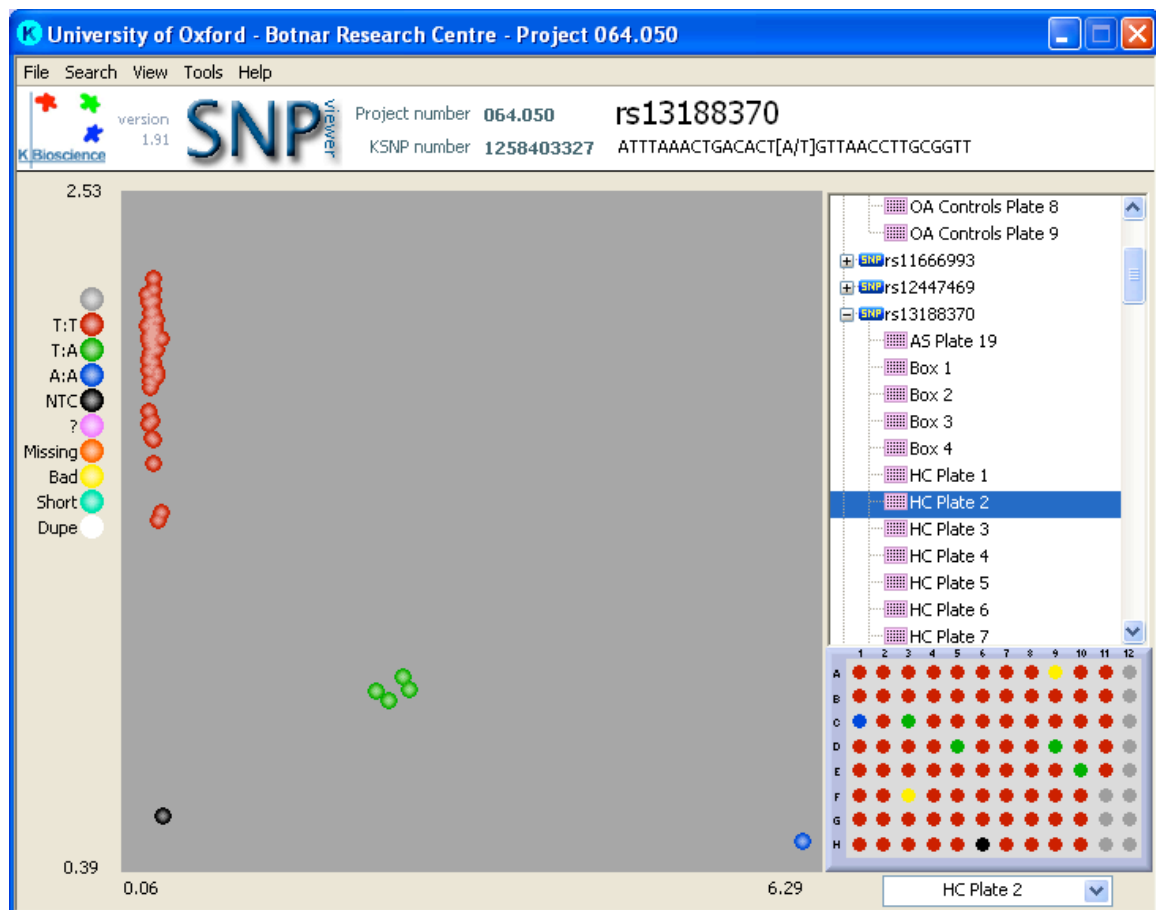


Figure 6.2. rs13188370 cluster plot showing three separate genotypes clusters.

This plot provides further support for the accuracy of the direct genotyping.

On the other hand, rs2013718 in the *IL12B* region has a MAF of 0.485 (G, allele count 610/1258, 1000 Genomes Project, August 2010, 629 individuals of mixed origin), and is located within a repeat sequence (NCBI SNP database). Repeat sequences can cause genotyping problems because of difficulty in designing specific primers. rs2013718 was genotyped twice using different DNA strands for primer design. In both cases, the SNP was monomorphic (allele A). The cluster plot for rs2013718 showed one distinct genotype cluster for A/A homozygotes suggesting that there was no problem with the genotype clustering

(Figure 6.3). Thus, it is more likely that the low concordance (14%) for rs2013718 is due to a genotyping error because of the repeat sequences surrounding the SNP. Therefore, association results from this SNP should not be trusted and treated with caution.

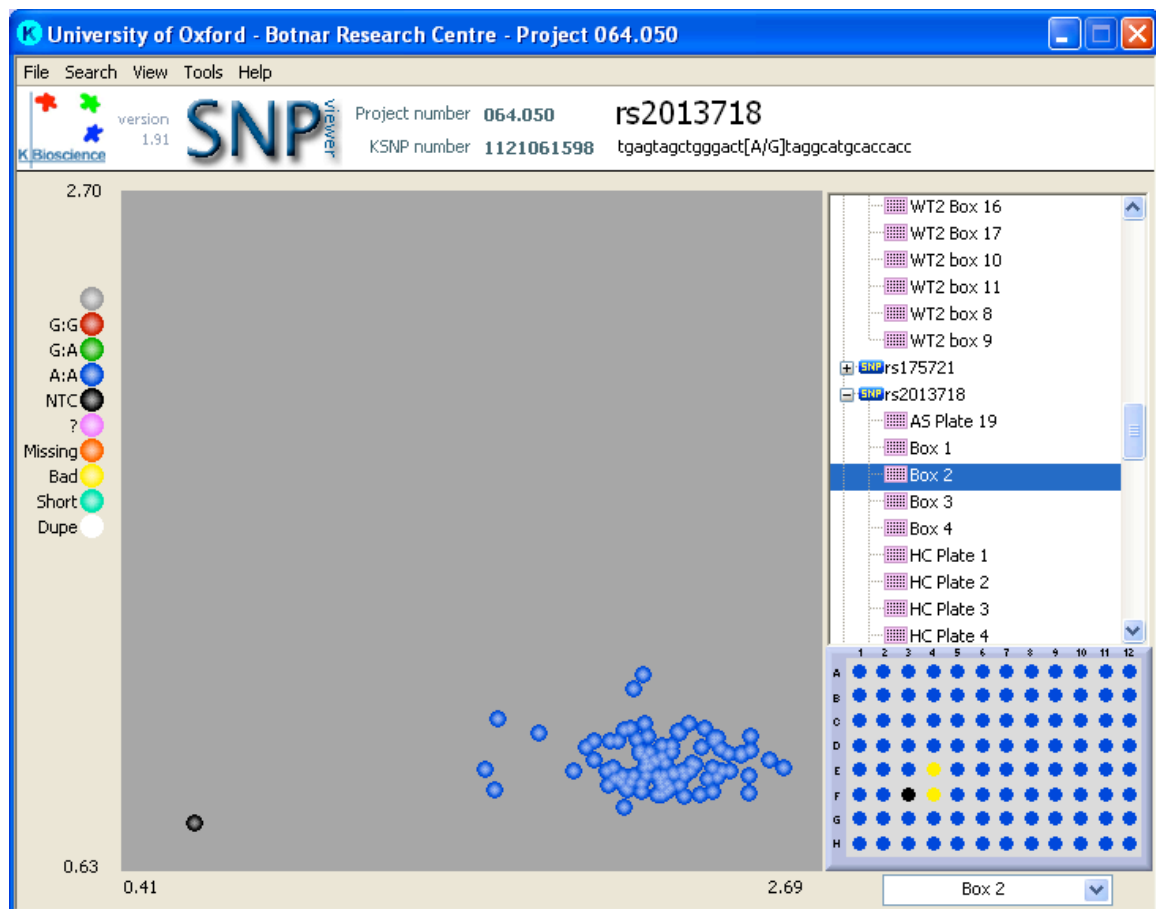


Figure 6.3. rs2013718 cluster plot showing only one genotype (A/A) cluster.

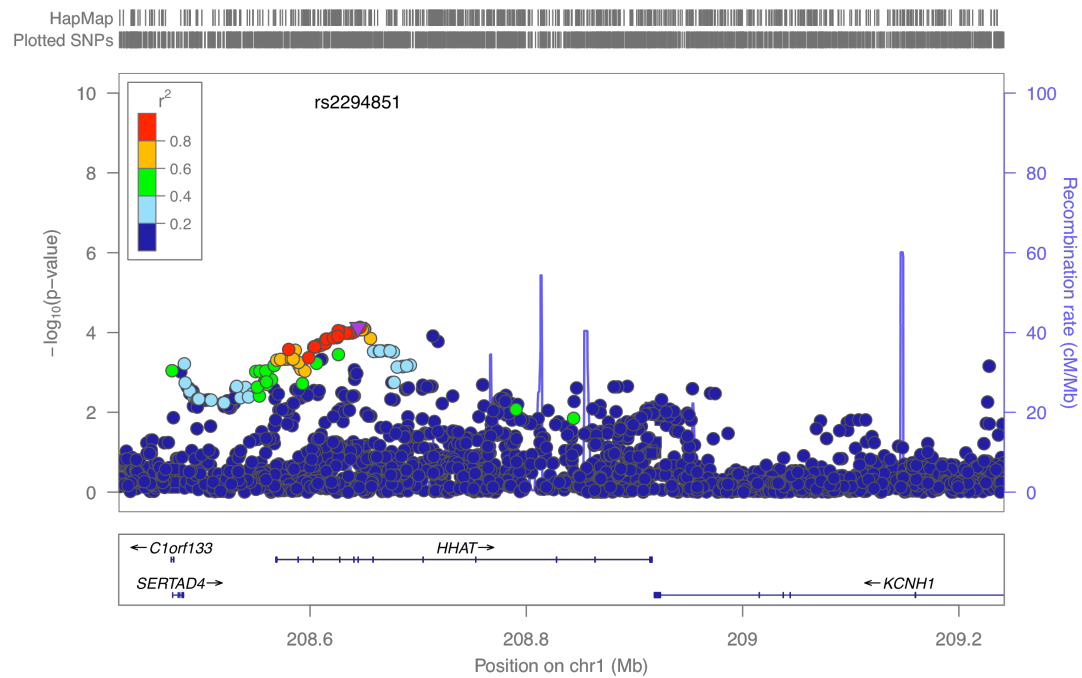
This plot illustrates that the SNP is monomorphic in the study population according to direct genotyping.

It is also important to note that imputation was carried out with the 1000 Genomes pilot project dataset acquired from only 61 CEPH (Centre d'Etude du Polymorphisme Humain; Utah residents of Northern and Western Europe ancestry) individuals (122 haplotypes, August 2009 data release). The August 2009 release of the 1000 Genomes Project was the only whole genome high resolution dataset available when this imputation project commenced. There are studies showing that increasing the size of the reference dataset increases the imputation accuracy (Pei *et al.* 2008; Li *et al.* 2009; Sung *et al.* 2012). Imputation with a more recent 1000 Genomes reference dataset (e.g. the most recent 2011 phase I release including 381 individuals of European descent), which contains a larger sample of rare/low frequency haplotypes and variants (as well as a more precise estimate of their frequencies), would be expected to produce a more accurate and informative imputation, particularly with respect to low frequency variants (Jostins *et al.* 2011). Further releases of the 1000 Genomes Project will likely increase imputation accuracy and information further as the size of the reference set and number of novel variants increase (Li *et al.* 2009; Genomes 2010).

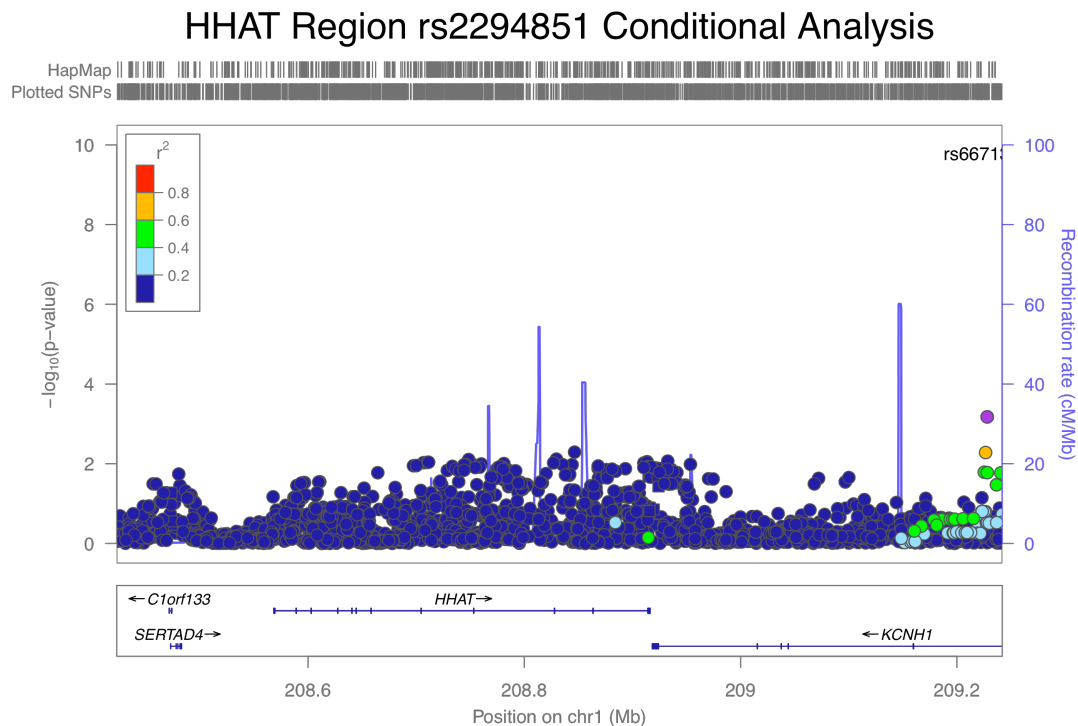
6.4.2. Chromosome 1q32 – *HHAT*

rs2294851, a non-synonymous coding SNP (Ser182Asn) which has been identified as a deleterious and probably damaging variant (SIFT score=0.02, Polyphen score=0.998; <http://www.ensembl.org>), and the top hit in *HHAT*, was associated with AS in the analysis of imputed genotypes (Table 6.2; $p=7.41 \times 10^{-5}$, OR=1.25; Plots 6.4 and 6.5). However, it showed no evidence of association with AS in the direct genotyping study (Tables 6.4 & 6.5). This might reflect the low statistical power of the replication study (Table 6.4, power=15%; further discussed in Section 6.4.7). However, imputation accuracy of rs2294851 was high with an overall genotype concordance of >99% (Table 6.3). Therefore, the association signal observed in the analysis of imputed genotypes is very likely to be real.

HHAT Region Unconditional Analysis



Plot 6.4. *HHAT* region unconditional analysis – showing the top hit SNP (rs2294851, purple) in the analysis of imputed genotypes, and the LD pattern relative to this top SNP.



Plot 6.5. *HHAT* region rs2294851 conditional analysis using imputed data only – showing that rs2294851 association was the only independent association in the region ($p < 5 \times 10^{-3}$).

HHAT encodes the hedgehog acyltransferase protein, which is also known as skinny hedgehog (SHH). This enzyme functions in hedgehog (HH) palmitoylation, a lipid modification leading to the addition of a hydrophobic palmitate to the N-terminal cysteine of HH (Chamoun *et al.* 2001; Buglino & Resh 2008). The palmitoylation step is important in HH modification as the non-palmitoylated form has a 30-fold decrease in potency i.e. lower signalling efficiency, and when unpalmitoylated, protein distribution (i.e. the morphogenic gradient) is also affected due to lack of formation of active soluble multimeric complex (Pepinsky *et al.* 1998; Chen *et al.* 2004). The non-synonymous mutation (rs2294851), Ser182Asn, is within a predicted loop region. According to *in vitro*

mutation analysis targeting highly conserved hydrophobic residues and mutating these residues to alanine, the Ser182Asn mutation does not affect the palmitoylation activity of HHAT. However, it is possible that this variant influences protein stability, which is also significant for protein function (Buglino & Resh 2010).

Mammals have three homologs of the hedgehog protein; sonic hedgehog (SHH), indian hedgehog (IHH) and desert hedgehog (DHH) (Stanton & Peng 2010). HH is a morphogen - a diffusible protein that exerts control over morphogenesis by forming a concentration gradient - and plays significant roles in many developmental processes such as development of digits on limbs, nervous system and various tissues such as osteoblastic and gastrointestinal cells via the hedgehog signaling pathway (Hooper & Scott 2005; Stanton & Peng 2010).

It has been shown that mouse embryos lacking skinny hedgehog activity are deficient in long-range SHH signaling and therefore, have defective neural tube and limb development. In addition, these mice have defective skeletal development with limbs displaying dwarfism and reduced chondrocyte proliferation due to defective long-range signaling by unpalmitoylated IHH (Chen *et al.* 2004). IHH also plays an important role in osteoblast differentiation in endochondral bone formation and the HH pathway has been suggested as a therapeutic target in osteoarthritis (Long *et al.* 2004; Hu *et al.* 2005; Rockel & Alman 2011).

It has also been shown that thymic epithelium employs SHH to regulate T cell proliferation, and SHH is also necessary for repertoire selection (Crompton *et al.* 2007). Moreover, it has been demonstrated that IHH produced by thymocytes regulates T cell development in the fetal and adult thymus (Outram *et al.* 2009). Mutations in *HHAT* are critical as they affect the essential roles played by HH. Defects in palmitoylation of HH by HHAT may either affect the role of HH in osteoblast differentiation, or in T cell development and maturation. Both of these scenarios could be important in AS pathogenesis.

6.4.3. Chromosome 5q31 – *IL12B*

IL12B encodes the p40 subunit of the cytokines interleukin (IL) 12 and 23 secreted by activated macrophages of the innate immune system. The p40 subunit forms heterodimers either with the p35 subunit to form IL12, or with the p19 subunit to form IL23. Cytokines IL12 and IL23 are required for the development of T_H1, and the maintenance of T_H17 T cell responses of cell-mediated immunity, respectively (Oppmann *et al.* 2000). Association has previously been reported and replicated between *IL23R* and AS (Burton *et al.* 2007; Karaderi *et al.* 2009). Polymorphisms in *IL12B* have also been associated with Crohn's disease (CD) (Libioulle *et al.* 2007; Cho 2008) and psoriasis (Capon *et al.* 2007; Cargill *et al.* 2007; Zhang *et al.* 2009).

There were 3 potentially independent hits upstream of *IL12B* in the analysis of the imputed genotypes (Table 6.2, rs13188370, rs2059131, rs2013718; Plots 6.1 - 6.3). None of these variants was associated with AS in the replication study (Tables 6.4 & 6.5). Statistical power was between 16%-20% for the analyses of

the SNPs rs13188370 and rs2059131, whereas rs2013718 was monomorphic in the replication sample. Errors in either the imputation or the genotyping may be responsible for these discordant results (Table 6.3; further discussed in Sections 6.4.1 & 6.4.7).

6.4.4. Chromosome 6 – rs4349859 (HLA-B27)

The HLA-B27 tagging SNP rs4349859 was strongly associated with AS in the replication study (Tables 6.4 & 6.5, $p < 0.00$, OR=15.6). The aim of rs4349859 genotyping was to stratify cases and controls according to HLA-B27 status in order to look for differential associations between HLA-B27 positive and negative individuals. No statistically significant associations were observed in the HLA-B27-stratified analyses.

6.4.5. Other Genes/Regions of Interest

In the analysis of imputed genotypes, statistically significant associations with rs77143393 (*IRAK4*), rs175721 (*BATF*), rs12447469 (*GAN-CMIP*), rs77587461 (*ATG4D*) and rs11666993 (*SMARCA4*) were observed (Table 6.2). However, these associations failed to replicate in this study (Tables 6.4 & 6.5). Statistical power of this study was low (Table 6.4; power<20%) and therefore, it is necessary to replicate these analyses in a larger study population (Further discussed in Sections 6.4.1 & 6.4.7).

6.4.5.1. Chromosome 19p13 – *JMJD2B*

rs60951857, the most strongly associated variant in *JMJD2B* in the imputation study (Table 6.2), has a MAF of 0.18 (allele count 227/1258) according to the 1000 Genomes Project (NCBI, dbSNP). Even though genotyping both positive and negative strands was attempted, rs60951857 genotyping failed possibly due to its location within a repeat sequence.

6.4.6. Control Group & Selection Bias: Healthy Controls & Osteoarthritis Cases

The control group included healthy controls from British blood donors and OA cases in order to increase the sample size and consequently, the statistical power of the replication study. However, it was necessary to consider the role of possible selection bias in this strategy. Selection bias refers to a statistical bias where selected groups (e.g. case and control groups) are not representative of the larger populations they belong to, and this biases estimation of the population parameters of interest (e.g. the odds ratios). For instance, use of certain groups of individuals as controls (e.g. OA cases in this study) may affect the overall allele frequencies in the controls, if allele frequencies at loci of interest are significantly different in OA cases than those of the healthy blood donors. However selection bias did not appear to influence the results of this study, since there were no statistically significant differences in allele frequency between OA cases and British blood donors at any of the SNPs genotyped in this study. It was therefore appropriate to combine the two samples in the analysis.

6.4.7. Further Considerations & Future Work

The replication study was underpowered (Table 6.4, power range: 5%-20%) to detect the genetic risk variants of small effect identified in the imputation study. This was because the WTCCC2 study population used for the discovery analysis was larger than the replication study (WTCCC2 1782 cases vs. 5167 controls; replication study 1804 cases vs. 1818 controls). Published guidelines for replicating genomewide association studies recommend that discovery and replication samples be of similar size (Chanock *et al.* 2007). However, it is also likely that the effect sizes observed in the WTCCC2 imputation study (Table 6.2) were inflated because of winner's curse and imputation errors, and hence the power calculations reported in this chapter probably overestimated the potential to detect an association (Tables 6.2 & 6.4). It is therefore necessary to perform a larger replication study to confirm the non-significant findings from this relatively small study, where power is likely to be an issue.

Meta-analysis is also a useful statistical tool to combine the results from independent genetic association studies into a single effect estimate [e.g. (Franke *et al.* 2010) for inflammatory bowel disease] particularly where effect sizes are small ($OR < 1.5$), similar to those observed in this study (e.g. $OR \sim 1.1$). Meta-analysis using the data from the current study and future studies may provide further evidence for possible novel associations. Moreover, repeating the imputation using a more recent and larger data release of the 1000 Genomes Project may also help confirm the current observations as well as identifying new associations.

To sum up, imputation is a reliable and cost-effective tool for *in silico* fine-mapping and determining regions of interest for further fine-mapping and functional studies. Imputation accuracy, specifically for low frequency variants, will be further improved by expanding reference populations and developing technologies such as DNA sequencing, genotype calling, and computational power to impute with large datasets (Li *et al.* 2009; Genomes 2010; Jostins *et al.* 2011). With respect to the replication of associations observed in the previous imputation study of the WTCCC2 GWAS, the only novel association that is likely to be real is with the rs2294851 SNP (*HHAT*), which has a high overall genotype concordance (>99%). This association still requires robust confirmation in replication studies. There is no obvious biological connection between this gene and pathways known to be involved in AS. However, either *HHAT* itself or another close by gene may prove important in AS pathogenesis and provide a new therapeutic target.

6.5. References

- Buglino, J. A. & Resh, M. D. (2008). "Hhat is a palmitoylacyltransferase with specificity for N-palmitoylation of Sonic Hedgehog." J Biol Chem **283**(32): 22076-22088.
- Buglino, J. A. & Resh, M. D. (2010). "Identification of conserved regions and residues within Hedgehog acyltransferase critical for palmitoylation of Sonic Hedgehog." PLoS One **5**(6): e11195.
- Burton, P. R., Clayton, D. G., *et al.* (2007). "Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants." Nat Genet **39**(11): 1329-1337.
- Capon, F., Di Meglio, P., *et al.* (2007). "Sequence variants in the genes for the interleukin-23 receptor (IL23R) and its ligand (IL12B) confer protection against psoriasis." Hum Genet **122**(2): 201-206.
- Cargill, M., Schrodi, S. J., *et al.* (2007). "A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes." Am J Hum Genet **80**(2): 273-290.
- Chamoun, Z., Mann, R. K., *et al.* (2001). "Skinny hedgehog, an acyltransferase required for palmitoylation and activity of the hedgehog signal." Science **293**(5537): 2080-2084.
- Chanock, S. J., Manolio, T., *et al.* (2007). "Replicating genotype-phenotype associations." Nature **447**(7145): 655-660.
- Chen, M. H., Li, Y. J., *et al.* (2004). "Palmitoylation is required for the production of a soluble multimeric Hedgehog protein complex and long-range signaling in vertebrates." Genes Dev **18**(6): 641-659.
- Cho, J. H. (2008). "The genetics and immunopathogenesis of inflammatory bowel disease." Nat Rev Immunol **8**(6): 458-466.
- Crompton, T., Outram, S. V., *et al.* (2007). "Sonic hedgehog signalling in T-cell development and activation." Nat Rev Immunol **7**(9): 726-735.
- Evans, D. M., Spencer, C. C., *et al.* (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**(8): 761-767.
- Franke, A., McGovern, D. P., *et al.* (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." Nat Genet **42**(12): 1118-1125.
- Genomes (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.
- Hao, K., Chudin, E., *et al.* (2009). "Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies." BMC Genet **10**: 27.
- Hooper, J. E. & Scott, M. P. (2005). "Communicating with Hedgehogs." Nat Rev Mol Cell Biol **6**(4): 306-317.
- Hu, H., Hilton, M. J., *et al.* (2005). "Sequential roles of Hedgehog and Wnt signaling in osteoblast development." Development **132**(1): 49-60.
- Jostins, L., Morley, K. I., *et al.* (2011). "Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets." Eur J Hum Genet **19**(6): 662-666.
- Karaderi, T., Harvey, D., *et al.* (2009). "Association between the interleukin 23 receptor and ankylosing spondylitis is confirmed by a new UK case-

- control study and meta-analysis of published series." Rheumatology (Oxford) **48**(4): 386-389.
- Li, Y., Willer, C., *et al.* (2009). "Genotype imputation." Annu Rev Genomics Hum Genet **10**: 387-406.
- Li, Y., Willer, C. J., *et al.* (2010). "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." Genet Epidemiol **34**(8): 816-834.
- Libioulle, C., Louis, E., *et al.* (2007). "Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4." PLoS Genet **3**(4): e58.
- Long, F., Chung, U. I., *et al.* (2004). "Ihh signaling is directly required for the osteoblast lineage in the endochondral skeleton." Development **131**(6): 1309-1318.
- Marchini, J. & Howie, B. (2010). "Genotype imputation for genome-wide association studies." Nat Rev Genet **11**(7): 499-511.
- Nothnagel, M., Ellinghaus, D., *et al.* (2009). "A comprehensive evaluation of SNP genotype imputation." Hum Genet **125**(2): 163-171.
- Oppmann, B., Lesley, R., *et al.* (2000). "Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12." Immunity **13**(5): 715-725.
- Outram, S. V., Hager-Theodorides, A. L., *et al.* (2009). "Indian hedgehog (Ihh) both promotes and restricts thymocyte differentiation." Blood **113**(10): 2217-2228.
- Pei, Y. F., Li, J., *et al.* (2008). "Analyses and comparison of accuracy of different genotype imputation methods." PLoS One **3**(10): e3551.
- Pepinsky, R. B., Zeng, C., *et al.* (1998). "Identification of a palmitic acid-modified form of human Sonic hedgehog." J Biol Chem **273**(22): 14037-14045.
- Pruim, R. J., Welch, R. P., *et al.* (2010). "LocusZoom: regional visualization of genome-wide association scan results." Bioinformatics **26**(18): 2336-2337.
- Rockel, J. S. & Alman, B. A. (2011). "Don't hedge your bets: hedgehog signaling as a central mediator of endochondral bone development and cartilage diseases." J Orthop Res **29**(6): 810-815.
- Stanton, B. Z. & Peng, L. F. (2010). "Small-molecule modulators of the Sonic Hedgehog signaling pathway." Mol Biosyst **6**(1): 44-54.
- Sung, Y. J., Wang, L., *et al.* (2012). "Performance of genotype imputations using data from the 1000 genomes project." Hum Hered **73**(1): 18-25.
- van der Linden, S., Valkenburg, H. A., *et al.* (1984). "Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria." Arthritis Rheum **27**(4): 361-368.
- Zhang, X. J., Huang, W., *et al.* (2009). "Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21." Nat Genet **41**(2): 205-210.

CHAPTER 7

7. CONCLUSIONS & FUTURE DIRECTIONS

7.1. General Conclusions

It is now very clear that ankylosing spondylitis (AS) is a complex disease in which numerous genetic factors are involved. The aim of my research was to identify and to confirm some of these genetic factors. My research included replication and refinement of associations initially identified by the Wellcome Trust Case Control and the Triple A –Australo-Anglo-American Spondylitis Consortia (WTCCC-TASC) (Burton *et al.* 2007). I performed an extended UK study and a meta-analysis of the association with *IL23R* and confirmed its association with AS. The non-synonymous coding SNP rs11209026 showed strong association with AS (OR=0.61, 1.5×10^{-9}) and is likely to be causally related. This SNP has also shown strong association with Crohn's disease (CD) and psoriasis (Duerr *et al.* 2006; Cargill *et al.* 2007). The IL23 receptor plays a crucial role in the expansion and maintenance of T_H17 T cells. These T cells are involved in tissue inflammation and production of IL17, which has been shown to have increased serum levels in AS patients (See Section 7.2.2) (Mei *et al.* 2011).

A fine mapping project involving re-sequencing of the upstream flanking sequences, exon-intron boundaries and exons of the *TNFRSF1A* gene to replicate

and refine its association with AS was also carried out. *TNFRSF1A* is a strong functional candidate as its variants have been previously associated with AS in the TASC GWAS and it encodes the p55 receptor in the TNF pathway (Reveille *et al.* 2010). This pathway plays an important role in AS, and TNF blockade with recombinant TNF receptor/Ig fusion protein (etanercept) is a highly effective treatment in AS (Davis *et al.* 2008). This study identified a rare SNP (MAF<1%) in *TNFRSF1A* and demonstrated the existence of weak association with AS (rs4149584, OR=1.60, p=0.01). Sub-group analysis, in which rs4149570 was associated with age of onset (age>25, p=0.0006, OR=1.4), also provided further clues about the nature of the association between *TNFRSF1A* and AS. However, such subgroup analysis is potentially misleading in the absence of an *a priori* hypothesis. Therefore, the association between rs4149570 and age of onset requires robust independent replication. *TNFRSF1A* variants have also shown association with AS in Han Chinese (Davidson *et al.* 2011), which provided additional support for the observed association. But, it still requires robust replication to refine and identify precisely which variant is functional.

I then performed *in silico* studies to follow up several positive associations identified by the genomewide association studies conducted by the WTCCC-TASC (Evans *et al.* 2011). *In silico* studies were also used to detect novel associations with low frequency variants. These *in silico* studies used techniques such as genotype imputation using the 1000 Genomes Pilot Project data as a reference population and statistical analysis methods that involved single locus testing and rare/low frequency variant collapsing strategies to test for association with AS (Li *et al.* 2009; Genomes 2010; Morris & Zeggini 2010). In addition to the

previously identified associations with variants in *ERAP1*, *IL23R*, *KIF21B*, *HLA-B*, *2p15*, *21q22*, *RUNX3*, *IL12B*, *LTBR-TNFRSF1A*, *ANTXR2*, *PTGER4*, *CARD9* and *TBKBP1* along with the *HLA-B27-ERAP1* interaction, a number of genes/regions of interest were identified based on evidence of association ($p < 10^{-4}$), biological/functional plausibility and associations with other inflammatory diseases such as CD and psoriasis. Replication studies in independent populations are necessary to confirm these associations. Furthermore, a novel and strong association with rs4869313 (*ERAP2*) was identified in this study ($OR=0.79$, $p=7.3 \times 10^{-8}$). This association was not apparent in the initial analysis due to the haplotype structure in the region. Similar to *ERAP1*, *ERAP2* functions as an aminopeptidase in trimming peptides for antigen presentation (Nguyen *et al.* 2011; Birtley *et al.* 2012).

Ten SNPs from the genes/regions of interest identified in the previous imputation project were genotyped in an independent replication study to determine the accuracy of imputation and to confirm these associations. The accuracy of imputation was high for common SNPs (overall genotype concordance $\sim 97\%$). However, lower frequency variants ($MAF < 5\%$) showed less accuracy when genotype concordances were stratified (heterozygote concordance $\sim 86\%$, minor allele homozygote concordance = 0%). Even though imputation of common SNPs is a reliable and cost-effective tool for *in silico* fine-mapping and determining regions of interest for further fine-mapping and functional studies, imputation accuracy for low frequency SNPs will be further improved by expanding reference populations and developing technologies such as DNA sequencing, genotype calling, and computational power to impute with

large datasets (Li *et al.* 2009; Genomes 2010; Jostins *et al.* 2011). With regards to the independent replication study, statistically significant ($p < 0.05$) associations were not observed with any of the SNPs. This may be due to the low statistical power given the small effect sizes ($OR < 1.2$). However, the association observed with the non-synonymous coding SNP rs2294851 (*HHAT*) in the previous imputation study ($OR = 1.25$, 7.41×10^{-5}) was shown to be accurate (overall genotype concordance $> 99\%$), which may indicate that this gene or another gene in the same region may play a role in the pathogenesis of AS.

The detection of genetic associations with AS is potentially valuable in the identification of pathological pathways that may harbour possible novel therapeutic targets for treating AS. For instance, confirmation of the *IL23R* association in AS has supported the use of anti-IL17A monoclonal antibodies (secukinumab) in AS treatment, which has already proved efficacious in preliminary studies (Toussirot 2012). Furthermore, functional studies have now demonstrated that loss of function variants of *ERAP1*, including rs30187 (R528K), are associated with protection against AS (Kochan *et al.* 2011). Work is already under way to develop small molecule inhibitors of ERAP1 that could have similar beneficial therapeutic effects. The resolution of the crystal structure of ERAP1 to $2\text{\AA}$ may facilitate the design of a more effective and specific inhibitor (Kochan *et al.* 2011).

More potential therapeutic targets are likely to be identified through these ongoing genetic studies with the analysis of larger study populations and the development of new statistical methods to achieve greater statistical power. This

is particularly true for rare to low frequency variants ($0.1\% < \text{MAF} < 5\%$) with modest to intermediate effect sizes ($1.5 < \text{OR} < 3.0$) that may prove to be functional. Functional variants are likely to be subject to natural selection limiting their frequency in the general population and are therefore likely to be of low frequency (Manolio *et al.* 2009; Marth *et al.* 2011). Thus, low frequency variants may be important in disease aetiology and point to pathways that are targets for novel treatments. However, there is a limit to identifying functional variants by genetics alone. This is sometimes because of linkage disequilibrium (LD) across a region of interest interfering with the identification of the actual causal variant (such the *ERAP1* regions). Therefore, functional studies are necessary to explore the structure and interactions of proteins encoded by strong candidate genes. On the other hand, it is also possible that detection of even weak genetic associations may lead to identification of biologically-relevant pathways that harbour therapeutic targets (e.g. IL17 in the T_H17 pathway).

7.2. General Discussion & Future Directions

Genome-wide association studies have identified several genes and loci with major involvement in AS, including *ERAP1*, *IL23R*, *KIF21B*, *HLA-B*, *2p15*, *21q22*, *RUNX3*, *IL12B*, *LTBR-TNFRSF1A*, *ANTXR2*, *PTGER4*, *CARD9* and *TBKBP1* using a single locus approach (Burton *et al.* 2007; Reveille *et al.* 2010; Evans *et al.* 2011). A strong gene-gene interaction between HLA-B27 and *ERAP1* has also been found (Evans *et al.* 2011). I used the WTCCC2 AS GWAS data to explore the use of other statistical approaches, including imputation of untyped genotypes and rare and low frequency variant analysis. Many common variants ($\text{MAF} > 5\%$) of

modest risk ($1.1 < OR < 1.5$) have been identified by GWAS, but it is likely that numerous variants, not tagged well by single common SNPs, still remain to be discovered (Manolio *et al.* 2009). I have identified additional associated variants by imputing with the 1000 Genomes Project pilot data (August 2009). This approach has improved the coverage of common variants, allowed fine-mapping by increasing the density of SNPs in associated regions and generated genotyping data on low frequency variants ($MAF < 5\%$).

However, imputation with the most recent and larger 1000 Genomes Project data and low frequency SNP analysis with newly developed more sophisticated programmes such as C-alpha, which uses novel clustering strategies, will increase confidence in the identified associations (Neale *et al.* 2011). By combining AS GWAS data (Burton *et al.* 2007; Reville *et al.* 2010; Evans *et al.* 2011) and using imputation to infer un-typed genotypes, a larger study population could also be obtained which should provide more statistical power to detect genetic associations with AS. This will be particularly useful for detecting associations with rare and low frequency variants ($MAF < 5\%$), as well as refining previously identified associations by meta-analysis (Willer *et al.* 2010). Even though validation and refinement of detected genetic associations in independent populations and in populations with different ethnic backgrounds is important, a major continuing problem with the analysis of low frequency variants using a single locus approach is the requirement for very large study populations to generate sufficient statistical power. Even with a combined dataset, loci with small effect sizes may be extremely difficult to detect.

7.2.1. Gene-Set Analysis

It may be possible to increase the power of genetic studies through gene-set analyses such as pathway and network-based analyses. In a pathway analysis, SNP data are combined in a biologically meaningful way making the interpretation of results less complicated. Such analyses combine SNPs of weak effects that would typically be undetectable by single-SNP tests. Moreover, such analyses that combine SNP data may decrease the number of tests carried out depending on the specific method used and therefore, require less stringent statistical correction measures (Fridley & Biernacka 2011; Khatri *et al.* 2012). Many network analyses also make use of gene co-expression data such as weighted gene co-expression network analysis (Langfelder & Horvath 2008). For instance, a network analysis carried out with the Ingenuity Pathway Analysis tool involving mouse model of MS, experimental autoimmune encephalomyelitis, identified protein networks mainly regulated by two common proteins (post-synaptic density protein 95 (DGL4) and calcium-activated potassium channel alpha 1 (KCNMA1)). Blocking of KCNMA1 in macrophages altered myelin phagocytosis that is important for disease pathology (Vanheel *et al.* 2012).

Some of these analyses employ statistical packages, such as R, for network analysis that can also be used for characterising the correlation structure between genetic marker data. For example, there are studies that make use of previous SNP genotyping data, such as WTCCC1 GWAS results, to construct a correlation structure between genes, which form gene clusters functioning in specific biological pathways that may play a role in disease susceptibility (Torkamani *et al.* 2008). For instance, in the study by Torkamani *et al.* (2008),

dopamine signaling was significantly associated with hypertension. However, this association was missed in the previous WTCCC1 GWAS (WTCCC1 2007). Considering the complex polygenic nature of AS, such multi-locus approaches to data analysis may reveal how genetic variations affect disease susceptibility. As AS, IBD and psoriasis all share some genetic background, the pathway analysis approach may also provide correlation information between genetic basis of different diseases (Torkamani *et al.* 2008; Fridley & Biernacka 2011). Likewise, it may be possible to identify gene of small effect by combining sample populations with AS, IB and psoriasis to increase statistical power of such a pathway study.

There are two different approaches to pathway analysis; 1) Hypothesis-driven candidate pathway analysis, 2) Data-driven genomewide pathway analysis. The main difference between these two approaches is the use of prior knowledge in pathway selection in candidate pathway analyses. Freely available online tools such as GenGen, GSA-SNP and GRAIL that employ different strategies including enrichment and text-mining, together with published protocols can be used to carry out pathway analysis of the data (Fridley & Biernacka 2011; Ramanan *et al.* 2012). Evident candidate pathways for such analysis are the immune-related pathways such as T_H17 pathway where AS-associated genes *IL23R*, *IL12B*, *STAT3* and *JAK2* play important roles (See Chapter 3, Table 3.4). However, application of the pathway analysis approach on the data generated by GWAS and imputation is potentially a more efficient way to determine the association of AS with various pathways.

7.2.2. Functional Studies

Even with the use of large sample sizes and sensitive statistical methods, confirmation of the candidacy of a particular gene may depend on functional studies. Functional studies are essential to confirm observed genetic associations and give biological interpretations to these associations prior to possible therapeutic targeting. For example, the study by Bowness *et al.* (2011) demonstrated a functional link between the expression of HLA-B27 homodimers (B27₂) on antigen presenting cells (APCs) and IL23R expressing T cells (Bowness *et al.* 2011). This study of B27₂ and T_H17 T cells provided evidence for the inflammatory role of T_H17 T cells in AS. B27₂ expressed on APCs bind to killer cell immunoglobulin-like receptor 3DL2 (KIR3DL2) expressed on CD4⁺ T cells. Even though only 15% of the CD4⁺ T cells expressed KIR3DL2 in peripheral blood of AS patients, these cells account for 70% of the increase in T_H17 T cell numbers seen in AS patients. On stimulation, these T cells also exhibited a 4-fold increase in IL17 production, the main pro-inflammatory cytokine of the T_H17 T cells, compared to the controls. Moreover, KIR3DL2⁺ T cells were responsible for most of IL23R expression, and produced more IL17 in the presence of IL23, the cytokine necessary for expansion and maintenance of T_H17 T cells. Thus, the KIR3DL2-B27₂ interaction is likely to lead to a pro-inflammatory outcome by inhibiting T cell death and promoting cytokine production (Figure 7.1).

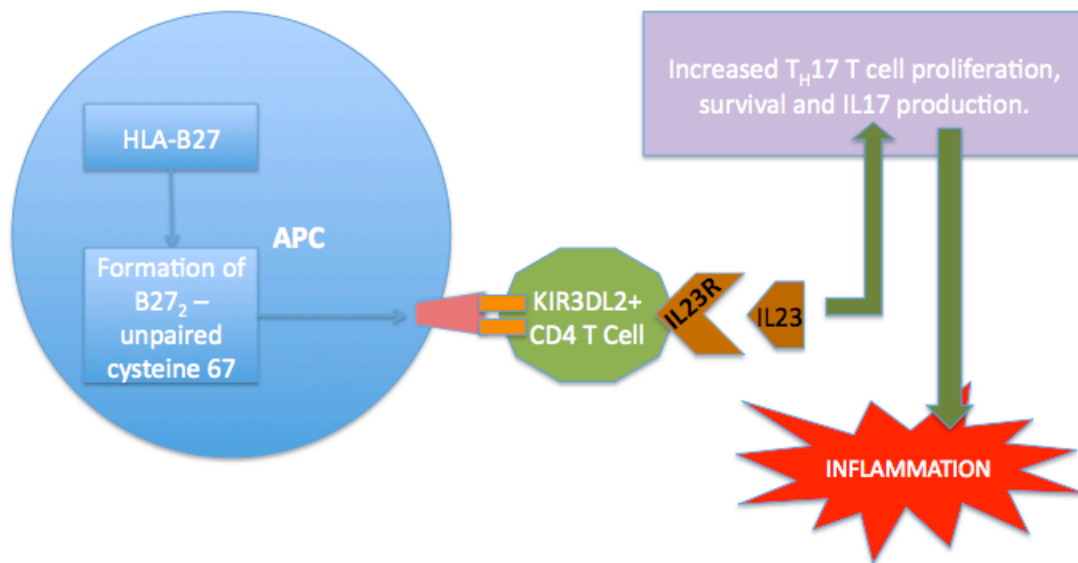


Figure 7.1. KIR3DL2+ CD4 T cells may be responsible for the increased T_H17 response and therefore, inflammation leading to AS.

Expression quantitative trait loci (eQTL) analysis is also useful for identifying the effects of SNPs associated with diseases in genetic studies i.e. translating genotype into phenotype. In general, it is a complicated task to determine the actual causal SNP in a genetic association study. Even though association signals with significantly low p-values are detected, linkage disequilibrium across gene regions makes it difficult to discover a truly causal variant. Since some of the causal variants affect gene expression levels, it is possible to use eQTL analysis, which employs SNP genotype data and gene expression (mRNA level) data, to detect variants that affect gene expression. For example, a genetic study of celiac disease identified a strong association with an *IL18RAP* variant (rs917997), which was also shown to correlate with the expression of *IL18RAP* in the eQTL analysis (AA homozygotes showed the lowest level of expression, whereas GG

homozygotes showed the highest level of expression of the IL18 receptor (IL18RAP)) (Hunt *et al.* 2008). A gene expression profiling has already been done in AS using mRNA expression levels in peripheral blood mononuclear cells from 18 cases and 18 (age-matched and gender-matched) controls (Duan *et al.* 2010). This study showed downregulation of immune-related genes, *NR4A2*, tumour necrosis factor AIP3 (*TNFAIP3*) and *CD69*. Future eQTL analysis could also be informative in determining genotype-phenotype correlations and thus, identifying variants affecting gene expression in AS. Results from such an eQTL study could complement the associations observed in the previous AS GWAS (Burton *et al.* 2007; Reveille *et al.* 2010; Evans *et al.* 2011).

7.2.3. DNA Sequencing

The 1000 Genomes Exon Pilot Project, involving the sequencing of ~1000 human genes in ~700 samples, contributed to the development of new technology for cost-effective DNA sequencing, which is now in general use and is subject to further development. As a result of this pilot project, robust protocols were developed for the 1000 Genomes Sequencing Project, which will inform the development of strategies for studying rare and low frequency variants. Based on 12,758 exonic SNPs discovered in 7 populations (70% novel, 74% with MAF<1%), it was found that most functional (i.e. causing amino acid change) variants are rare (MAF<1%) and are much less likely to be shared between different populations than rare intergenic variants. This reduction in variant sharing disappears when comparing exonic variants of relatively higher frequency (MAF>1%) with intergenic variants (Marth *et al.* 2011).

Second generation DNA sequencing involves the rapid generation of sequence data from hundreds of thousands of short DNA fragments in parallel. Exome sequencing of an individual generally generates 15,000-25,000 variants. These variants are then filtered and reduced by comparison to known variants in public databases and according to their functional importance (i.e. synonymous vs. non-synonymous amino acid changes) by using programs such as SIFT, PolyPhen and VarioWatch (Jaffe *et al.* 2011; Singleton 2011). Exome sequencing is now considered an efficient method for discovery of rare causal variants. It has been successful in identifying causal variants in rare and familial forms of diseases such as Parkinson's disease (Vilarino-Guell *et al.* 2011; Zimprich *et al.* 2011). A similar strategy is also being used in the identification of non-inherited/de-novo mutations in diseases such as the Proteus syndrome (Lindhurst *et al.* 2011) and mental retardation disorders (O'Roak *et al.* 2011).

Whole-genome sequencing, which involves sequencing of exons, introns, intergenic regions as well as large gene deserts that may contain variants affecting transcription and translation, is still too costly for study of complex traits that require large populations (i.e. thousands of individuals). However, low coverage sequencing (e.g. 2X) of a sample population and then imputation of these genotypes in a larger population is an appealing strategy. When deep sequencing (16X) of 60 individuals was compared with low coverage (2X) sequencing of 400 individuals, it was observed that low coverage sequencing provided more variant discovery power as well as information content after imputation. Fourteen per cent more SNPs ($1\% < \text{MAF} < 2\%$) were imputed (imputation accuracy $\text{RSQR} > 0.3$) when low coverage haplotypes from 400

individuals were used as a reference panel, compared to using the deep sequencing (16X) of 60 individuals as a reference panel. Therefore, there is a case for the use of low coverage sequencing data for imputation in association studies of complex diseases such as AS and it is a cost-effective method of generating dense genotype data for SNPs of low frequency. However, with further improvements in sequencing technologies both in terms of accuracy and read lengths to reduce errors, it will also be possible to design association studies based on sequencing data (Li *et al.* 2011).

In the past 5 years of GWAS era, many genetic associations (>1000) have been identified in complex diseases such as AS, MS, PS and IBD (UC/CD). However, it is somewhat sobering that a substantial proportion (~60%) of the heritability of these diseases remains to be explained and has led to differences of opinion on the success of the GWAS design to deliver disease-associated genetic variants (Visscher *et al.* 2012). For some diseases such as psoriasis and AS the identification of an associated gene (*IL23R*) has lead to a new therapeutics, such as treatment with anti-IL17 (secukinumab) and anti-IL12Rp40 (ustekinumab). However, translation of GWAS results into therapy will not generally be simple (Manolio *et al.* 2009). Common variants (MAF>5%) do not explain all the heritability and common disease-common variant (CD-CV) hypothesis used in agnostic GWAS studies is an incomplete explanation for these diseases. It ignores the effects of rare variants that may collectively play a major role in complex diseases. Furthermore, biological interpretation of genetic associations observed by statistical analyses is an intricate task requiring functional studies as already been discussed earlier in this chapter. For the AS-associated gene,

ERAP1, this work is already under way with the resolution of the ERAP1 crystal structure and studies on the functional effects of the associated variants (Kochan *et al.* 2011). For other associations such as the intergenic regions at 2p15 and 21q22, establishing a pathological role is far more challenging. However, the genetic studies in AS discussed in this thesis, provide substantial improvements in our understanding of the aetiology of AS as a direct result of the GWAS data. Future technological advances and use of sequencing data, which will lead to inclusion of both common and rare/low frequency variants in genetic association studies, is likely to identify new variants and pathways as therapeutic targets, while further advancing the understanding of AS aetiology.

7.3. References

- Birtley, J. R., Saridakis, E., *et al.* (2012). "The crystal structure of human endoplasmic reticulum aminopeptidase 2 reveals the atomic basis for distinct roles in antigen processing." *Biochemistry* **51**(1): 286-295.
- Bowness, P., Ridley, A., *et al.* (2011). "Th17 cells expressing KIR3DL2+ and responsive to HLA-B27 homodimers are increased in ankylosing spondylitis." *J Immunol* **186**(4): 2672-2680.
- Burton, P. R., Clayton, D. G., *et al.* (2007). "Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants." *Nat Genet* **39**(11): 1329-1337.
- Cargill, M., Schrodi, S. J., *et al.* (2007). "A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes." *Am J Hum Genet* **80**(2): 273-290.
- Davidson, S. I., Liu, Y., *et al.* (2011). "Association of STAT3 and TNFRSF1A with ankylosing spondylitis in Han Chinese." *Ann Rheum Dis* **70**(2): 289-292.
- Davis, J. C., Jr., van der Heijde, D. M., *et al.* (2008). "Efficacy and safety of up to 192 weeks of etanercept therapy in patients with ankylosing spondylitis." *Ann Rheum Dis* **67**(3): 346-352.
- Duan, R., Leo, P., *et al.* (2010). "Gene expression profiling reveals a downregulation in immune-associated genes in patients with AS." *Ann Rheum Dis* **69**(9): 1724-1729.
- Duerr, R. H., Taylor, K. D., *et al.* (2006). "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene." *Science* **314**(5804): 1461-1463.
- Evans, D. M., Spencer, C. C., *et al.* (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**(8): 761-767.
- Fridley, B. L. & Biernacka, J. M. (2011). "Gene set analysis of SNP data: benefits, challenges, and future directions." *Eur J Hum Genet* **19**(8): 837-843.
- Genomes (2010). "A map of human genome variation from population-scale sequencing." *Nature* **467**(7319): 1061-1073.
- Hunt, K. A., Zhernakova, A., *et al.* (2008). "Newly identified genetic risk variants for celiac disease related to the immune response." *Nat Genet* **40**(4): 395-402.
- Jaffe, A., Wojcik, G., *et al.* (2011). "Identification of functional genetic variation in exome sequence analysis." *BMC Proc* **5 Suppl 9**: S13.
- Jostins, L., Morley, K. I., *et al.* (2011). "Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets." *Eur J Hum Genet* **19**(6): 662-666.
- Khatri, P., Sirota, M., *et al.* (2012). "Ten years of pathway analysis: current approaches and outstanding challenges." *PLoS Comput Biol* **8**(2): e1002375.
- Kochan, G., Krojer, T., *et al.* (2011). "Crystal structures of the endoplasmic reticulum aminopeptidase-1 (ERAP1) reveal the molecular basis for N-terminal peptide trimming." *Proc Natl Acad Sci U S A* **108**(19): 7745-7750.
- Langfelder, P. & Horvath, S. (2008). "WGCNA: an R package for weighted correlation network analysis." *BMC Bioinformatics* **9**: 559.

- Li, Y., Sidore, C., *et al.* (2011). "Low-coverage sequencing: implications for design of complex trait association studies." *Genome Res* **21**(6): 940-951.
- Li, Y., Willer, C., *et al.* (2009). "Genotype imputation." *Annu Rev Genomics Hum Genet* **10**: 387-406.
- Lindhurst, M. J., Sapp, J. C., *et al.* (2011). "A mosaic activating mutation in AKT1 associated with the Proteus syndrome." *N Engl J Med* **365**(7): 611-619.
- Manolio, T. A., Collins, F. S., *et al.* (2009). "Finding the missing heritability of complex diseases." *Nature* **461**(7265): 747-753.
- Marth, G. T., Yu, F., *et al.* (2011). "The functional spectrum of low-frequency coding variation." *Genome Biol* **12**(9): R84.
- Mei, Y., Pan, F., *et al.* (2011). "Increased serum IL-17 and IL-23 in the patient with ankylosing spondylitis." *Clin Rheumatol* **30**(2): 269-273.
- Morris, A. P. & Zeggini, E. (2010). "An evaluation of statistical approaches to rare variant analysis in genetic association studies." *Genet Epidemiol* **34**(2): 188-193.
- Neale, B. M., Rivas, M. A., *et al.* (2011). "Testing for an unusual distribution of rare variants." *PLoS Genet* **7**(3): e1001322.
- Nguyen, T. T., Chang, S. C., *et al.* (2011). "Structural basis for antigenic peptide precursor processing by the endoplasmic reticulum aminopeptidase ERAP1." *Nat Struct Mol Biol* **18**(5): 604-613.
- O'Roak, B. J., Deriziotis, P., *et al.* (2011). "Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations." *Nat Genet* **43**(6): 585-589.
- Ramanan, V. K., Shen, L., *et al.* (2012). "Pathway analysis of genomic data: concepts, methods and prospects for future development." *Trends Genet.*
- Reveille, J. D., Sims, A. M., *et al.* (2010). "Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci." *Nat Genet* **42**(2): 123-127.
- Singleton, A. B. (2011). "Exome sequencing: a transformative technology." *Lancet Neurol* **10**(10): 942-946.
- Torkamani, A., Topol, E. J., *et al.* (2008). "Pathway analysis of seven common diseases assessed by genome-wide association." *Genomics* **92**(5): 265-272.
- Toussiot, E. (2012). "The IL23/Th17 Pathway as a Therapeutic Target in Chronic Inflammatory Diseases." *Inflamm Allergy Drug Targets* **11**(2): 159-168.
- Vanheel, A., Daniels, R., *et al.* (2012). "Identification of protein networks involved in the disease course of experimental autoimmune encephalomyelitis, an animal model of multiple sclerosis." *PLoS One* **7**(4): e35544.
- Vilarino-Guell, C., Wider, C., *et al.* (2011). "VPS35 mutations in Parkinson disease." *Am J Hum Genet* **89**(1): 162-167.
- Visscher, P. M., Brown, M. A., *et al.* (2012). "Five years of GWAS discovery." *Am J Hum Genet* **90**(1): 7-24.
- Willer, C. J., Li, Y., *et al.* (2010). "METAL: fast and efficient meta-analysis of genomewide association scans." *Bioinformatics* **26**(17): 2190-2191.
- WTCCC1 (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* **447**(7145): 661-678.

Zimprich, A., Benet-Pages, A., *et al.* (2011). "A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease." Am J Hum Genet **89**(1): 168-175.

APPENDICES

Appendix 1

Possible novel SNPs 2, 3, 4, 5 and 9 detected by polyphred were excluded from RFLP analysis following visual inspection of the electropherograms as well as because of their lower polyphred probability scores (score<90).

Electropherograms of these and other detected known SNPs are provided below in Figures 4.5-4.7.

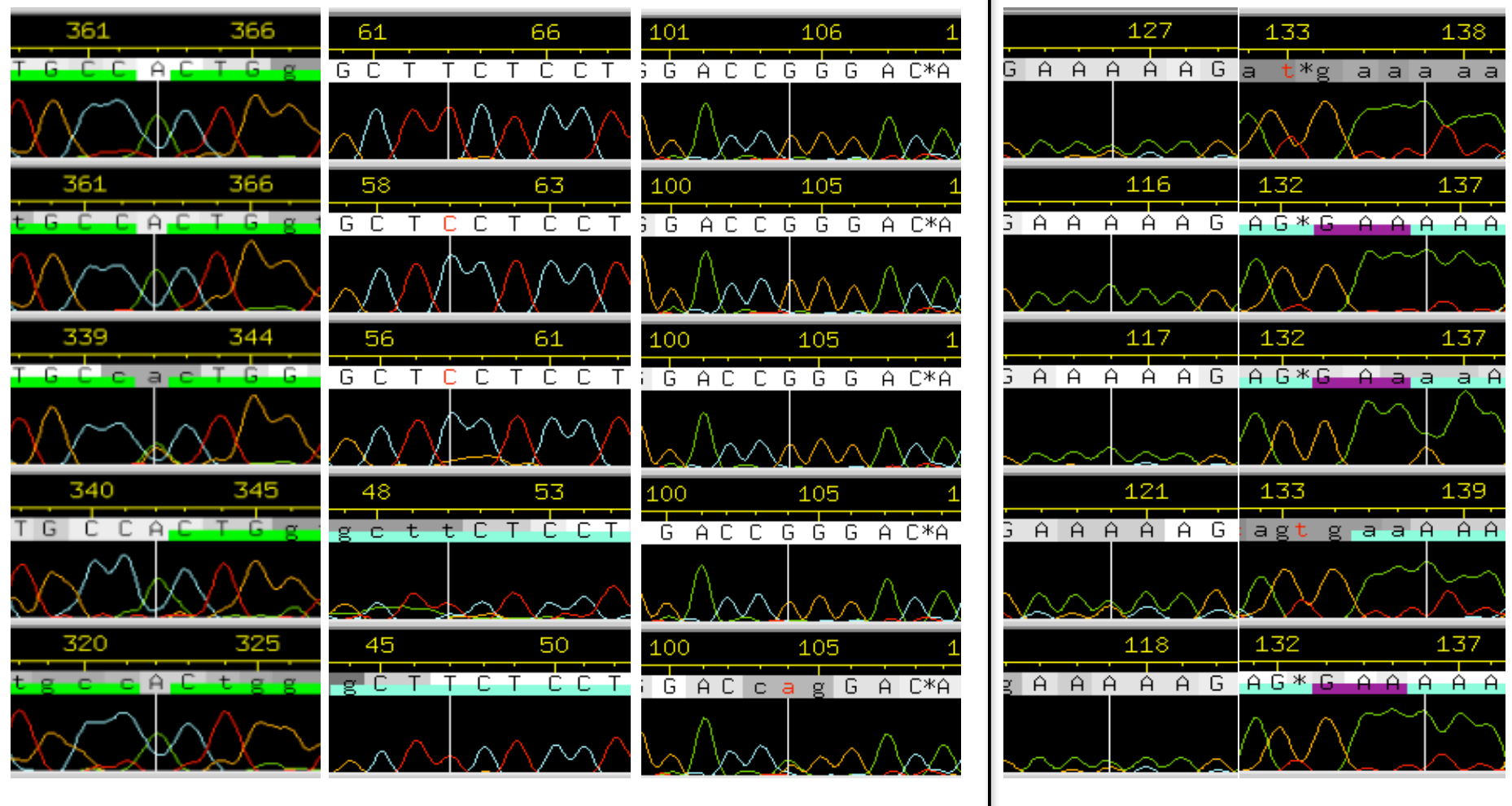


Figure 4.5. (A) rs767455 (A/G) (B) rs1800692 (C/T) (C) rs4149584 (A/G),

(D) rs12426675 (A/G)

(D) Novel 6 (score=79).

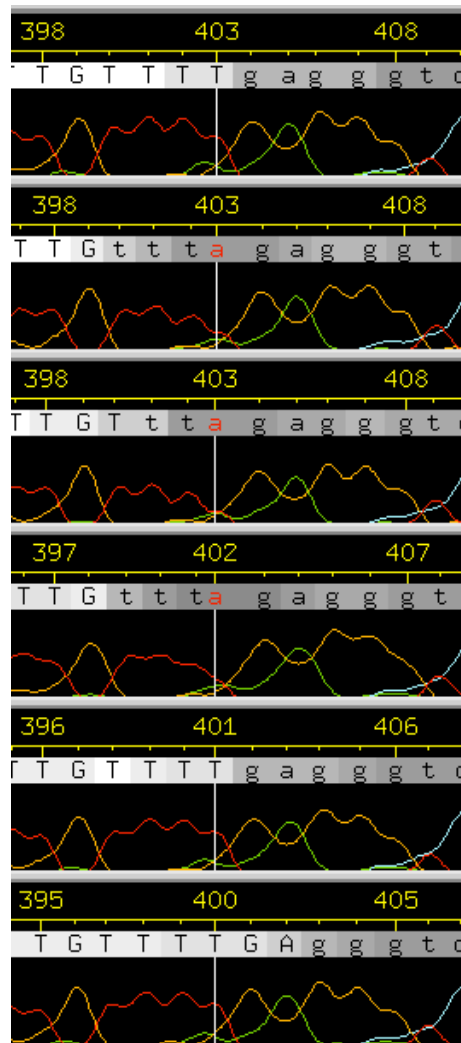
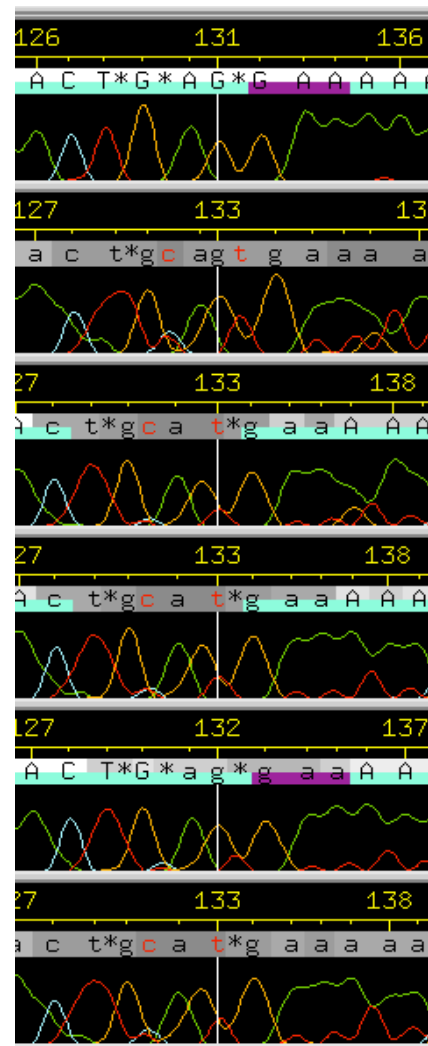
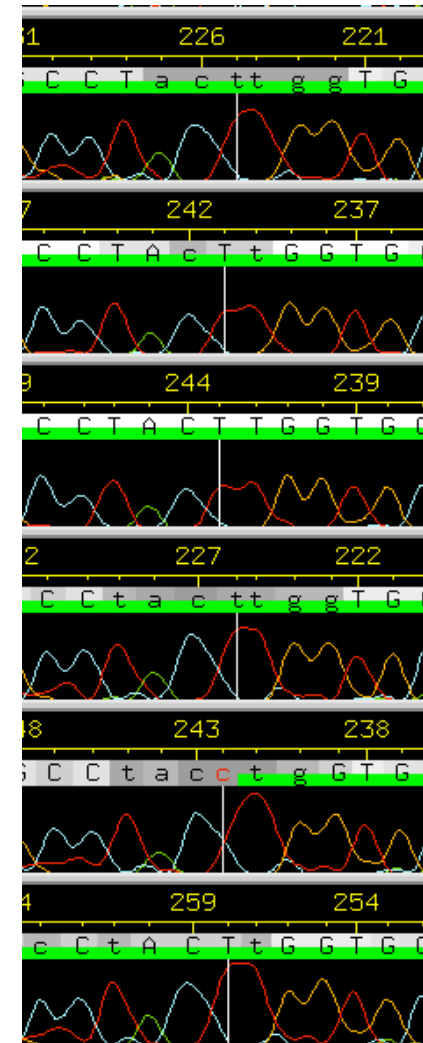


Figure 4.7. (A) Novel 1 (score>90),

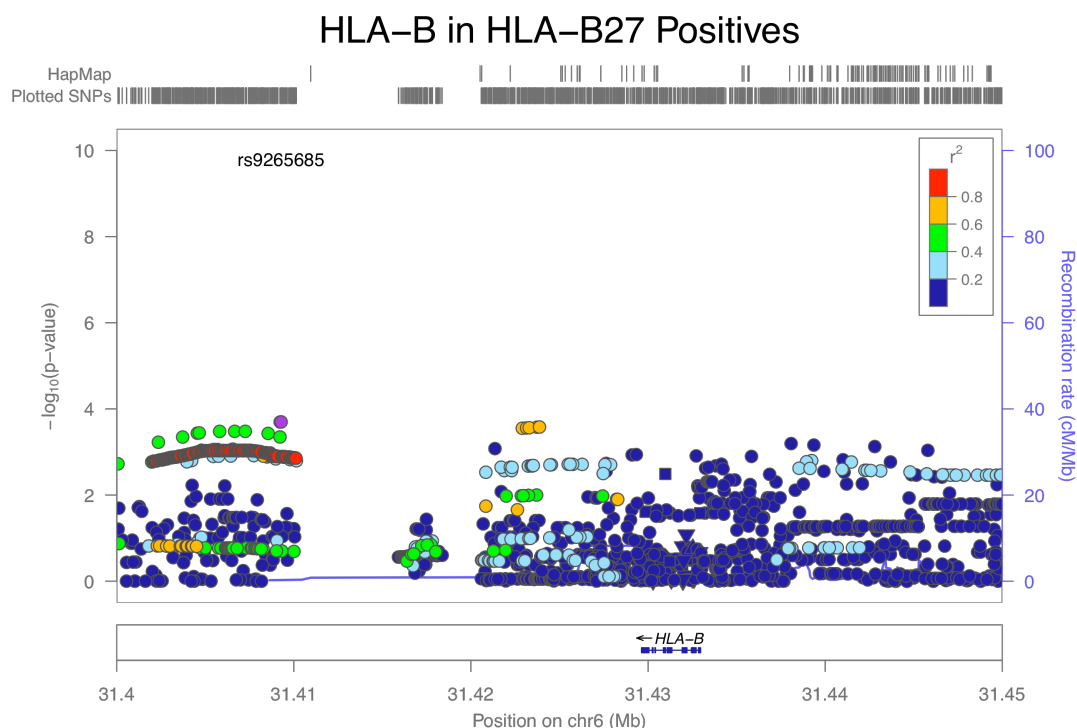


(B) Novel 7 (score>90),



(C) Novel 8 (score>90).

Appendix 2



Plot 5.24. Residual association in the *HLA-B* region in HLA-B27 positive individuals (as defined by rs4349859). The strong association observed in HLA-B27 positive individuals is absent in HLA-B27 negative individuals. The strongest association was observed with the SNP rs9265685 (Purple). The strength of LD relative to the top SNP (Purple) is shown with different colours ($r^2 > 0.8$ – red, $0.8 > r^2 > 0.6$ – orange, $0.6 > r^2 > 0.4$ – green, $0.4 > r^2 > 0.2$ – light blue, $r^2 < 0.2$ – dark blue). (Triangle: frame/splice variant; inverted triangle: non-synonymous variant; square: untranslated region/synonymous variant)

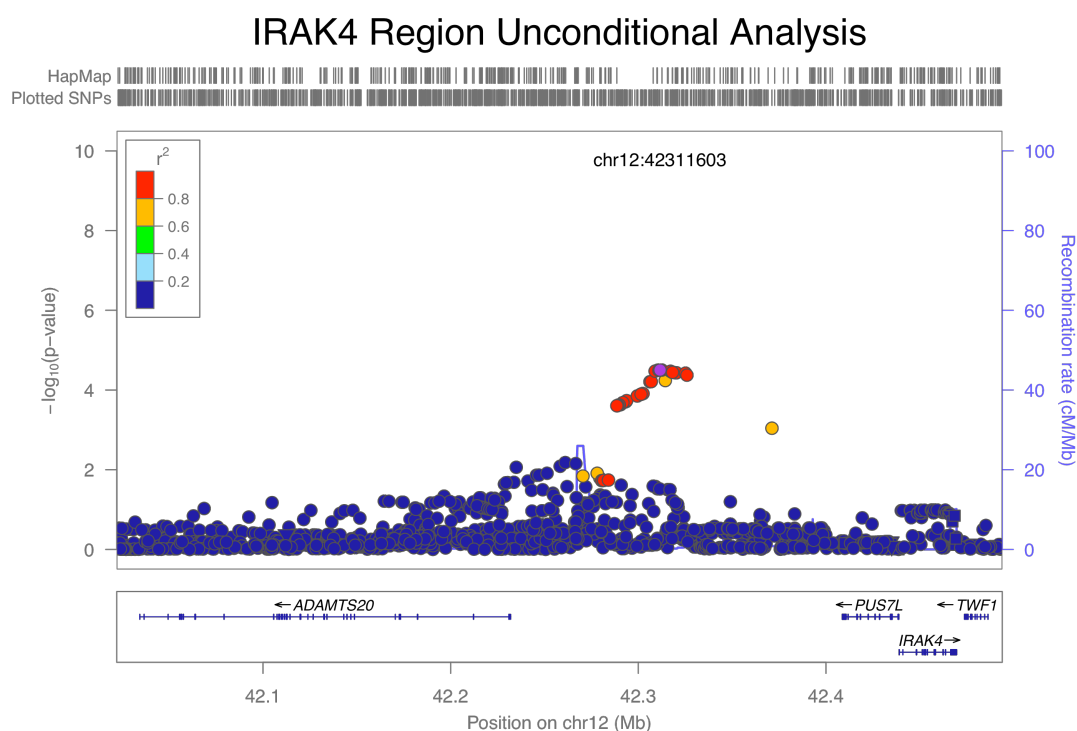
CHR	GENE	MARKER COUNT	SAMPLE COUNT	RARE VARIANT COUNT	BETA	P-VALUE
1	<i>SLC35D1</i>	13	6949	4635	0.796474	0.000173353
1	<i>IL12RB2</i>	39	6949	12707	-0.846924	0.0044853
3	<i>RHOA</i>	15	6949	2388	-1.68455	0.000629735
5	<i>KCNMB1</i>	3	6910	323	-0.752361	0.000824054
6	<i>HLA-B</i>	13	6949	4015	-2.59	4.4 x 10 ⁻¹⁴
7	<i>ZNF786</i>	5	6949	659	0.9519	0.000456796
7	<i>IL6</i>	2	6949	1229	-0.270398	0.0081329
10	<i>C10orf67</i>	12	6949	3486	0.75591	0.000303745
10	<i>LOC729398</i>	2	6949	398	0.599986	0.000335085
10	<i>C10orf115</i>	15	6949	6670	0.457299	0.000630667
13	<i>TNFSF13B</i>	5	6949	656	-0.837734	0.00675807
14	<i>SOCS4</i>	2	6949	77	0.845779	0.000221605
16	<i>BTBD12</i>	8	6949	1376	0.715129	3.5 x 10 ⁻⁵
16	<i>NOD3</i>	14	6949	3025	0.810292	0.0243776
17	<i>C17orf46</i>	2	6949	107	-0.759918	0.000321262
18	<i>TNFRSF11A</i>	23	6949	7444	-0.859322	0.0314204
18	<i>PTPN2</i>	18	6949	4703	0.837856	0.0431238
19	<i>SPIB</i>	3	6949	51	1.14371	0.000356989
19	<i>RGL3</i>	4	6949	474	0.828338	0.000965077

Table 5.16. A more complete list of the results of the GRANVIL analysis. (CHR: Chromosome; Marker count: Number of rare markers in gene region; Sample count: Number of samples in analysis; Rare variant count: Count of rare alleles found in individuals; Beta: Effect size)

Appendix 3

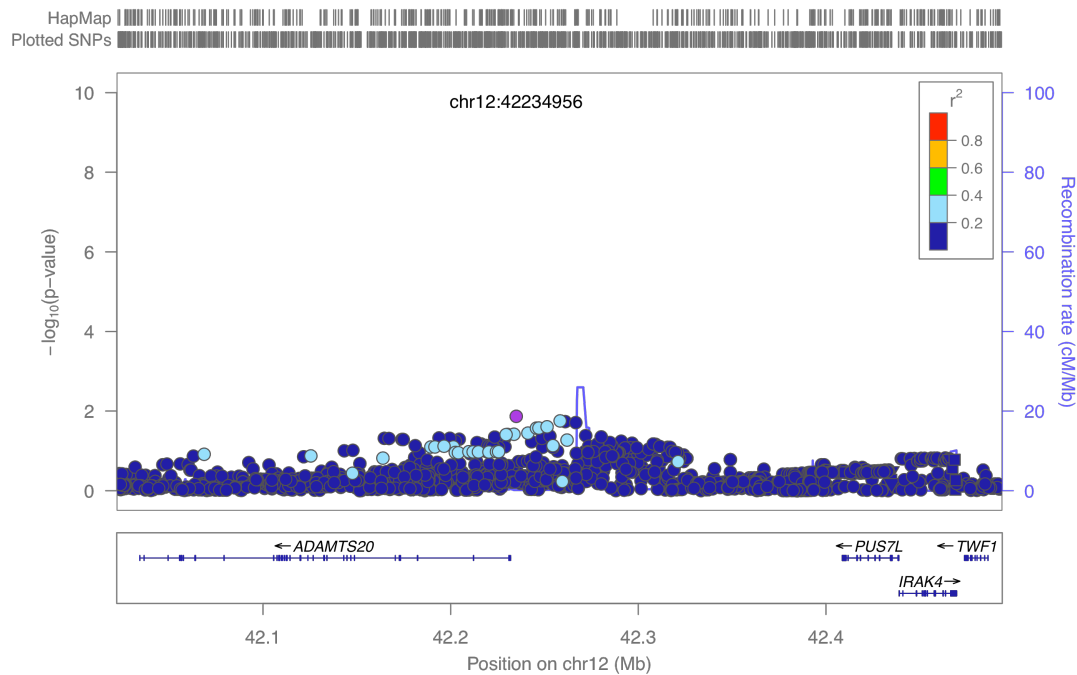
Chromosome 12 – *IRAK4*

Interleukin-1-receptor (IL1R)-associated kinases (IRAK) are important mediators in the signal transduction of Toll-like receptor (TLR) and IL1R family members. IRAK4 functions in this signal transduction pathway (Li et al. 2002). The IL1 gene complex, encoding both IL1R and its ligands, has been associated with AS in previous studies in various ethnicities such as British, Canadian, Taiwanese Chinese and Han Chinese (Chou et al. 2006; Maksymowych et al. 2006; Guo et al. 2010; Reveille et al. 2010).



Plot 6.6. *IRAK4* region unconditional analysis – showing the top hit SNP (rs77143393, purple) in the analysis of imputed genotypes, and the LD pattern.

IRAK4 Region rs77143393 Conditional Analysis

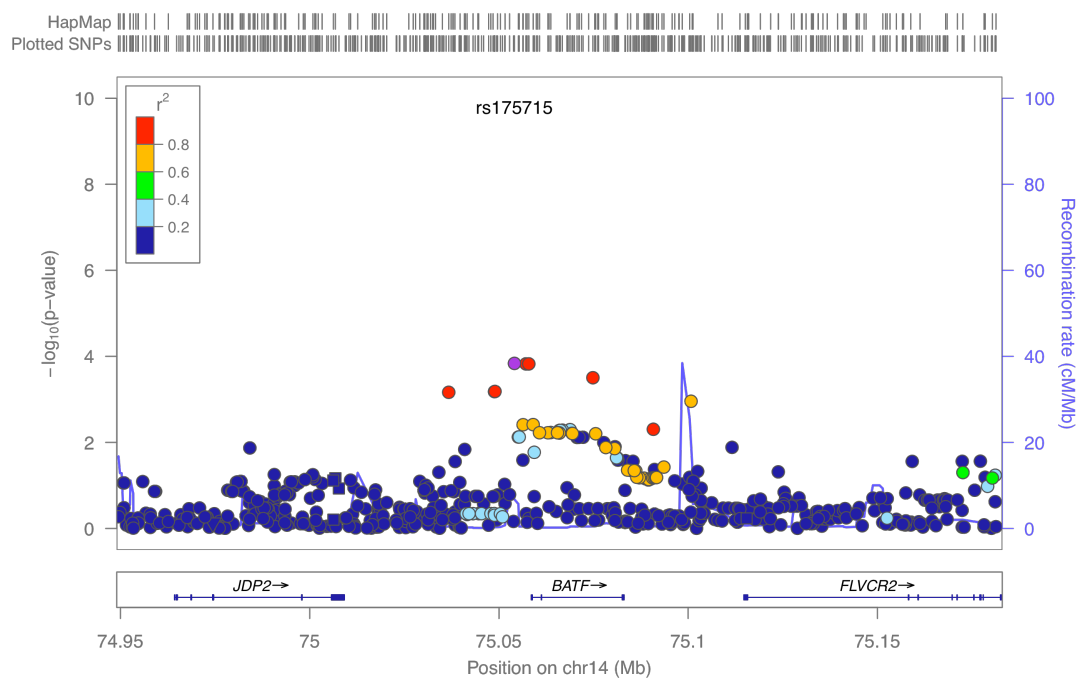


Plot 6.7. *IRAK4* region rs77143393 conditional analysis using imputed data only – showing that rs77143393 association was the only independent association in the region.

Chromosome 14 – *BATF*

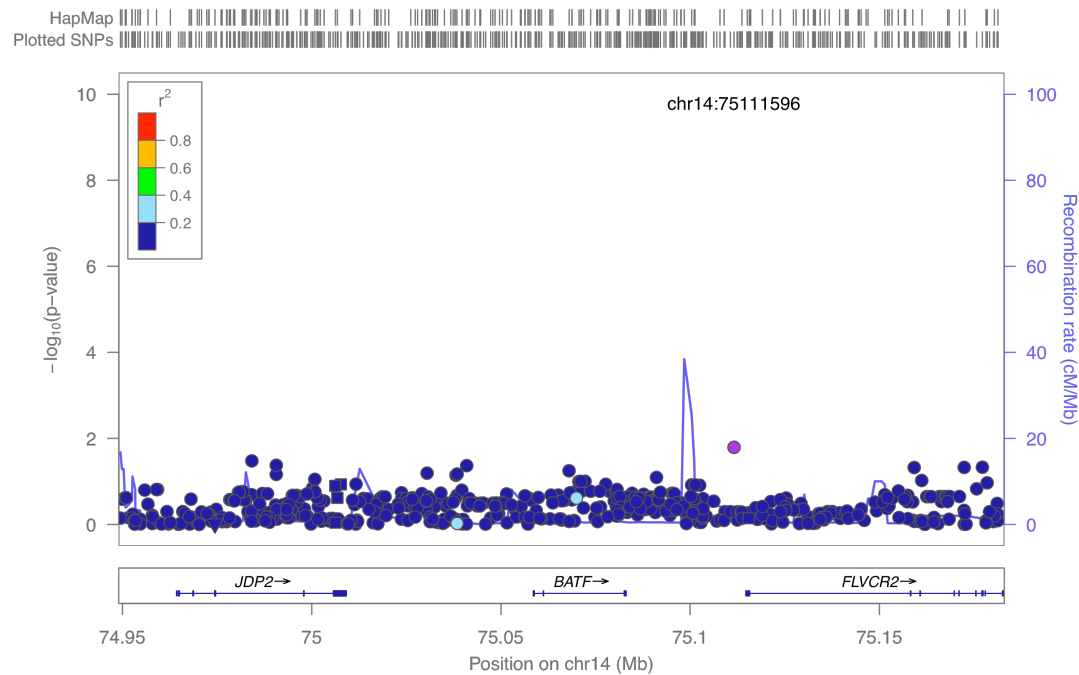
This gene encodes BATF, which forms heterodimers with Activating Protein 1 (AP-1) inhibiting its transcription factor function. BATF has been shown to play an important role in IL-17-producing T cell (T_H17) differentiation. *Batf*^{-/-} mice show normal T_H1 and T_H2 differentiation but decreased IL17 production under T_H17 conditions (Schraml *et al.* 2009) Evidence from genetic and functional studies has provided support for the role of T_H17 lymphocytes in inflammatory diseases such as IBD and AS. IL23R - essential for maintenance of T_H17 pathway – has been associated with AS previously (Burton *et al.* 2007; Karaderi *et al.* 2009).

BATF Region – Imputed based on 1000G August 2009 Pilot Data



Plot 6.8. *BATF* region unconditional analysis – showing the top hit SNP (rs175715, purple) in the analysis of imputed genotypes, and the LD pattern.

BATF Region rs175721 Conditional Analysis



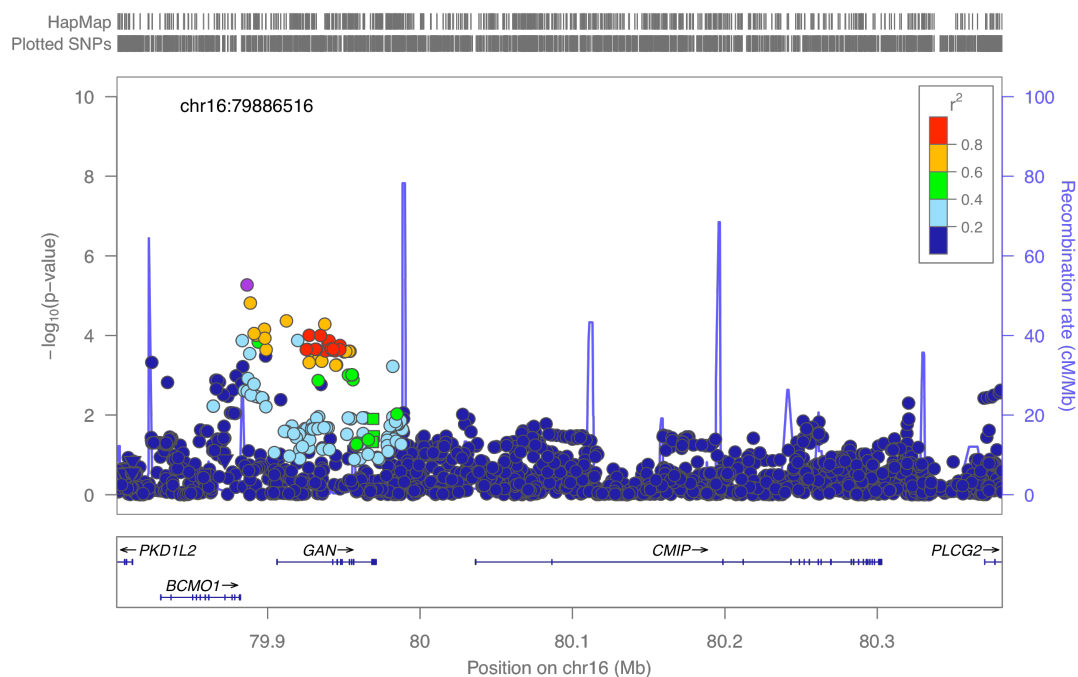
Plot 6.9. *BATF* region rs175721 conditional analysis using imputed data only – showing that rs175721 association was the only independent association in the region (rs175715-rs175721 $r^2 > 0.8$).

Chromosome 16 – *GAN* – Upstream *CMIP*

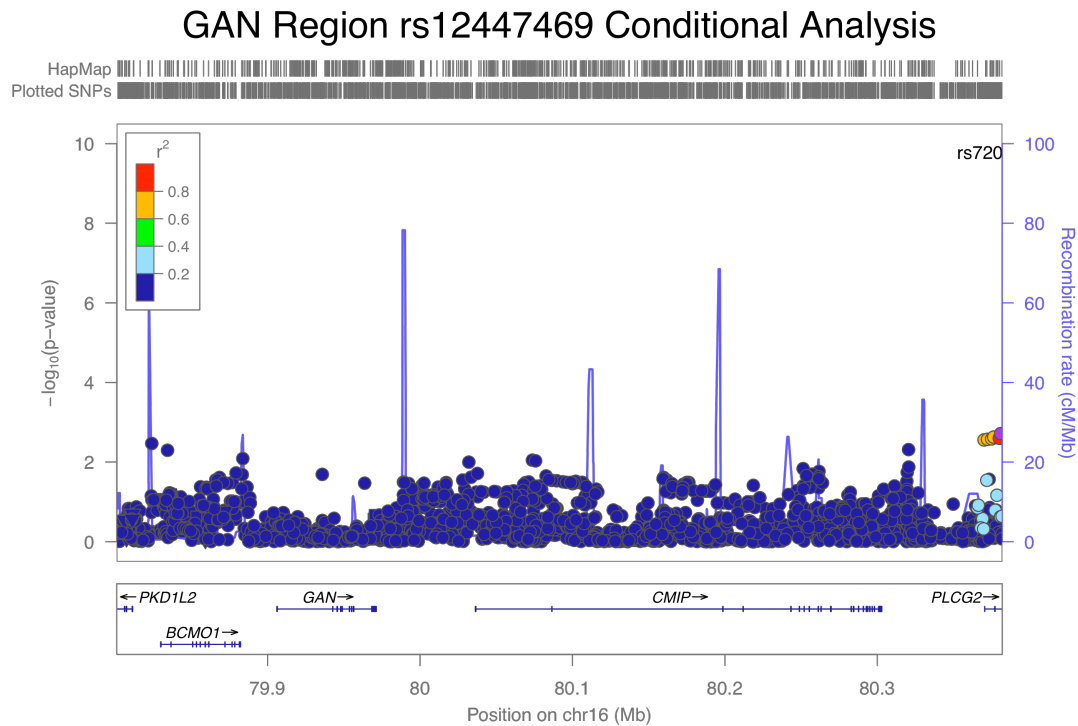
GAN encodes a protein that potentially plays a role in autophagy by interacting with ATG16L1 in the ubiquitin-like (UBL) transfer cascade network responsible for vesicle elongation and autophagosome assembly (Behrends *et al.* 2010). A non-synonymous SNP, rs2242880, in *ATG16L1* has been associated with Crohn's disease in a genomewide association study (UK replication; $p < 0.0004$, OR~1.35) (Hampe *et al.* 2007). Autophagy is important for protein/organelle degradation and has been implicated in pathogenesis of human diseases such as cancer, neurodegenerative diseases (e.g. Alzheimer's disease and Huntington's disease),

as well as in innate and adaptive immunity (Shintani & Klionsky 2004; Kundu & Thompson 2008).

***GAN-CMIP* Region Unconditional Analysis**



Plot 6.10. *GAN-CMIP* region unconditional analysis – showing the top hit SNP (rs58532255, purple) in the analysis of imputed genotypes, and the LD pattern.

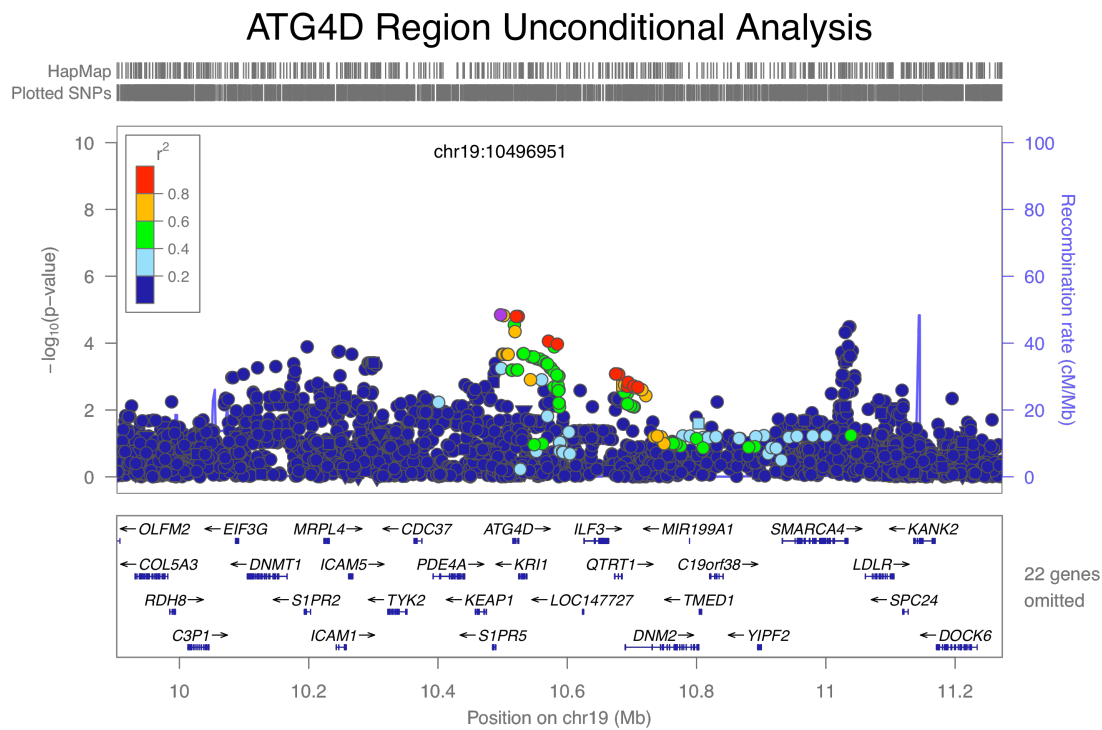


Plot 6.11. *GAN-CMIP* region rs12447469 conditional analysis using imputed data only – showing that rs12447469 association was the only independent association in the region (rs58532255-rs12447469 $r^2 > 0.6$).

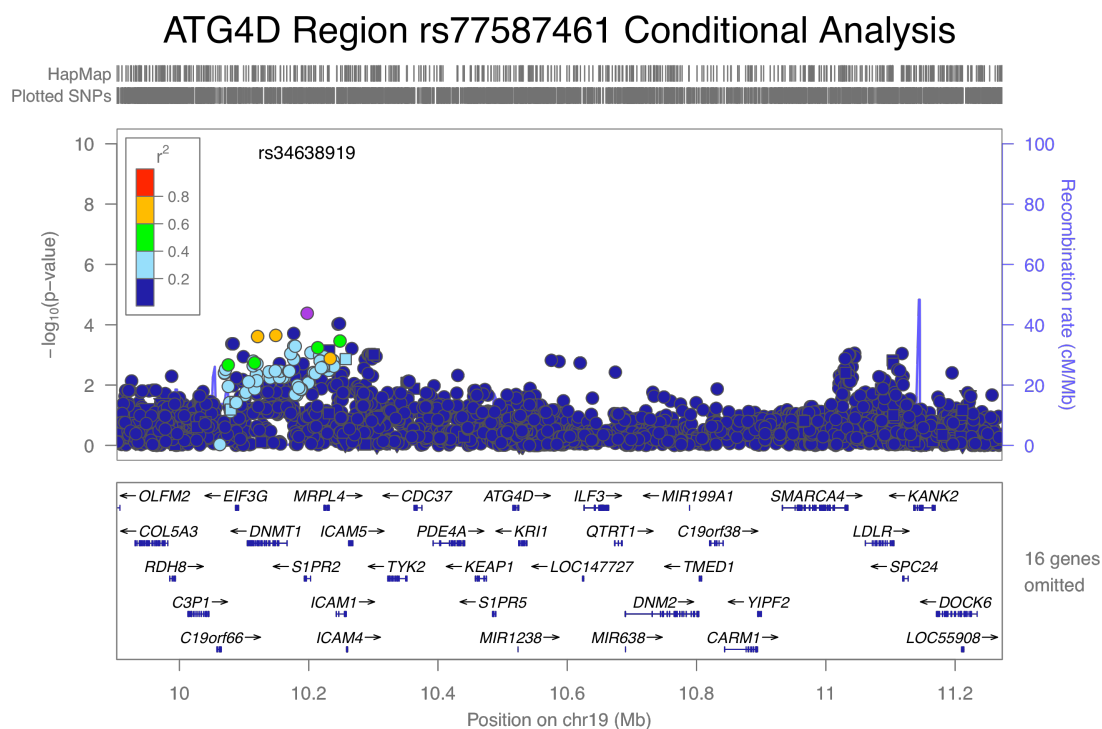
Chromosome 19 – *ATG4D*

ATG4D encodes a cysteine protease that plays an important role in autophagy (Scherz-Shouval *et al.* 2007). Autophagy is the only process by which whole organelles can be recycled. It is essential for cell survival i.e. cellular homeostasis to maintain internal stability via adjusting physiological processes, as well as cell death (type II (autophagic) programmed cell death) (Shintani & Klionsky 2004). It has been hypothesised that *ATG4D* is an essential cell survival factor, and caspase-3 cleaves *ATG4D* under apoptotic conditions allowing it to act as a

regulator of autophagy coupling mitochondrial dysfunction to apoptosis (Betin & Lane 2009).



Plot 6.12. *ATG4D* region unconditional analysis – showing the top hit SNP (rs77587461, purple) in the analysis of imputed genotypes, and the LD pattern.



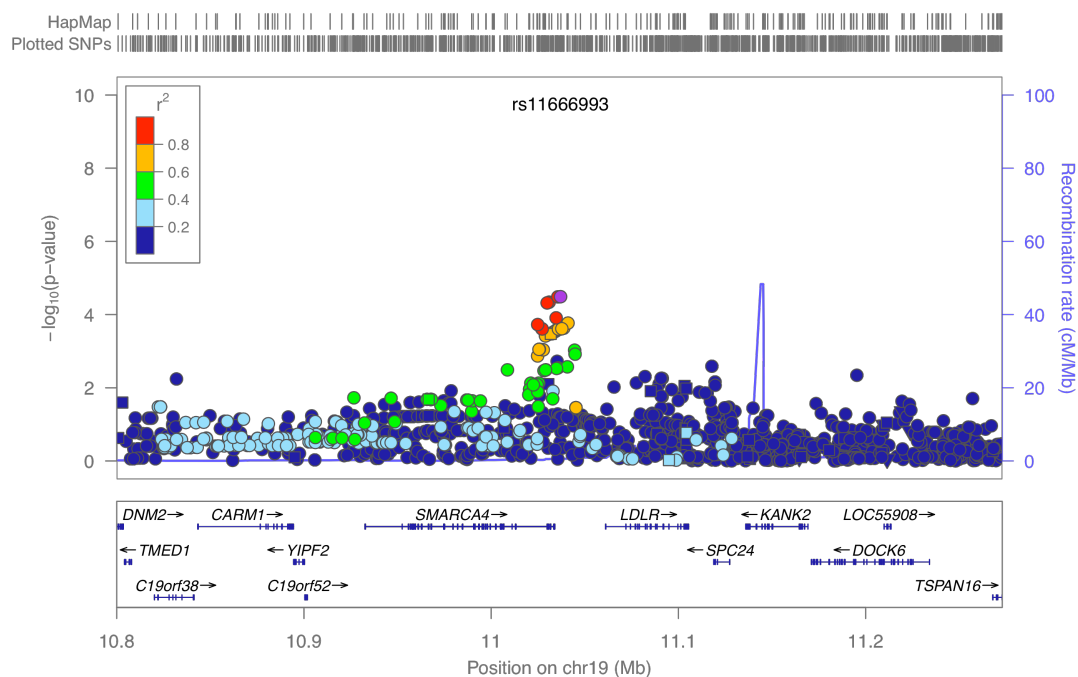
Plot 6.13. *ATG4D* region rs77587461 conditional analysis using imputed data only – showing that rs77587461 association was the only independent association in the *ATG4D* region. The association signal with rs34638919 is potentially an independent association in the distant upstream region.

Chromosome 19 – *SMARCA4*

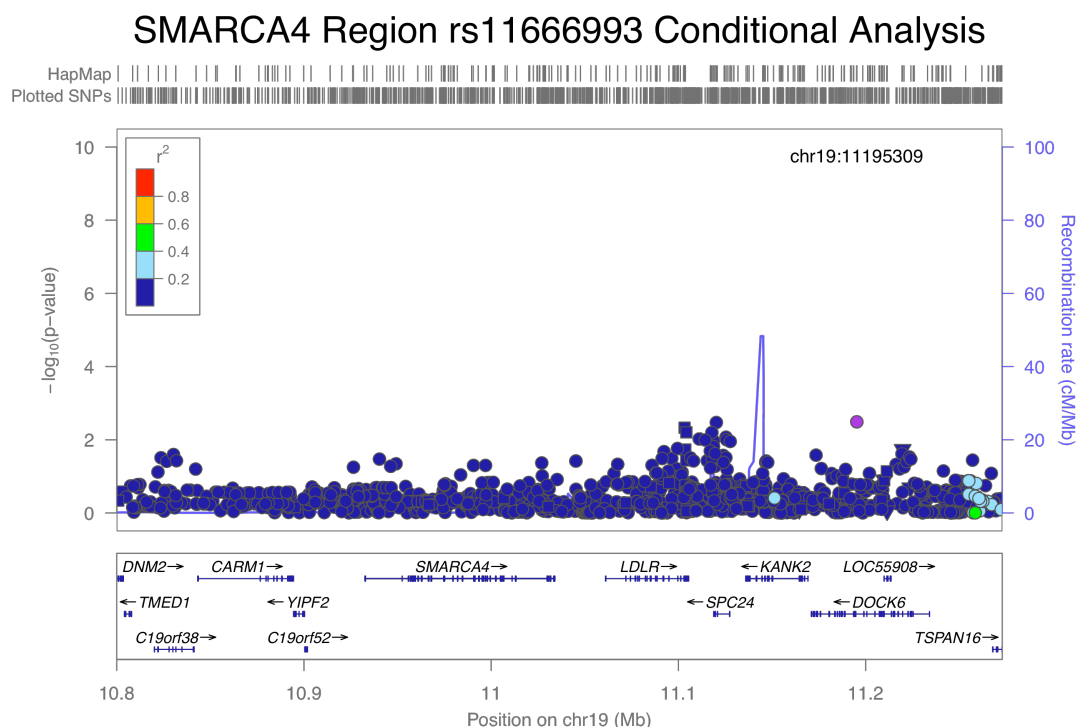
SMARCA4 encodes the highly conserved Brahma-related gene-1 (BRG1) protein that acts as an ATPase in the multimeric SWI/SNF complex for chromatin remodeling that regulates gene expression (Phelan *et al.* 1999). BRG1 regulates transcription of various proteins such as Smad3, Stat1, Stat2, beta-catenin, p53 and pRb (Trotter & Archer 2008). Its expression is tightly controlled as changes lead to cell senescence and apoptosis (Alessio *et al.* 2010). Moreover, chromatin

remodeling and Brg1 ATPase function have been shown to be essential for T cell development and immune response in mice (Gebuhr *et al.* 2003).

***SMARCA4* Region Unconditional Analysis**



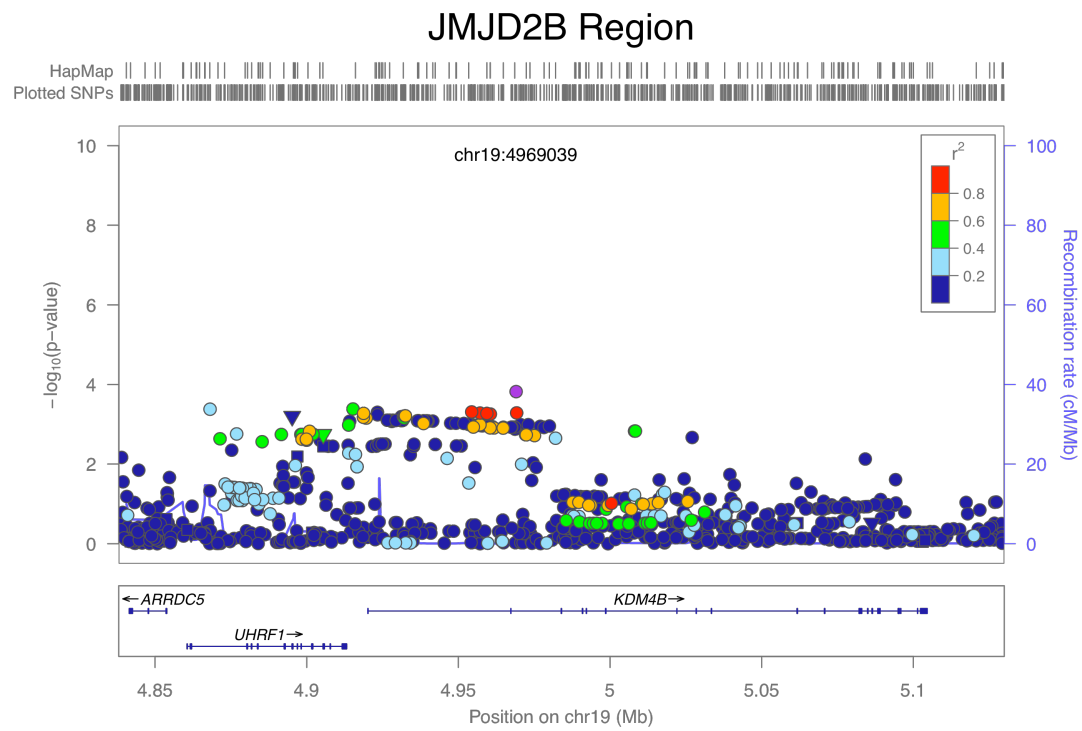
Plot 6.14. *SMARCA4* region unconditional analysis – showing the top hit SNP (rs11666993, purple) in the analysis of imputed genotypes, and the LD pattern.



Plot 6.15. *SMARCA4* region rs11666993 conditional analysis using imputed data only – showing that rs11666993 association was the only independent association in the region.

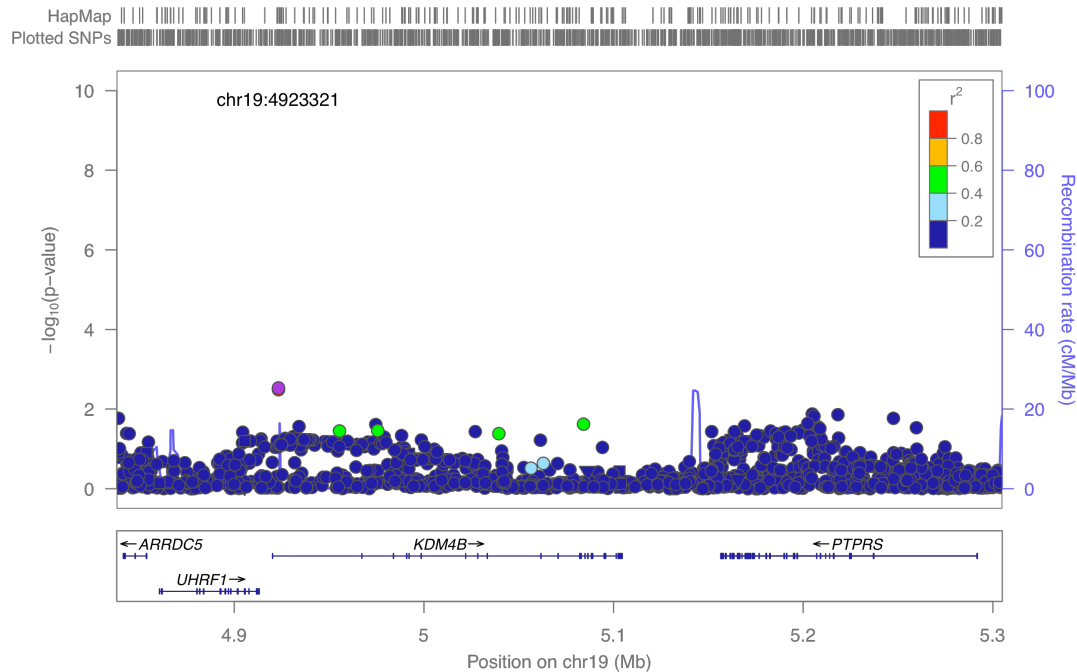
Chromosome 19 – *JMJD2B*

JMJD2B (*KDM4B*) encodes a member of the JMJD2 family of histone demethylases (Kato 2004; Tan *et al.* 2008). Histone tail (i.e. epigenetic) modifications play an important role in the control of gene expression. JMJD2B specifically demethylates H3K9 trimethyl at pericentric heterochromatin, where trimethylated lysine is principally associated with transcription repression (Fodor *et al.* 2006).



Plot 6.16. *JMJD2B* region unconditional analysis – showing the top hit SNP (rs60951857, purple) in the analysis of imputed genotypes, and the LD pattern.

JMJD2B Region rs60951857 Conditional Analysis



Plot 6.17. *JMJD2B* region rs60951857 conditional analysis using imputed data only – showing that rs60951857 association was the only independent association in the region.

References for Appendix 3

- Fodor, B. D., Kubicek, S., *et al.* (2006). "Jmjd2b antagonizes H3K9 trimethylation at pericentric heterochromatin in mammalian cells." *Genes Dev* **20**(12): 1557-1562.
- Gebuhr, T. C., Kovalev, G. I., *et al.* (2003). "The role of Brg1, a catalytic subunit of mammalian chromatin-remodeling complexes, in T cell development." *J Exp Med* **198**(12): 1937-1949.
- Guo, Z. S., Li, C., *et al.* (2010). "Association of IL-1 gene complex members with ankylosing spondylitis in Chinese Han population." *Int J Immunogenet* **37**(1): 33-37.
- Hampe, J., Franke, A., *et al.* (2007). "A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1." *Nat Genet* **39**(2): 207-211.
- Karaderi, T., Harvey, D., *et al.* (2009). "Association between the interleukin 23 receptor and ankylosing spondylitis is confirmed by a new UK case-control study and meta-analysis of published series." *Rheumatology (Oxford)* **48**(4): 386-389.

- Katoh, M. (2004). "Identification and characterization of JMJD2 family genes in silico." *Int J Oncol* **24**(6): 1623-1628.
- Kundu, M. & Thompson, C. B. (2008). "Autophagy: basic principles and relevance to disease." *Annu Rev Pathol* **3**: 427-455.
- Li, S., Strelow, A., *et al.* (2002). "IRAK-4: a novel member of the IRAK family with the properties of an IRAK-kinase." *Proc Natl Acad Sci U S A* **99**(8): 5567-5572.
- Maksymowych, W. P., Rahman, P., *et al.* (2006). "Association of the IL1 gene cluster with susceptibility to ankylosing spondylitis: an analysis of three Canadian populations." *Arthritis Rheum* **54**(3): 974-985.
- Phelan, M. L., Sif, S., *et al.* (1999). "Reconstitution of a core chromatin remodeling complex from SWI/SNF subunits." *Mol Cell* **3**(2): 247-253.
- Reveille, J. D., Sims, A. M., *et al.* (2010). "Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci." *Nat Genet* **42**(2): 123-127.
- Scherz-Shouval, R., Shvets, E., *et al.* (2007). "Reactive oxygen species are essential for autophagy and specifically regulate the activity of Atg4." *EMBO J* **26**(7): 1749-1760.
- Schraml, B. U., Hildner, K., *et al.* (2009). "The AP-1 transcription factor Batf controls T(H)17 differentiation." *Nature* **460**(7253): 405-409.
- Shintani, T. & Klionsky, D. J. (2004). "Autophagy in health and disease: a double-edged sword." *Science* **306**(5698): 990-995.
- Tan, H., Wu, S., *et al.* (2008). "The JMJD2 members of histone demethylase revisited." *Mol Biol Rep* **35**(4): 551-556.
- Trotter, K. W. & Archer, T. K. (2008). "The BRG1 transcriptional coregulator." *Nucl Recept Signal* **6**: e004.

PUBLICATIONS

Karaderi T, Cohen CJ, Pointon JJ, Wordsworth BP, *A CCR6 variant strongly associated with rheumatoid arthritis in two populations is not associated with ankylosing spondylitis*. Rheumatol Int. 2012 Apr 8.

Karaderi T, Pointon JJ, Wordsworth TWH, Harvey D, Appleton LA, Reveille JD, Farrar C, Harin A, Brown MA, Wordsworth BP, *Limited genetic association between TNFRSF1A, encoding the p55 tumour necrosis factor receptor, and ankylosing spondylitis in the UK Caucasian population*. Clinical & Experimental Rheumatology. 2012 Jan-Feb;30(1):110-3.

Harvey D, Pointon JJ, **Karaderi T**, Appleton LH, Farrar C, Wordsworth BP. *A common functional variant of endoplasmic reticulum aminopeptidase 2 (ERAP2) that reduces major histocompatibility complex class I expression is not associated with ankylosing spondylitis*. Rheumatology (Oxford). 2011 Sep;50(9):1720-1.

Australo-Anglo-American Spondyloarthritis Consortium (TASC) and the Wellcome Trust Case Control Consortium 2 (WTCCC2). *Genome-wide association study in ankylosing spondylitis identifies further non-MHC associations, and demonstrates that the ERAP1 association is restricted to HLA-B27 positive cases implicating peptide presentation as the likely mechanism underlying the association of HLA-B27 with the disease*. Nat Genet. 2011 Jul 10;43(8):761-7.

Pointon JJ, Harvey D, **Karaderi T**, Appleton LH, Farrar C, Wordsworth BP. *The histone demethylase JARID1A is associated with susceptibility to ankylosing spondylitis*. Genes Immun. 2011 Jul;12(5):395-8.

Pointon JJ, Harvey D, **Karaderi T**, Appleton L, Farrar C, Stone MA, Sturrock RD, Brown MA, Wordsworth BP. *Elucidating the Chromosome 9 Association with AS; CARD9 is a Candidate Gene*. Genes and Immun. 2010 Sep;11(6):490-6.

Australo-Anglo-American Spondyloarthritis Consortium (TASC). *Genomewide association study of ankylosing spondylitis identifies multiple non-MHC susceptibility loci*. Nat Genet. 2010 Feb;42(2):123-7.

Pointon JJ, Harvey D, **Karaderi T**, Appleton L, Farrar C, Stone MA, Sturrock RD, Reveille JD, Weisman MH, Ward MM, Brown MA, Wordsworth BP. *The chromosome 16q region associated with ankylosing spondylitis includes the candidate gene TRADD (TNF receptor type 1-associated death domain)*. Ann Rheum Dis. 2010 Jun;69(6):1243-6.

Harvey D, Pointon JJ, Evans DM, **Karaderi T**, Farrar C, Appleton LA, Sturrock RD, Stone MA, Oppermann U, Brown MA, Wordsworth BP. *Investigating the genetic association between ERAP1 and ankylosing spondylitis*. Hum Mol Genet. 2009 Nov 1;18(21):4204-12.

Karaderi T, Harvey D, Farrar C, Appleton L, Stone MA, Sturrock RD, Brown M, Wordsworth BP and Pointon JJ. *Association between the interleukin 23 receptor and ankylosing spondylitis is confirmed by a new UK case-control study and meta-analysis of published series*. Rheumatology (Oxford). 2009 Apr;48(4):386-9.