GENETICS OF RHEUMATOID ARTHRITIS IN BLACK SOUTH AFRICANS

Nimmisha Harilall Govind

A thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, in fulfilment of the requirements for the degree of

Doctor of Philosophy

Johannesburg, 2013

DECLARATION

I, Nimmisha Harilall Govind declare that this thesis is my own work. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other university.

.....

.....day of.....2013

To my loving parents, Harilall and Tara Govind

PUBLICATIONS AND PRESENTATIONS ARISING FROM THIS PROJECT

Publications:

Genetic diversity in black South Africans from Soweto. Andrew May[1,2], Scott Hazelhurst[3], Yali Li[4], Marie Waldvogel[5], Juerg Eigenburger[5], Shane Norris[6], Nimmisha Govind[1,7], Mohammed Tikly[7], Claudia Hon[4], Keith J. Johnson[4], Nicole Hartmann[5], Frank Staedler[5] and Michèle Ramsay[1,2], BMC Genomics, September 2013

Oral Presentations:

Nimmisha Govind[1], Richard J. Reynolds[2], Bridget Hodkinson[1], Claudia Ickinger[1], Jacqueline Frost[3], Miche´le Ramsay[3], Zenoria Causey[2], S. Louis Bridges[2], Jr, Mohammed Tikly[1]. **HLA DRB1 Amino Acid position 11 is strongly associated with Rheumatoid Arthritis in Black South Africans.** The European League against Rheumatism congress, Madrid, Spain, 13 June 2013

Nimmisha Govind[1], Richard J. Reynolds[2], Bridget Hodkinson[1], Claudia Ickinger[1], Jacqueline Frost [3], Michèle Ramsay[3], Zenoria Causey[2], S. Louis Bridges, Jr[2], Mohammed Tikly[1]. **Valine at position 11 of HLA DRB1 is strongly associated with Rheumatoid Arthritis in Black South Africans.** African League of Associations for Rheumatology and South African Rheumatism and Arthritis Association congress, Durban, South Africa, 4 April 2013

Nimmisha Govind[1], Ananyo Choudhury[2,3], Bridget Hodkinson[1], Claudia Ickinger[1], Jacqueline Frost [2], Annette Lee[4], Peter K. Gregersen[4], Richard J. Reynolds[5], S. Louis Bridges, Jr[5], Scott Hazelhurst[3], Michèle Ramsay[2], Mohammed Tikly[1]. **Genotyping of Risk Loci in Black South Africans with Rheumatoid Arthritis using the Immunochip: An Association study.** African League of Associations for Rheumatology and South African Rheumatism and Arthritis Association congress, Durban, South Africa, 5 April 2013

Nimmisha Govind, Ananyo Choudhury, Bridget Hodkinson, Claudia Ickinger, Jacqueline Frost, Annette Lee, Peter K. Gregersen, Richard J. Reynolds, S. Louis Bridges, Jr, Scott Hazelhurst, Michèle Ramsay, Mohammed Tikly. **Dense genotyping of risk loci in Black**

South Africans with rheumatoid arthritis. University of the Witwatersrand, Faculty of Health Sciences, Research Day, Johannesburg, South Africa, September 2012

Nimmisha Govind. **The Genetics of Rheumatoid Arthritis in Black South Africans,** Carnegie Symposium, Johannesburg, 19 March 2013

Poster presentation:

Nimmisha Govind, Ananyo Choudhury, Bridget Hodkinson, Claudia Ickinger, Jacqueline Frost, Annette Lee, Peter K. Gregersen, Richard J. Reynolds, S. Louis Bridges, Jr, Scott Hazelhurst, Michèle Ramsay, Mohammed Tikly (2012). **Dense genotyping of risk loci in Black South Africans with rheumatoid arthritis.** Poster number 975. American College of Rheumatology annual meeting, Washington DC, USA, November 2012

ABSTRACT

Introduction

The association of the HLA shared epitope with rheumatoid arthritis (RA) in black South Africans is well established. The aims of the thesis were to identify non-HLA risk loci associated with RA using the Immunochip array and to assess the association of specific amino acids at specific positions of the *DRB1* locus with the risk for developing RA in black South Africans.

Methods

Ethnically and geographically matched RA cases (n=435) and controls (n=463) were genotyped using the Immunochip array which has ~196 000 single nucleotide polymorphisms (SNPs) previously associated with 12 autoimmune diseases. After quality control, 103 700 SNPs were tested for association in 263 cases and 374 controls. For the amino acid study, DNA sequencing of exon 2 of the *DRB1* locus was performed on the seropositive cases (n=261) using the Allele SEQR HLA-DRB1 (Abbot) assay and four digit HLA typing was assigned using the Assign software (Conexio Genomics). The amino acid sequences were inferred from the called allele type. Association testing and conditional analysis were performed.

Results

A total of 72 HLA SNPs reached statistical significance ($p< 5x10^{-8}$) of which 71 SNPs locate to the *HLA-DR* or *DQ* genes or to the intergenic region between these two genes. Specifically, 4 SNPs in the intergenic region between *HLA-DRB1* and *HLA-DQA1* (rs3104413, OR=3.88, $p=5.49x10^{-21}$; rs3129769, OR=3.91, $p=4.60x10^{-21}$; rs6931277, OR=3.97, $p=1.03x10^{-21}$; rs9272353, OR=4.1, $p=3.01x10^{-21}$) were highly associated. Ten non-*HLA* SNPs on chromosome 6 reached

significance of which 7 SNPs locate to the intergenic region between *BTNL2* and *HLA-DRA* and confer protection. Four “SNPs” on chromosome 1 locating to a copy number variant region of the intergenic region between *PLD5* and *LOC400*723 were significantly associated with RA in this study (rs2809867, OR=0.33, $p=5.01x10^{-08}$; rs12738883, OR=0.33, $p=3.98x10^{-08}$; rs78352781, OR=0.33, $p=3.98x10^{-08}$; rs12739315, OR=0.33, $p=3.98x10^{-08}$) and showed an even stronger association with the rheumatoid factor positive subgroup. Valine at amino acid position 11 of *DRB1* demonstrated the strongest associated with RA (OR=5.1(3.7, 7.0), $p=1.63x10^{-27}$) and serine at this position was protective (OR=0.4(0.3, 0.5), $p=2.46x10^{-16}$). Interestingly, the frequency of valine in black South African RA cases was much lower than in Caucasians with RA. Using principal component analysis, black South Africans were found to be genetically distinct from west and east African populations.

Conclusion

This study demonstrated both similarities and differences in the risk for RA between Caucasians and black South Africans. Similar to Caucasians, the HLA region confers the strongest genetic risk for RA in black South Africans. More specifically, valine at position 11 of DRB1 confers the strongest risk for RA in black South Africans and serine confers protection. Four novel non-HLA RA-associated SNPs in the intergenic region between *LOC400723* and *PLD5* were identified. Variants of the *PTPN22* gene, although strongly associated with RA in Caucasians, are not associated with RA in this population. Since this population is genetically distinct, the findings need to be independently validated in other African populations.

ACKNOWLEDGEMENTS

This dissertation is a culmination of the hard work and dedication of a remarkable group of people. It is with the utmost humility that I express my gratitude to those individuals who have been integral to this project.

To my supervisor and mentor, Professor Mohammed Tikly, I am eternally grateful for your tutelage. Your guidance and support through the years has greatly enriched my work. As a fledgling in the field of genetics, I could not ask for a more understanding supervisor than Professor Michèle Ramsay, whose encouragement and supervision allowed me to grow as a scientist.

Certainly there were times when I felt out of my depth, and not more so than when waddling through the murky waters of bioinformatics. Had it not been for the “life-saving” efforts of Ananyo Choudhury, Scott Hazelhurst and Andrew May of the University of the Witwatersrand, and Richard Reynolds of the University of Alabama, I would probably not have made it to shore. It was a joy working together with these gentlemen, who were always eager to help and share their knowledge.

I would like to express my appreciation to Jacqueline Frost, who schooled me in the performance of the specialised genetic laboratory techniques, essential to my work.

A heartfelt thank you to my rheumatology colleagues, Dr. Claudia Ickinger and Dr. Bridget Hodkinson, who provided me with invaluable clinical data for the study.

The hard work and dedication of Ms.Ntabiseng Monai in the administrative aspects of data collection and processing did not go unnoticed.

I am truly indebted to Professor Louis Bridges and his team at the University of Alabama, including Zenoria Causey, whose laboratory proficiency was

indispensible for the HLA genotyping. I will never forget the gracious hospitality that was shown to me during my stay in Birmingham.

One of the greatest privileges, and highlight of my study, was meeting Professor Peter Gregersen, a name synonymous with the genetics of rheumatoid arthritis. I am grateful for his invaluable intellectual input, but also his sheer enthusiasm for my endeavours and validation of my work. My experience at the Feinstein Institute was unforgettable, and I would like to acknowledge all those involved in the laboratory processing of our data, notably Annette Lee. She graciously took time away from her own work to demonstrate the Immunochip array.

I would also like to express my appreciation to Professor Soumya Raychaudhuri of Harvard University for endulging my questions and responding generously with valuable information.

I was afforded a great head start with a fellowship from the Carnegie Corporation of New York. My sincere gratitude to the directors at the University of Witwatersrand, Professor Pettifor and Professor Veriawa, as well as Professor Kramer of the post-graduate office at the University of the Witwatersrand. The opportunity enabled me to focus completely on my research.

This project would not be possible without the financial contributions of the Connective Tissue Diseases Research Fund of the University of Witwatersrand and the Division of Clinical Immunology and Rheumatology of the University of Alabama.

Finally, it would be amiss not to mention my closest friends and family, who have remained steadfast in my corner, despite my sometimes obsessive pursuit, of a dream that has at long last come to fruition. My sincerest gratitude to my brothers

Sanjay and Pravesh, sister-in law Versha and my nephew and niece, Veer and Anjani.

This thesis is dedicated to my dearest parents. The completion of this body of work is borne of the endless strength, support and love you have bestowed upon me. For this, and so much more, I am forever in your debt.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

American Colleague of Rheumatology	ACR
anti-citrullinated protein antibody	ACPA
antigen presenting cells	APC
ACPA positive	ACPA+
ACPA negative	ACPA-
Black South Africans	BSA
Han Chinese in Beijing, China	CHB
Hardy Weinberg Equilibrium	HWE
human leucocyte antigen	HLA
genome wide association studies	GWAS
Gujarati Indians in Houston, Texas	GIH
complement factor genes 2 and 4	C2, C4
interleukin	IL
Encyclopedia of DNA Elements	ENCODE
linkage disequilibrium	LD
lymphotoxin α and β	LTA, LTB
leukotriene	LTβ
Luhya in Webuye, Kenya	LWK
major histocompatibility complex	MHC
Maasai in Kinyawa, Kenya	MKK
matrix metalloproteinase	MMP
minor allele frequency	MAF
natural killer cell	NKC
nuclear factor kappa-light-chain-enhancer of activated B cells	NFκB
oral contraceptive pill	OCP
polymerase chain reaction	PCR
protein kinase C θ	PKC θ
RA synovial fibroblast	RASF
regulatory T-cells	Tregs
rheumatoid arthritis	RA
rheumatoid factor	RF
single nucleotide polymorphism	SNP
shared epitope	SE
T-cell receptor	TCR
T-helper	T_H
tumour necrosis factor	TNF
transmission disequilibrium test	TDT
systemic lupus erythematosus	SLE
Utah residents with Northern and Western European ancestry from the CEPH collection	CEU
vascular endothelial growth factor	VEGF
Yoruba in Ibadan, Nigeria	YRI

Chapter 1. INTRODUCTION AND LITERATURE REVIEW

Rheumatoid arthritis (RA) is a chronic autoimmune rheumatic disease characterised by inflammation and progressive destruction of synovial joints. Poorly controlled disease is associated with significant physical disability and reduced life expectancy of between 5-10 years. The last few decades have brought progress in the understanding of the pathogenesis, genetics and thus development of treatment for RA which has resulted in better outcomes for RA patients. Today RA is considered the commonest potentially treatable cause of disability in the western world (Emery and Salmon, 1995). However, despite these developments, there remains much to be understood about the pathogenesis of the disease and many unmet needs for patients at a clinical level.

The aims of the introduction of this thesis are to briefly describe the history of RA, explain the pathogenesis and clinical manifestations, highlight the progress in genetic studies and then describe the known HLA and non-HLA genes associated with RA, especially how they relate to sub-Saharan Africa.

1.1 History of rheumatoid arthritis

Rheumatoid arthritis is considered to be a modern disease although there is evidence of an arthritic disease suggestive of RA in the skeletal remains of Native Americans in Alabama dating back as early as 450BC (Feitsma et al., 2008). In contrast, there is little evidence of the existence of RA in the Old World prior to the 16th century. Historical characters such as Erasmus, considered as the ‘Prince of Humanism’, Louis XIV, Kant and others had suffered some form of arthritis. The paintings of Rubens, Renoir, Klee and Dufy depict physical changes which many physicians have described as being consistent with ‘modern day’ RA (Dieppe and Rogers, 1986).

Paul Rubens (1577-1640) who is considered to be the master of European baroque paintings lived in the city of Antwerp, a centre of European culture, which was frequented by sailors from the New World. It is hypothesised that shipments from the New World

> ‘revolutionized eating habits, doubled the available therapeutic assortment of medicinal plants, mixed genes of populations, swapped infectious plants...This series of conditions taken together, enabled an infectious, but in principle not aggressive, American agent to find genetically more receptive European hosts, with mixed blood, triggered from certain cities which were transit points in an ‘epidemic’ of rheumatoid arthritis.’ (Appelboom, 2005)

In 1800 a French physician, Dr Augustin Jacob Landre-Beauvais, who was based in the famed Salpetriere Hospital in Paris described a type of arthritis distinct from the commonly occurring monoarticular gout. Unlike gout which occurred in otherwise healthy men, this type of arthritis occurred in females. The disease ran a chronic course resulting in frail and weak patients. It affected multiple joints at the time of disease onset and resulted in deformities. The title of his doctoral thesis asks the question ‘*Doit-on admettre une nouvelle espece de goutte sous la denomination de goutte asthenique primitive?*’ (Should we recognise a new type of gout to be called primary asthenic gout?) (Landre-Beauvais, 2001). Landre-Beauvais’s thesis is recognised as the first clinical description of RA. The term ‘rheumatoid arthritis’ itself was coined in 1859 by a British rheumatologist, Dr Alfred Baring Garrod (Keitel, 2009). The name was based on the term ‘rheumatic fever’, an illness which includes joint pain. The suffix *-oid* ("resembling") gives the translation as *joint inflammation that resembles rheumatic fever*.

1.2 Epidemiology and clinical features

Rheumatoid arthritis is one of the most common autoimmune systemic diseases, afflicting 0.5-1% of adults worldwide. It affects twice as many females than males. The peak age of onset is the 4^{th}-5^{th} decade of life. The prevalence increases after the age of 60 years (Gabriel et al., 1999). The disease is known to occur worldwide, although there is considerable geographical and ethnic variation, with respect to both prevalence and severity (Silman and Pearson, 2002).

The disease is characterised by persistent synovitis usually involving peripheral joints in a symmetrical distribution. Classical RA causes swelling of the small joints of the hand but can affect any peripheral synovial joint (Fig1.1).

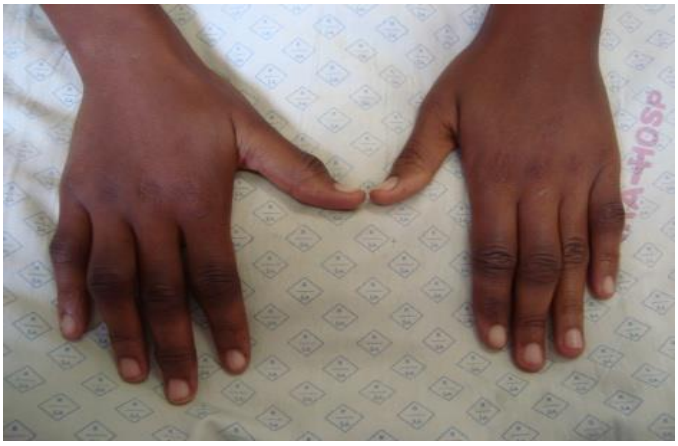

Figure 1.1 Swelling of the proximal interphalangeal joints in a patient with rheumatoid arthritis.

Extra-articular inflammation is not uncommon in RA. Inflammation of the organs may occur, especially in rheumatoid factor positive (RF+) patients. The commonest extra-articular manifestations are rheumatoid nodules and keratoconjunctivitis sicca which occur in 20-25% and 10% of patients respectively. Other extra-articular features include rheumatoid lung (interstitial lung disease, pleural effusions, pulmonary nodules and bronchiolitis), scleritis, serositis,

anaemia of chronic diseases, secondary Sjogren's syndrome, Felty's syndrome and vasculitis (Cojocaru et al., 2010).

Other than the extra-articular features, RA patients are at an increased risk of co-morbidities like osteoporosis, cardiovascular disease, infections and lymphomas (Michaud and Wolfe, 2007).

1.3 Health and economic burden

Cost-of-illness studies have shown that RA is associated with a high health and economic burden (Boonen and Mau, 2009). The economic 'burden' of disease relates to both direct and indirect costs.

The direct cost relates to the actual payment made for diagnosis, therapy, complications and prevention of the disease. The direct costs of RA are 2-3 times higher than the average cost for individuals of a similar age and gender (Lubeck, 2001).

The indirect costs of RA are substantial and relate to loss of resources to the individual, their family and society. These costs relate to work disability and early retirement. Approximately a quarter of patients will discontinue paid work within 2 years of RA diagnosis (Kvien, 2004). A South African study showed that only 24% of RA patients were employed and 35% were dependent on financial support from the State (Mody et al., 1988).

Indirect cost of RA is also linked to an increased mortality due to RA. Life expectancy is decreased by 5-10 years (Kvien, 2004). The five-year survival of patients in the American College of Rheumatology (ACR) functional class four is similar to those patients of triple vessel disease or stage IV Hodgkin's lymphoma (Pincus et al., 1984). The major causes for death in RA are infections,

cardiovascular and cerebrovascular disease (Douglas et al., 2006, Meune et al., 2009).

The psychosocial burden of the disease includes loss of quality of life, pain, frustration and lack of self esteem. Studies done in South Africa have shown that RA has a significant effect on health-related quality of life (Benitha and Tikly, 2007) and it is also associated with a negative impact on the individual's psychological and social functioning (Schneider et al., 2008).

1.4 Aetiopathogenesis

The aetiology of RA is not entirely understood but twin studies suggest that there is a substantial heritable contribution to the risk for RA. The heritability of RA is estimated to be as high as 60% (MacGregor et al., 2000). It is now widely accepted that RA occurs in a genetically susceptible host that is exposed to an environmental factor.

1.4.1 Environmental factors

No single environmental factor has been found to be causative in all RA patients. However, there is evidence of varying strength to suggest that microbes, smoking and dietary factors play role in the causation of the disease.

1.4.1.1 Infections

Microbes such as *Epstein-Barr virus* (Ollier, 2000, Saal et al., 1999), *Parvovirus B19* (Murai et al., 1999) and *Mycobacteria* (Tsoulfa et al., 1989) have been thought to increase the susceptibility for RA, probably through molecular mimicry of specific microbial peptides, with autologous molecules such as the rheumatoid or cartilage-derived epitopes. However, convincing evidence of a specific pathogen-derived antigen or of cross-reactivity of self-antigen specific T or B-cells

with pathogen-derived peptides is lacking. Periodontitis caused by *Porphyromonas gingivalis* is associated with anti-citrullinated peptide antibodies (ACPA) (Hitchon et al., 2010). Citrullination is the post-transcriptional modification of the positively charged amino acid arginine to the neutral citrulline which is mediated through one of the 5 isoforms of the enzyme peptidylarginine deiminase (PADI). It is hypothesised that citrullination occurs in the oral mucosa following infection with *Porphyromonas gingivalis* and in the lung as a result of smoking.

1.4.1.2 Smoking

Smoking is the strongest known environmental risk factor associated with RA and is also the most prominent example of a gene-environment interaction in RA pathogenesis. This association was first made by several epidemiological studies (Silman et al., 1996, Stolt et al., 2003, Uhlig et al., 1999). Later, an association with the HLA DRB1 shared epitope (SE) alleles and smoking was made with RF+ RA (Padyukov et al., 2004). Subsequently, Klareskog et al. (2006) showed that smoking predisposes to RA by increasing citrullinated peptides in lungs of patients carrying the HLA DRB1 SE alleles. Their findings were that smokers who do not carry the SE alleles have a 1.5-fold increased risk of developing ACPA positive (ACPA+) RA compared to non-smokers who do not carry the SE alleles. However smokers who carry 2 copies of the SE alleles have a 21-fold increase in developing ACPA+ RA compared to non smokers who do not carry the SE alleles. Later, it was shown that the risk for RA increases further when smokers carry a combination of the HLA DRB1 SE alleles and the R620W polymorphism of the *PTPN22* gene (Fig 1.2) (Kallberg et al., 2007). The risk increases with both increasing duration of smoking and increasing number of cigarettes smoked. Smokers with a greater than 40 pack year smoking history have a 2-fold increased

risk of RA compared to non-smokers. Tobacco exposure increases the risk for ACPA+ RA, but only in SE positive patients (Linn-Rasker et al., 2006). A similar gene-environmental risk has been observed in RF+ males and smokers (Sugiyama et al., 2010).

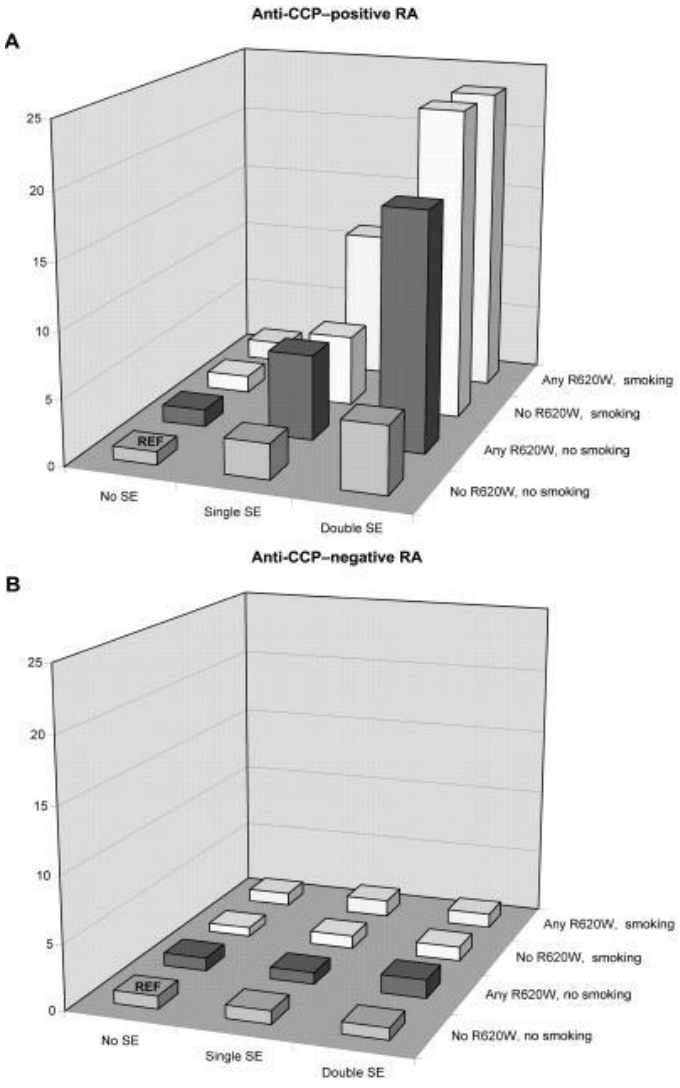

Figure 1.2 Histograms of odds ratios (OR) for developing ACPA+ (A) and ACPA- (B) rheumatoid arthritis and the dose effect of the SE alleles and the R620W polymorphism of the *PTPN22* gene and smoking status (Kallberg et al., 2007).

1.4.1.3 Hormones

Rheumatoid arthritis occurs more frequently in women than men. It has been hypothesised that this difference in frequency is due to the proinflammatory milieu created by oestrogen. Oestrogen is known to exert a stimulatory effect on the

immune system by inhibiting T-cell suppressor function and facilitating T-cell maturation (Cutolo, 2004). Women with older age at menarche (>15 years) have an almost 2-fold increased risk of developing RA as compared to women with early menarche (<12). Pregnancy increases the risk of developing RA. Twelve percent of women with RA develop RA within a year after pregnancy. Furthermore, females having multiple pregnancies are at risk of severe RA. Longer duration of breastfeeding increases the risk of RA in individuals especially those who are ACPA+ or IgM RF+ (Berglin et al., 2010). This is in contrast with previous studies which show that long term breast feeding is associated with a decrease risk of RA (Karlson et al., 2004). The use of the oral contraceptive pill (OCP) delays onset of disease rather than protects against its occurrence (Doran et al., 2004). Furthermore, in asymptomatic females there is an inverse relation between RF seropositivity and the use of the OCP (Bhatia et al., 2007).

1.4.1.4 Dietary factors

In general dietary factors seem to play a minor role in RA susceptibility but some associations have been described. A diet low in antioxidants, e.g. vitamin C and β-cryptoxanthin which is a constituent of yellow peeled food such as citrus, has been found to increase the risk of RA, as has excessive coffee consumption, but only in ACPA+ RA (Pattison et al., 2005). Alcohol was found to decrease the risk of developing RA (Kallberg et al., 2009). The role of red meat and animal protein is less clear, some studies show an increased risk for RA and other studies suggest otherwise (Pattison et al., 2004, Benito-Garcia et al., 2007). The role of vitamin D in RA risk remains equivocal with conflicting results from different studies (Costenbader et al., 2008, Nielen et al., 2006).

Higher birth weight (Jacobsson et al., 2003, Mandl et al., 2009) and lower socio-economic status (Bengtsson et al., 2005, Pedersen et al., 2006) increase the risk of developing RA. Occupational exposure to mineral oils was found to increase the risk of RA in a Swedish study (Sverdrup et al., 2005).

1.4.2 Genetic factors

Familial clustering, with RA occurring more frequently in first degree relatives of RA and a higher concordance in monozygotic twins (15%) versus dizygotic twins (3.5%) has been strongly suggestive of heritability or genetic contribution to RA (MacGregor et al., 2000). The increased risk of the disease in siblings of patients with RA compared with that of the general population has been estimated to be between 2 and 17 fold (Seldin et al., 1999).

1.4.2.1 HLA associations with rheumatoid arthritis

The HLA DRB1 gene is estimated to contribute approximately a third of the genetic risk for RA which is the strongest genetic association. Alleles that encode for a specific sequence of amino acids on the third hypervariable region of the HLA DRB1 chain called the ‘shared epitope’ alleles demonstrate the strongest association with RA and confer risk in most of the ethnic groups studied thus far (Holoshitz, 2010). The HLA association with RA will be discussed in detail in chapter 3.

1.4.2.2 Non-HLA genes associated with rheumatoid arthritis

The non-HLA genes however show less consistent risk for RA across the major ethnic groups. A summary of some of the polymorphisms associated with RA is briefly described in Table 1.1.

Candidate gene and genome wide association studies (GWAS) have identified numerous risk loci associated with RA. The function of these genes relate to T- and B-cell activation such as *HLA*, *PTPN22, STAT4, CTLA4, CD40* or to cytokine expression such as *IL-2, IL-21, CCL21* or to the NFκB pathway such as *TRAF1, TRAF6, TNFAIP3, REL*, co-receptor stimulation (*CTLA4* and *CD28*), transcription signalling (*STAT4*), citrullination (*PADI4*), T-cell receptor signalling (*PTPN22*), cytokine regulation (*TRAF1/C5*). Most of the loci identified are associated with ACPA+ RA in Caucasians.

Table 1.1 Non-HLA genes associated with ACPA+ rheumatoid arthritis in Caucasians and their function, the genetic variant and the effect size of the risk it confers and their association with other autoimmune diseases

Gene (chromosome)	**Protein function**	**SNP or allele**	**OR (95% CI)**	**Reference**	**Autoimmune disease**
PTPN22 (1p13)	Encodes a lymphocyte specific phosphatase involved in the regulation of T-cells	rs2476601 rs6679677	1.94 (1.81-2.08) 2.06 (1.68-2-53)	(Stahl et al., 2010) (Kurreeman et al., 2011)	T1D, SLE, GD, Vitiligo
CTLA4 (2q33)	The *CTLA-4* gene is a member of the immunoglobulin family. It binds to the CD80/86 molecule expressed on antigen presenting cells. It down-regulates T-cells by preventing the binding of the co-stimulatory molecule CD28 to CD80/86.	rs3087243	1.15 (1.10-1.20)	(Stahl et al., 2010)	GD, HT, SS, T1D, WG
STAT4 (2q32)	The *STAT4* gene is a member of the STAT family of transcription factors. They act as transcription activators of cytokines especially interferon γ involved in the differentiation of CD4 T-cell into Th-1 and Th-17 cells. STAT4 also signals in dendritic cells.	rs7574865	1.16 (1.10-1.23)	(Stahl et al., 2010)	CD, JIA, SLE, SS, SSc, T1D, UC
FCGR (1q21-23)	Encode receptors for the Fc fragment of IgG and is involved in clearing of immune complexes	FCGR2B rs1050501 FCGR3A rs393991 FCGR2A rs12746613	1.35 (1.09-1.66) 1.3 (1.01-1.55) 1.13(1.06-1.21)	(Chen et al., 2008) (Thabet et al., 2009) (Raychaudhuri et al., 2009)	
CD28 (2q33)	CD28 is a co-stimulatory molecule required for T-cell activation. CD28 binds to CD80 and CD86 molecules.	rs1980422	1.12(1.06-1.18)	(Stahl et al., 2010)	
RBPJ (4p15)	*RBPJ* is a transcription factor that is vital to the Notch signalling pathway that is involved with many cell-cell interactions and cell fate pathways.	rs874040	1.18(1.12-1.24)	(Stahl et al., 2010)	

TRAF1/C5 (9q33-34.1)	The TNF associated factor 1(*TRAF*) gene is an adaptor molecule that signals for TNF. The Complement 5 (*C5*) gene lies close to the *TRAF1* gene.	rs3761847 rs2900180 rs1011835 7	1.13 (1.08-1.18) 1.28 (1.13-1.44) 1.19 (1.05-1.34)	(Stahl et al., 2010) (Patsopoulos and Ioannidis, 2010) (Kurreeman et al., 2011)	AS, JIA, SLE
BLK (8p23)	The B-lymphoid tyrosine kinase gene encodes a tyrosine kinase of the src family which is involved in B-cell receptor signalling and B-cell development.	rs1327711 3 rs2736340	1.15 (1.00-1.33) 1.12 (1.07-1.18)	(Begovich et al., 2004) (Stahl et al., 2010)	

T1D = Type 1 diabetes, CD = Crohns disease, JIA = Juvenile inflammatory arthritis, SLE = Systemic lupus erythematosis, SSc = Systemic sclerosis, SS = Sjogren syndrome, GD = Graves' disease, WG = Wegners granulomatosis, PsA = Psoriatic arthritis, CeD = Celiac Disease, MS = Multiple sclerosis, AS = Ankylosing spondylitis

Unlike ACPA+ RA, only a few loci have been associated with ACPA negative (ACPA-) RA. In Caucasians the HLA DR3 allele which is part of the conserved ancestral haplotype (HLA A1-B8-DR3-DQ2) is associated with ACPA- RA. A GWAS study of ACPA- RA patients found the *IRF5* and *CLEC16A* genes to be associated with RA (Ruyssen-Witrand et al., 2012).

1.4.2.3 Genetic diversity between major ethic groups

Genetic diversity exists between major ethic groups, therefore genetic associations with disease in one population cannot be assumed to be causal in another. Humans are thought to have originated in Africa and migrated out from the East African gene pool, the so called 'Out of Africa model' (Tishkoff and Verrelli, 2003). Africans have the largest number of population specific alleles and non-Africans have much less genetic diversity (Gabriel et al., 2002, Verrelli and Tishkoff, 2004). This could be as a result of a genetic bottleneck in populations outside of Africa (Calafell et al., 1998). Apart from the differences in genetic diversity, the level of LD in Africans and non-Africans differ significantly, with Africans having lower levels of LD (Tishkoff and Verrelli, 2003, Tishkoff and Williams, 2002). Furthermore, the genetic diversity within African populations has been well established (Ramsay, 2012, Tishkoff et al., 2009). Specifically, black South Africans are genetically distinct from other African populations (May et al., 2013). This is evidenced by the significantly lower frequency of the SE alleles in the west African, Cameroonians (Singwe-Ngandeu et al., 2010) compared to black South Africans (Meyer et al., 2011). The genetic diversity and LD structure both between and within ethnic groups emphasise the need for population specific studies.

1.4.2.4 Genetic variants associated with rheumatoid arthritis in different ethnic groups

Genome wide association studies have been predominately performed on 2 broad ethnic groups, the Caucasians and Asians. These studies indicate that shared and ethnic specific risk loci exist. An example of the differences in risk between these two ethnic groups is a functional single nucleotide polymorphism (SNP), Arg620Trp (rs2476601), in the *PTPN22* gene which confers the second strongest genetic risk in Caucasians (Begovich et al., 2004) but does not confer risk in Asians (Lee et al., 2009, Ikari et al., 2006). Interestingly, this *PTPN22* gene variant is non-polymorphic in the black South Africans (Tikly et al., 2010). Likewise genetic variants in the *PADI4* gene confer risk in the Asians but not the Caucasians (Stahl et al., 2010). However, using the Immunochip, a group from Manchester in the UK, recently found an association with a SNP in intron 9 of the *PADI4* gene in a Caucasian population (Eyre et al., 2012). In addition, the *RUNX1* and *SLC22A4* (Tokuhiro et al., 2003) genes confer risk in the Asian but not the European populations (Karlson et al., 2008, Barton et al., 2004b). Interestingly, the *STAT4* genetic variants confer risk in both the European and Asian populations (Lee et al., 2009). In the admixed African American population it was found that of 27 candidate SNPs that confer risk in European populations, 23 also conferred risk in African Americans (Hughes et al., 2010). Variants of the *FCRL3* and *CD244* genes confer risk for RA in Japanese but not Caucasians (Suzuki et al., 2008). However, one study showed a *FCRL3* gene marker to be a risk factor for ACPA+ RA susceptibility and severity in a Norwegian population (Maehlen et al., 2011). Genetic variants that confer risk for RA in Caucasians such as markers associated with *PTPN22, TRAF/C5, CD40, CCL21, 6q23* and the *4q27* regions were either

non-polymorphic or did not show association with RA in Koreans (Lee et al., 2009). A study of North Indians of India showed limited replication of known European and Asian RA risk loci (Prasad et al., 2012).

Together these results provide evidence that genetic diversity and RA susceptibility differences exist between the major ethnic groups. These contrasting findings suggest that further research is required to examine the role of the non-HLA genetic variants in RA in various ethnic groups and may eventually provide insight in the quest for personalised therapy.

1.4.2.5 Genetic variants associated with several autoimmune diseases

Numerous autoimmune diseases share similarities in terms of demographics, clinical features and chronic inflammation of organs. In addition, GWAS have identified over 200 genetic loci associated with one or more autoimmune diseases, many of which are shared amongst various autoimmune diseases suggesting common biological pathways (Fig 1.3).

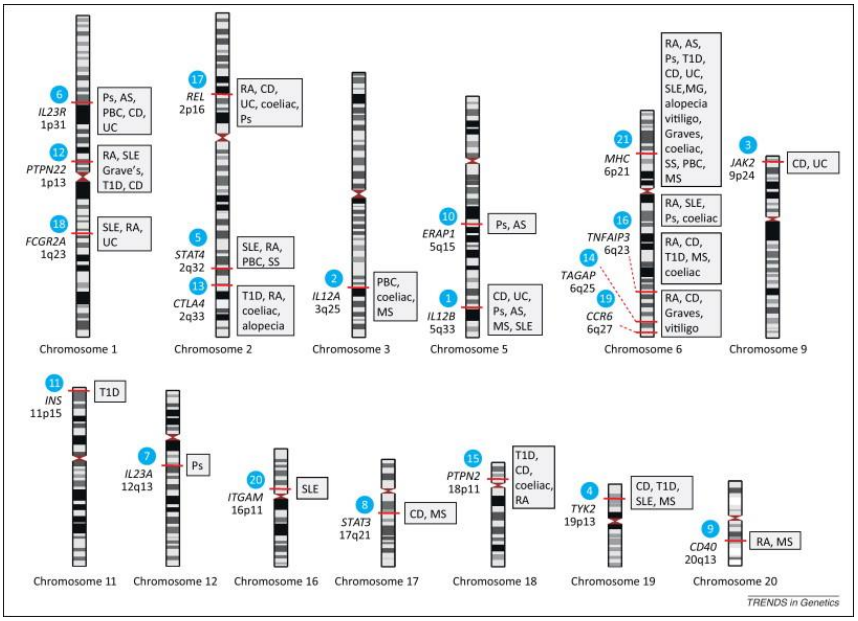

Figure 1.3 Chromosomes with overlapping risk regions amongst various autoimmune diseases (Knight, 2013).
T1D = Type 1 diabetes, CD = Crohns disease, SLE = Systemic lupus erythematosis, SS = Sjogren syndrome, Ps = Psoriasis, MS = Multiple sclerosis, AS = Ankylosing spondylitis, PBC = Primary biliary cirrhosis, MG = Myasthenia gravis

However, some diseases have distinct pathways of pathogenesis. These shared and distinct pathways are found in the innate and adaptive immune systems and more recently found to be involved in the threshold signals for cell activation. The implications of these findings is that therapies can be tested for effect in different autoimmune diseases known to have common biological pathways and more sophisticated genotyping techniques are being designed such as the Immunochip which take advantage of the knowledge of shared risk loci.

The HLA locus remains the strongest association with most autoimmune diseases. The R620W risk allele in the *PTPN22* gene confers risk for several autoimmune diseases such as RA (Begovich et al., 2004), systemic lupus erythematosus (SLE) (Orozco et al., 2005, Kyogoku et al., 2004), type 1 diabetes (Smyth et al., 2004, Zheng and She, 2005), and Graves' disease in populations of European descent (Velaga et al., 2004, Skorka et al., 2005). Variants in the *CTLA4* gene confers risk for Addison's disease, autoimmune pancreatitis (Chang et al., 2007), autoimmune hypothyroidism, celiac disease, chronic inflammatory arthritis (Suppiah et al., 2006), multiple sclerosis (Heggarty et al., 2007) and RA (Plenge et al., 2007b). Rheumatoid arthritis and SLE have been associated with the *BANK1* gene (Orozco et al., 2009, Kozyrev et al., 2008). The *TNFAIP3* locus has been associated with RA (Plenge et al., 2007a, Thomson et al., 2007), SLE (Graham et al., 2008) and type 1 diabetes (Fung et al., 2009).

1.4.2.6 Genetic variants associated with rheumatoid arthritis severity

Risk factors for severe RA include RF and ACPA positivity (Niewold et al., 2007), genetic susceptibility, smoking (Baka et al., 2009, Nyhall-Wahlin et al., 2009), poor functional class at presentation (Pincus et al., 1994), poor socio-economic

background (Camacho et al., 2012), older age at presentation (Bukhari et al., 2007), and the presence of extra-articular disease (Cojocaru et al., 2010).

Over recent years, GWAS have identified numerous genetic variants associated with RA susceptibility. However, few have focused on the contribution of genetic markers to the severity of radiological damage. Current biomarkers only explain a modest proportion of the variance in radiological damage. Apart from the HLA DRB1 SE alleles, only few other genes have been associated with severe erosive RA (van der Linden et al., 2009, Orozco et al., 2010, Marinou et al., 2010). A polymorphism in the *CD40* gene was found to be associated with joint destruction in ACPA+ RA (van der Linden et al., 2009). A study of polymorphisms in the *IL-1α* gene (Jouvenne et al., 1999) and the +3954 E2 allele of the *IL-1β* gene have shown associations with severe radiographic damage (Cantagrel et al., 1999, Buchs et al., 2001, Pawlik et al., 2005a). However the IL-1β 511 allele 2 is associated with milder radiographic disease (Genevay et al., 2002). The association of the *IL-6* 174G polymorphism and RA severity shows conflicting evidence for (Marinou et al., 2007) and against its risk for RA severity (Oen et al., 2005). The 1082 GG polymorphism in the *IL-10* gene has been associated with high radiographic damage scores in one study (Huizinga et al., 2000). However some studies show no association with radiographic disease (Cantagrel et al., 1999, Oen et al., 2005, Pawlik et al., 2005b). Two studies showed no association of the 590T allele in the *IL-4* gene and erosive RA (Genevay et al., 2002, Cantagrel et al., 1999). Genetic variants of the *TGFβ* (Oen et al., 2005, Mattey et al., 2005, Kim et al., 2004) and the *MMP3* (Dorr et al., 2004, Constantin et al., 2002, Mattey et al., 2004) genes are also associated with radiographic severity. The role of the R620W polymorphism in the *PTPN22* gene and disease severity is

conflicting (Pierer et al., 2006, Steer et al., 2005, Lie et al., 2007). Carriers of the A allele of the rs10818488 SNP in the *TRAF1* gene had a twofold higher radiographic damage score at 2 years than non-carriers (Kurreeman et al., 2007). A polymorphism in the *CCR5* gene was not found to be associated with juxta-articular erosions (Zapico et al., 2000). A study showed that the amino acid at position 70 in the epitope sequence can predict its risk for disease severity. Glutamine or arginine increases the risk of RA, whereas aspartic acid not only decreases the risk but also the severity of RA. The carriers had less radiographic erosions at 30 months of disease onset compared to non-carriers (Carrier et al., 2009).

1.4.2.7 Epigenetic factors

Despite the major role of genetic factors in RA susceptibility, the concordance of this disease in monozygotic twins is only 15% (Silman et al., 1993). It has therefore been hypothesized that whereas RA susceptibility is determined genetically, disease onset may depend on non-genetic or epigenetic events. Epigenetics refers to the changes in genomic function without alteration in the DNA sequence. The biological modifications of epigenetics processes include DNA methylation, histone methylation, acetylation, phosphorylation and sumoylation and miRNA expression levels (Fig 1.4). Due to their role in propagation of the disease many of the epigenetic studies have focused on the changes to the RA synovial fibroblast (RASF). These cells are globally hypomethylated which account for their activated phenotype (Karouzakis et al., 2009b), in addition they reside in a hyperacetylated synovial milieu (Huber et al., 2007) and are resistant to apoptosis due to sumoylation (Meinecke et al., 2007).

The expression of a distinct miRNA, miR-203, in RASF has been shown to increase the production of MMP-1 and IL-6 (Stanczyk et al., 2008).

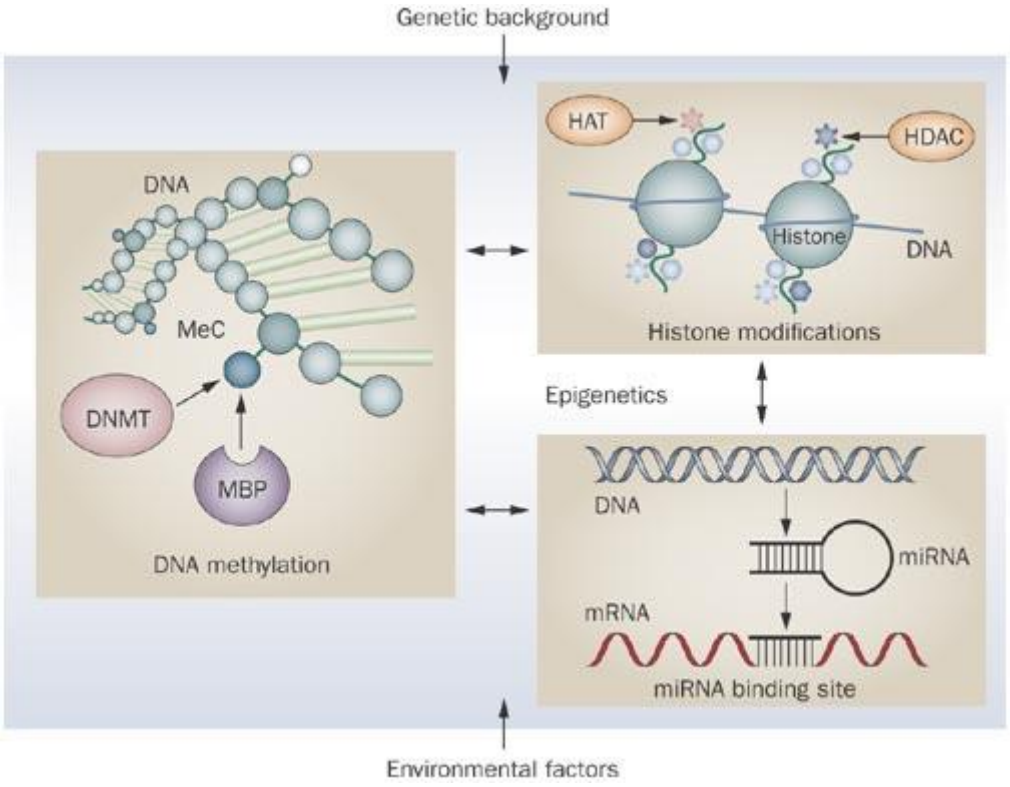

Figure 1.4 Epigenetic modification by DNA methylation, histone modification, and miRNA (Karouzakis et al., 2009a).

For the first time an association with an X chromosome locus in RA, IRAK1 (encoding interleukin-1 receptor-associated kinase 1) was demonstrated. This locus has been shown to escape X-inactivation in humans, pointing to a possible epigenetic mechanism underlying the sex bias in RA (Eyre et al., 2012).

1.5 Immunopathogenesis and pathology

The immune response in RA is thought to be initiated by the presentation of the putative antigen, possibly a citrullinated peptide, by antigen presenting cells (APC) including dendritic cells, macrophages and B-cells to the T-cell receptor (TCR) of the CD4+ T-cells (Huppa and Davis, 2003). In RA citrullinated peptides are presented by APC through HLA class II molecules to the TCR of CD4+ T-cells (Fig

1.5). This immunological synapse allows for presentation of a putative antigen to the TCR resulting in proliferation of T-cells in lymph nodes and joints. The antigen alone is not adequate to stimulate T-cells, co-stimulation from CD 80/86 on APC and CD28 on T-cells is required for complete activation. A negative regulator of T-cell signalling, cytotoxic T lymphocyte antigen 4 (CTLA4), competes with CD28 for binding with CD80/86.

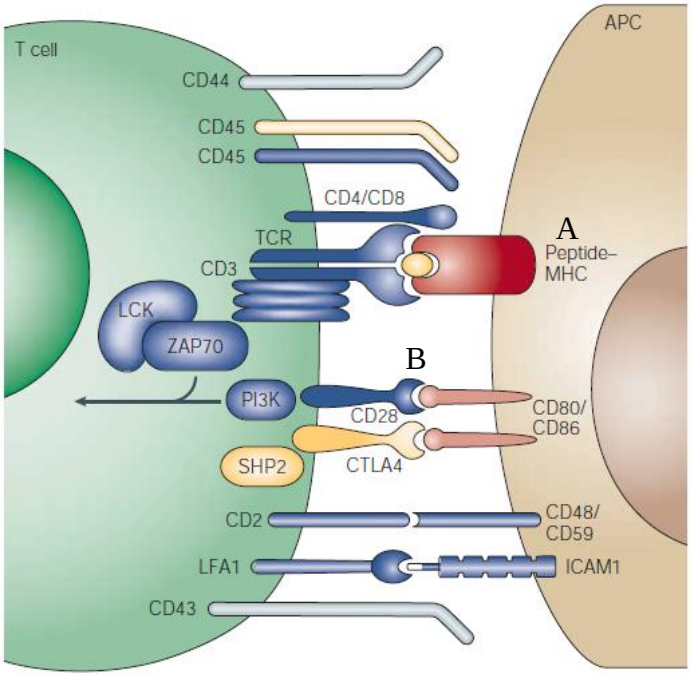

Figure 1.5 The immunological synapse showing the first signal (A) of T-cell activation initiated by the peptide being presented to the TCR by the HLA molecule and the second co-stimulatory signal (B) from one or more receptors (Huppa and Davis, 2003).

The T-cell phenotypes that predominate in RA are the T helper (T_H)1(T_H1) and T_H17 subtypes (Cooles and Isaacs, 2011). Within the T-cell there are intracellular transcription factor pathways which shuttle information about inflammatory stimuli to the cell nucleus. These pathways include the mitogen-activated protein kinases (MAPK) (Schett et al., 2008) and the janus kinase-signal transducers and activators of transcription (Jak-STAT) pathway, Syk, PI3K, BTK and NF-κb (McInnes and Schett, 2011). Transcription molecules stimulate the production of proinflammatory molecules such as interleukin (IL) -1, 6, 17, vascular endothelial

growth factor (VEGF) and tumor necrosis factor (TNF). These cytokines form an autocrine and paracrine feedback loop for the further interaction of T-cells, macrophages and B-cells. Although these cytokines have some unique roles, collectively their actions result in amplifying the immune response by increasing adhesion molecules, activation of monocytes, increasing the release of matrix metalloproteinases (MMP), promoting angiogenesis, stimulating T- and B-cell proliferation and differentiation and osteoclast activation. The upregulation of MMP results in cartilage destruction. The proinflammatory cytokines, IL-6 and TNF, upregulate receptor activator of nuclear factor kappa-B ligand (RANKL) and also result in osteoclast differentiation which leads to joint erosions.

Regulatory T-cells (Tregs) are vital in the regulation of peripheral T-cells. In RA, Tregs seem to have impaired regulatory function (McInnes and Schett, 2007). The role of B-cells in the pathogenesis of RA is through the production of autoimmune antibodies, cytokines and chemokines such as IL-6, IL-10 and leukotriene (LTβ) and the ability to present antigen.

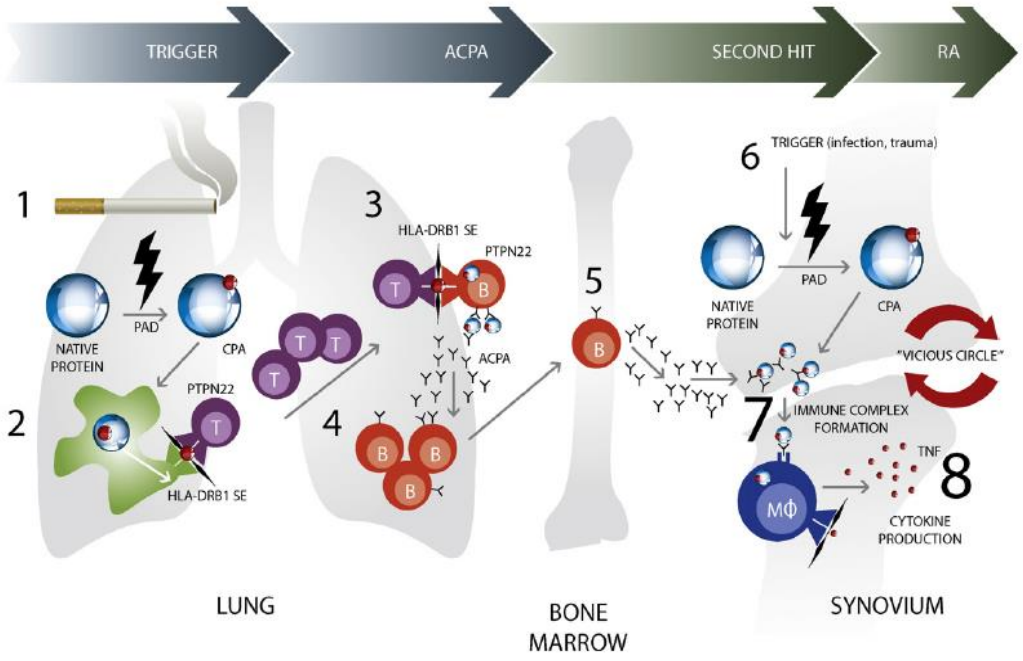

Figure 1.6 Hypothesis of the development of RA showing citrullination of peptides in the lung secondary to smoking (1), presentation of antigen to T-cells by APC (2), proliferation and stimulation of B-cells to produce anti-citrullinated antibodies (3-5) followed by a second hit resulting in local joint inflammation (6-8) (Klareskog et al., 2011).

1.5.1 Pathology of rheumatoid arthritis synovium

The pathological hallmark of RA is synovial hyperplasia known as ‘pannus’. The synovial lining is inflamed by the influx of T- and B-cells, plasma cells, dendritic cells, macrophages and mast cells. The pannus invades into cartilage and bone, secreting proteinases resulting in erosions and damage of surrounding joint structures such as the joint capsule and tendon (Pope, 2002) (Fig 1.7).

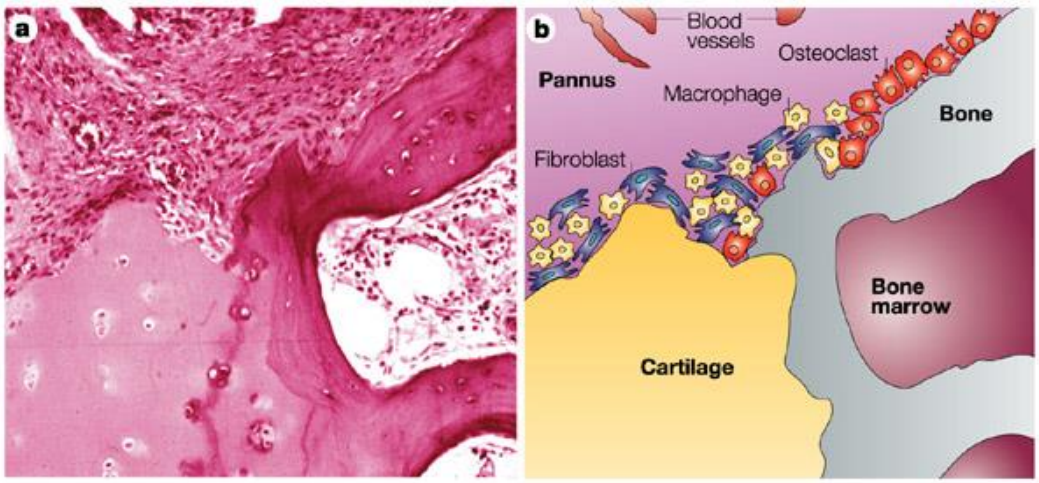

Figure 1.7 Histology of an ankle joint affected by rheumatoid arthritis (Pope, 2002).

1.5.2 Autoantibodies

Rheumatoid arthritis is characterised by the presence of autoantibodies of which rheumatoid factor (RF) are antibodies directed to the Fc component of IgG and have been the principal antibody in RA diagnosis. For RA, the specificity of RF is 87% (Shmerling and Delbanco, 1992). More recently, however anti-citrullinated protein antibodies have shown a better specificity of 93%-98% in RA (Niewold et al., 2007). In black South Africans, however, the specificity of RF (90.7%) is higher than the ACCP test (84.9%) (Hodkinson et al., 2010).

1.6 Rheumatoid arthritis in Sub-Saharan African

1.6.1 Prevalence and severity

Rheumatoid arthritis in Africans is thought to be less prevalent than in Caucasian populations. These observations date back to as early as 1956 in Malawi (Goodall, 1956), Uganda (Shaper and Shaper, 1958) and Kenya (Hall, 1966). Furthermore, earlier reports suggested that in Africans, RA was a mild disease and severe radiographic changes were uncommon. It was also thought that deformities were rare, extra-articular features were unusual and only symptomatic therapy was necessary to control symptoms in most patients (Adebajo and Reid, 1991). A study by Greenwood in 1969 described Nigerians with RA as having ‘a low incidence of nodules, vascular lesions, and peripheral neuritis’. He also found that RF was found no more frequently than in controls, radiological changes were mild, and the prognosis was good when compared to English patients (Greenwood, 1969).

However later studies showed discrepancies in the prevalence and severity of RA in Africans compared to those reported in earlier studies. A study in Uganda described a higher frequency than reported in Nigeria but not as high as in

Caucasians. In this study the radiological and clinical features of RA were similair to Caucasians. In Kenya the prevalence and clinical presentation were similar to Caucasians but there were no extra-articular manifestations and the disease had less functional impairment compared to Caucasians (Bagg et al., 1979, Chikanza et al., 1994, Ravindran et al., 2008). Later reports suggested the prevalence of RA in urban African populations to be similar to that in Western Europe and with severe disease more prominent than described in earlier reports (Kanyerezi et al., 1970, Bagg et al., 1979, Moolenburgh et al., 1984, Ouedraogo et al., 2011, Bileckot and Malonga, 1998). Rheumatoid arthritis, once a rarity in Africa, is now reported in large numbers from many parts of Africa. Further epidemiological studies are needed to confirm these findings in other parts of Africa and identify factors contributing to this difference to provide a better understanding for the emergence of RA in Africa.

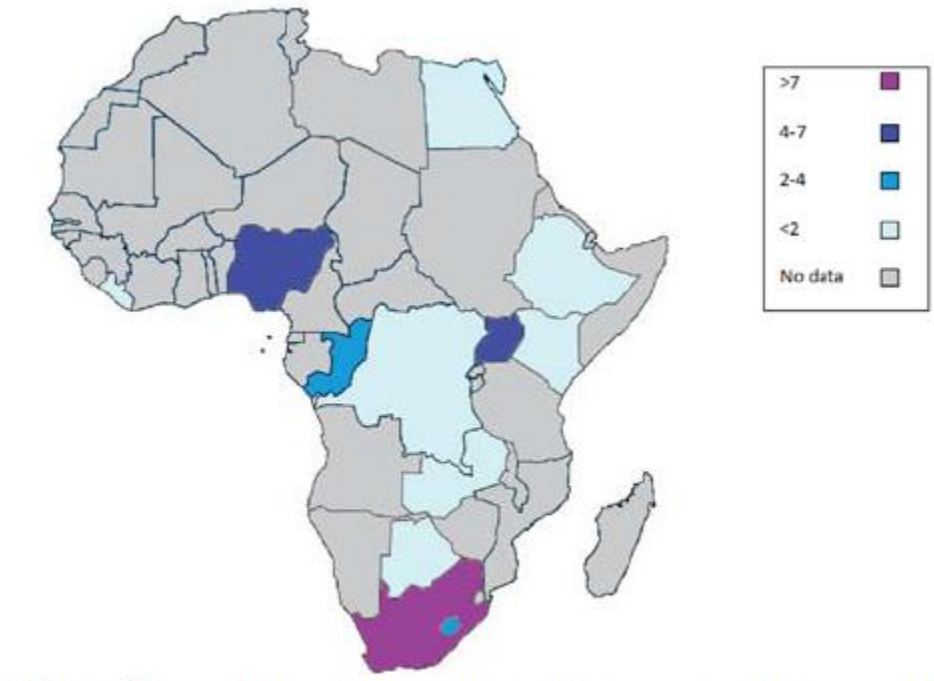

Figure 1.8 The number of prevalence studies of rheumatoid arthritis in different African countries (Dowman et al., 2012).

A recent and the only systematic literature review of the prevalence of RA in different African countries revealed that the estimated prevalence is 0.4% in 2010 (Dowman et al., 2012). This is less than the estimated 1% observed in Caucasians. Most of these studies were conducted in southern Africa (Fig 1.8). There was a difference in the prevalence between community and hospital-based studies, suggesting that less people in Africa present themselves to hospital with RA. There was a constant trend amongst all the studies of increasing prevalence with age.

1.6.2 Rheumatoid arthritis in black South Africans

1.6.2.1 Clinical and epidemiology aspects

Rheumatoid arthritis in black South Africans was initially thought to be uncommon and milder than reported in Caucasians. Epidemiological studies in the 1970s showed a higher prevalence of RA in urban compared with rural black South Africans. The prevalence of RA in the rural Tswana speakers was 0.12% (Beighton et al., 1975) which contrasted with the higher prevalence of 1% documented in Caucasians. However, a study done in urban blacks of Soweto, Johannesburg, showed that the prevalence was 0.9-1.4% and urban dwellers had more severe disease and features resembling classical RA observed in the Caucasian population (Solomon et al., 1975). Figure 1.9 shows the severe articular disease commonly observed in black South Africans. The difference in the prevalence of RA between rural and urban South African populations could be explained by differences in environmental factors in the two areas. A higher prevalence of smoking has been documented in urban versus rural populations (Peer et al., 2013), which in part could account for the differences in RA prevalence.

Recent experience shows that severe disease with deformities and radiographic changes are seen in black South Africans and a wide spectrum of extra-articular features are noted, although they may be less common than in Caucasians. Rheumatoid arthritis in black South Africans has a profound impact on functional disability and health related quality of life (Benitha and Tikly, 2007, Tikly et al., 2003). A third of black South Africans with RA require financial assistance from the state (Mody et al., 1988). Furthermore, significant radiographic changes of RA are frequently observed in black South Africans (Mody and Meyers, 1989) and the cardiovascular risk of RA is not decreased (Solomon et al., 2005).

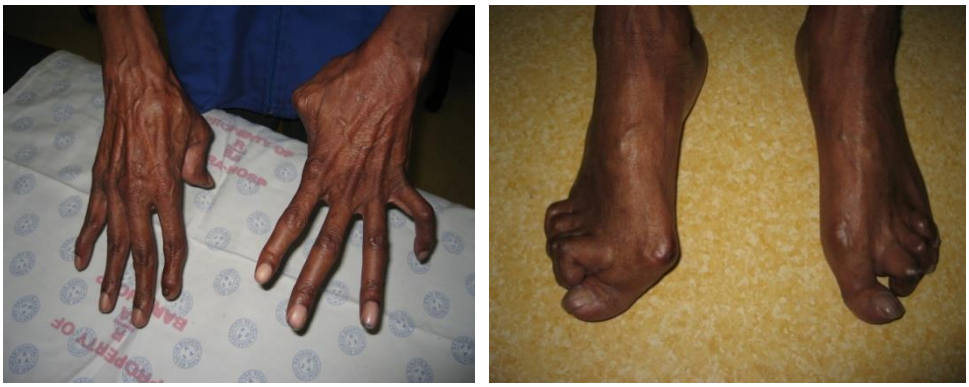

Figure 1.9 Severe deformities of hand and feet of a black South Africans with rheumatoid arthritis.

Interestingly, a proportion of black South Africans present with a unique phenotype of RA (Fig 1.10). Unlike classical RA some black South Africans develop predominately larger joint disease with relative sparing of the small joints of the rest of the hands (Maritz et al., 2005).

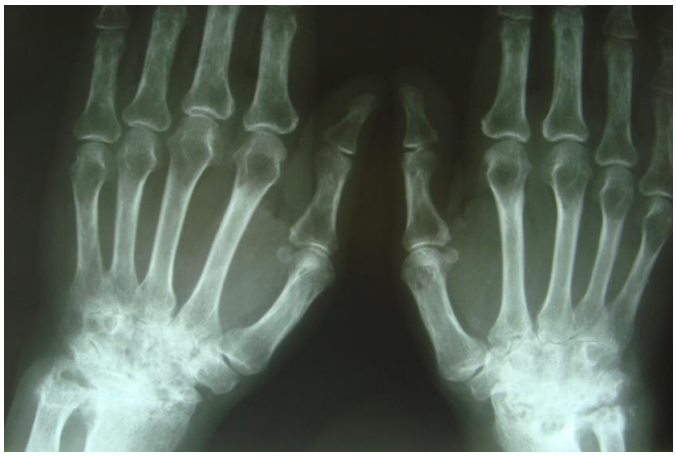

Figure 1.10 Radiography of the hands showing the African variant of rheumatoid arthritis with predominate wrist disease and sparing of the small joints of the hands.

1.6.2.2 Genetics of rheumatoid arthritis in black South Africans

Large scale genetic studies are lacking in African populations with RA. Studies in black South Africans with RA showed a strong association with HLA-DR4 haplotypes, DR1 and DR4 (Martell et al., 1989, Mody et al., 1989, Pile et al., 1992). It is thought that additional risk alleles within the major histocompatibility complex (MHC) exist, but these alleles remain to be pinpointed precisely. The SE still remains the only identified and confirmed genetic risk factor that confers susceptibility for RA across all ethnic groups. Genetic heterogeneity exists within Africans populations. Using low resolution HLA typing, 90% of black South Africans with RA carry at least copy of the SE alleles (Meyer et al., 2011), this is in contrast to only 30% in the Cameroonians of West Africa (Singwe-Ngandeu et al., 2010).

Genetic variants of the non-HLA genes however have failed to show consistency across ethnic groups. Recently a polymorphism in the *PTPN22* gene, R620W was found to be non-polymorphic in black South Africans (Tikly et al., 2010). Furthermore, there are differences in susceptibility for RA of the interleukin 10 (IL10) (MacKay et al., 2003), the corticotrophin releasing hormone (CRH) gene

(Baerwald et al., 2000) and p53 gene (Moodley et al., 2010) between black South Africans and Caucasians.

In black South Africans certain alleles in the HLA region, HLA DRB1 *0401 and *0404, are associated with severe erosive disease (Meyers et al., 2004). Polymorphisms of *IL-1* receptor antagonist have been associated with RA severity in black South Africans (Lubbe et al., 2008).

Rheumatoid arthritis clearly carries a significant health burden in Africa. Despite the progress made in understanding the genetics of RA in Africans, there has been a lack of large scale genetic studies which are warranted in this unique and understudied population.

The objective of the study was to identify genetic contributions to susceptibility for rheumatoid arthritis in a black South African population resident in Soweto.

The objectives of the present research were:

1) To test for known and novel genetic factors associated with RA in black South Africans by dense SNP genotyping using the Immunochip array
2) To investigate the role of specific amino acid positions in the HLA DRB1 molecule as a risk factor for RA in black South Africans

Since these two aims have independent methodologies and discussions, I have decided to present them as two chapters each with its own introduction, methods, results and discussion.

Chapter 2. DENSE GENOTYPING OF RISK LOCI USING THE IMMUNOCHIP SNP ARRAY

2.1 Introduction

Studies of major ethnic groups such as Caucasians and Asians show both shared and ethnic specific RA risk loci. It follows that the same would be found in Africans. However this pattern has very rarely been demonstrated among Africans principally because of the paucity of genetic studies performed in African patients with RA. With the huge burden of communicable diseases in Africa such as infections with the human immune-deficiency virus and tuberculosis, research into non-communicable diseases, amongst them RA, has unfortunately not been given priority. Often the infrastructure for conducting clinical research and recruiting large numbers of patients with accurate phenotype data is lacking. Over the last decade high-throughput genotyping techniques and improved bioinformatics analyses, coupled with large multi centre studies, mainly in Caucasian populations, have led to the discovery of several novel genetic risk loci for RA.

2.1.1 Progression of genetic studies for rheumatoid arthritis

There are two main types of genetic study designs to identify genetic variants associated with diseases, linkage studies and association studies. Linkage is usually the use of families to identify genetic variants segregating with a phenotype, whereas association is a relationship between alleles and a phenotype usually in a case:control design.

2.1.1.1 Linkage studies

Linkage analysis assesses alleles or chromosomal segments that are shared by family members with the same disease. Linkage studies in complex diseases

require no *a priori* hypothesis for the involvement of a particular gene. Thousands of genetic markers are used to scan the genome in sibling pairs. Evidence of linkage is obtained if analysis of large numbers of affected siblings show an increased sharing of haplotypes inherited above the expected under random segregation. Linkage studies of complex diseases have had limited success mainly due to the low power and resolution to detect the modest effect size of variants associated with complex diseases. The limitation with studying siblings is that they have long linkage regions with many genes and identifying the causal variant under a peak is difficult. Therefore a higher density of genetic markers is used to refine linkage peaks. Fourteen candidate regions including genes encoding the CD80 and CD86 molecules, which are involved in antigen specific T-cell recognition, have been identified using linkage studies in RA (Cornelis et al., 1998). In a Japanese study three principal chromosome regions of linkage, D1S253/214, D8S556 and DXS1232, have been identified as RA disease loci (Shiozawa et al., 1998). Another linkage study of multiplex RA families showed association between a number of non-HLA loci on chromosomes 1 (D1S235), 4 (D4S1647), 12 (D12S373), 16 (D16S403), and 17 (D17S1301), and RA and other autoimmune diseases (Jawaheer et al., 2001). Amos et al. studied 642 Caucasian families with RA and described linkage peaks in 2q33 and 11p12 (Amos et al., 2006). This led to the identification of the association with genetic variants of the *STAT4* gene and RA (Remmers et al., 2007).

2.1.1.2 Association testing

2.1.1.2.1 Transmission disequilibrium tests

The transmission disequilibrium test (TDT) was proposed by Spielman, McGinnis and Ewens (Spielman et al., 1993) as a family-based association test to detect the

presence of genetic linkage between a genetic marker and a trait. Although case control association studies have the benefit of identifying genes of modest effect size, the disadvantage is that it is subject to population stratification. Association tests may find association with disease merely due to population stratification and not due to real association. One way of eliminating the problem of population stratification is to use internal controls. The TDT starts with couples who have one or more affected offspring. The families will be included in the analysis studied if they have a parent that is heterozygote for an allele. The allele frequency of the marker locus in random unrelated patients and controls is compared to the frequency in the affected offspring. A Chi squared test is performed and if a specific allele of the suspect SNP is in higher frequency among affected children it is considered a risk allele. The advantage of TDT is that an affected sibling or multiple affected family members are not required, only the parents and the affected individual are needed in the TDT analysis. The disadvantage is that in complex diseases the effect is modest. Hence, a large sample size is required. Therefore many families have to be recruited to detect risk alleles. The TDT has been used to test for association of the *TNF* and *HLA-I* and *HLA-III* genes with RA (Kilding et al., 2004, Mulcahy et al., 1996, Martinez et al., 2000).

2.1.1.2.2 Case-control association testing

Candidate gene association studies rely on *a priori* knowledge about disease aetiology. In contrast, genome wide association studies (GWAS) combine the genomic coverage of linkage analysis with the power of association to have a much better chance of finding complex trait susceptibility variants.

2.1.1.2.3 Candidate gene approach

Prior to the GWAS era, the candidate gene approach was the most widely used study approach. The candidate gene approach studies selected genes with a plausible biological role in the disease of interest or genes that have been shown to have some association in another study. The candidate gene approach has the advantage of showing an association even if the sample size is small and is thus suited for replication studies. In small sample sizes, associations often do not withstand multiple testing corrections typically done for GWAS. The disadvantage though is that the selection of genes depends on prior knowledge of the biology which is not always known and many candidate genes for complex traits are weak candidates with weak prior probabilities. Candidate gene studies have identified important susceptibility loci for RA that have been validated repeatedly such as *PADI4* (Suzuki et al., 2003), *PTPN22* (Begovich et al., 2004, Gregersen, 2005), *CTLA4* (Plenge et al., 2005) and *TRAF1/C5* (Kurreeman et al., 2007).

2.1.1.2.4 Genome wide association studies (GWAS)

Humans inherit genomic regions in blocks called linkage disequilibrium blocks. GWAS and candidate-gene studies use information from high density SNP genotyping and sequencing initiatives such as the HapMap and 1000 Genomes projects to select SNPs that tag other SNPs that are in close linkage with it. In populations of European origin, several hundreds of thousands of SNPs can be used to tag for most of the genome. The tag SNPs may vary according to ethnicity, particularly in African populations that generally have smaller blocks of linkage disequilibrium and therefore would require more SNPs to provide genome-wide coverage (Teo et al., 2010). Unlike the candidate gene approach, GWAS is not hypothesis driven. Association studies are done to test for allele frequency

differences in the diseased compared to the unaffected group. Genome wide association studies have been performed for many diseases (GWAS catalogue, NCBI). Genotyping in GWAS uses array-based formats. The number of SNPs per array has increased dramatically over the last few years.

Genome wide association studies have made marked progress in identifying risk loci in complex diseases such as RA. Genome wide association studies, unlike candidate gene studies, are non-hypothesis driven and have therefore been successful in identifying loci and biological pathways not previously thought to be associated with RA. The first GWAS of European RA patients was performed in 2007 by the Wellcome Trust Case Control Consortium (WTCCC). The study aimed to identify genetic variants in 7 autoimmune diseases (2007). The findings confirmed the prior association of RA with the HLA region and variants of the *PTPN22* gene and discovered 5 novel associations with variants of the *TNFAPI3*, *PRKCQ*, *KIF5A*, *IL2RB* and *AFF3* genes. A number of successful GWAS of RA patients of European decent soon followed the first. The second GWAS of RA found an association with the gene *TRAF1-C5* (Plenge et al., 2007b). The third GWAS published within months of the first showed an association with markers on *6q23* (Plenge et al., 2007a). A Spanish group showed association with the *KLF12* gene (Julia et al., 2008). The following year *REL,* which encodes for members of the NFκB family of transcription factors, was identified as a risk locus for RA (Gregersen et al., 2009). Within just 5 years GWAS have identified more than 35 RA associated loci (Fig 2.1).

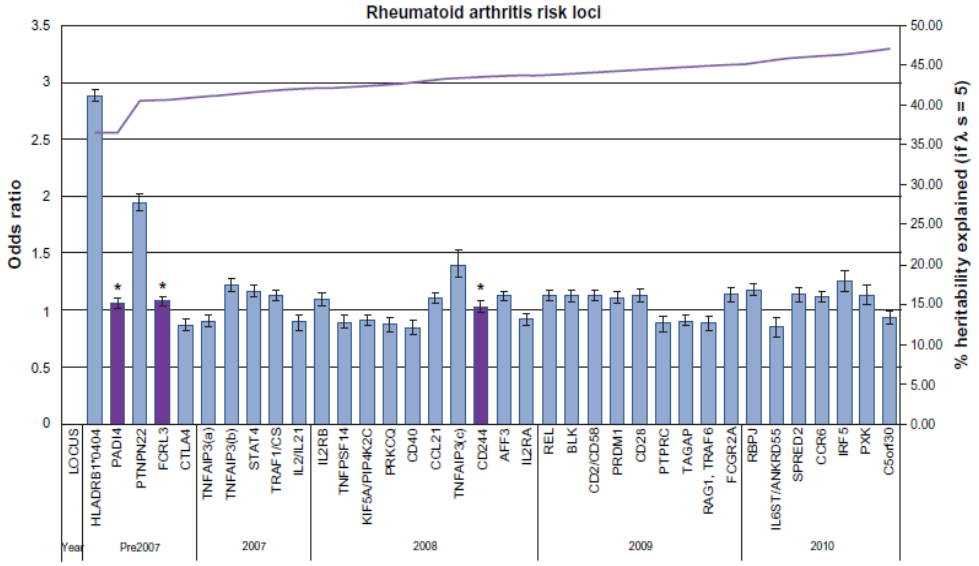

Figure 2.1 Rheumatoid arthritis risk loci identified in Caucasians (McAllister, 2011).

Furthermore, loci common to multiple autoimmune diseases have been identified suggesting common biological pathways. However despite the phenomenal progress made by GWAS in the last 5 years, a large proportion of the heritability of complex diseases including RA remain unexplained. Currently only 51% of the heritability of RA can be explained. The missing heritability can in part be explained by poor capture of unmapped common and rare variants by the current GWAS chips. The current GWAS chips do not cover all common SNPs especially in African populations and even though chips with better coverage are being produced, they will not cover all the variations in the 3.3 billion bases in the human genome. Moreover, the sample size required to detect the modest effects of the loci is ‘massive’ (e.g. >100 000) and with the price of GWAS chips averaging $250/per sample, it would be too expensive to perform a well powered study even if a chip that does cover most of the variation is developed. Another limitation of GWAS is that although numerous RA risk loci have been identified, the causal variant rarely is. These shortfalls can be overcome by fine mapping of the risk loci. This prompted the design of custom made arrays, such as the Immunochip.

2.1.2 Genotyping techniques

The ability to perform large scale genetic studies has improved rapidly over the last decade. Sophisticated high throughput genotyping technology now makes it possible for millions of SNPs to be tested simultaneously and with reduced cost. Depending on the scale of the project, genotyping techniques are chosen accordingly. In general terms low, medium and high throughput genotyping techniques are available and determine the number of SNPs that can be genotyped at once.

The techniques appropriate for candidate gene approachs include a 5' nuclease assay, TaqMan SNP Genotyping Assays (Applied Biosystems) for individual SNPs and for low level multiplexing of up to 10 SNPs the SNaPshot Multiplex Systems (Applied Biosystems) which uses a primer extension based method with detection by capillary electrophoresis. The SNPlex Genotyping System (Applied Biosystems) uses the oligonucleotide ligation assay followed by polymerase chain reaction (PCR) and capillary electrophoresis and can multiplex up to 48 SNPs. The LightTyper system (Roche Applied Science) uses fluorescently labelled oligonucleotides and melting curves analysis to discriminate from individual SNPs. Pyrosequencing is a DNA sequencing technique that is based on the detection of released pyrophosphate (PPi) during DNA synthesis. Flashes of light are released when nucleotides are incorporated. Custom microarrays with thousands of SNPs can be used or the Illumina Golden Gate assay which can test 384-1536 SNPs. The need for denser mapping of the genome prompted the design of BeadChips which can test hundreds of thousands to millions of SNPs. In addition, there are custom chips for specific groups of diseases such as the Metabochip for

cardiovascular and metabolic disorders, and the Immunochip for autoimmune disorders.

2.1.2.1 The Immunochip array

The Immunochip was designed by a world-wide panel of experts in the genetics of autoimmune disease who recognised that these diseases are genetically 'related'. The Immunochip was designed to allow for replication of the top ranked SNPs and to identify pleiotrophic genes which are associated with more than one disease for which the chip was designed. The Immunochip allows for fine mapping of loci identified in immunogenetic studies in order to allow for better coverage of the genetic variants in these regions (Cortes and Brown, 2011). Thus, the Immunochip represents an extraordinary opportunity for fine definition of associated loci across many autoimmune phenotypes, and to compare this information across diseases. The Immunochip is an Illumina Infinium genotyping chip that was designed using European GWAS data from studies of autoimmune diseases. The Immunochip has 196 524 polymorphisms (718 small insertion deletions and 195 806 SNPs) from 186 loci that reached genome wide significance criteria ($p< 5 \times 10^{-8}$) for 12 autoimmune diseases: autoimmune thyroid disease (AITD), ankylosing spondylitis (AS), celiac disease (CeD), Crohn's disease (CD), IgA deficiency, multiple sclerosis (MS), primary biliary cirrhosis (PBC), psoriasis (PS), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), type 1 diabetes (T1D) and ulcerative colitis (UC). The chip was designed by the Wellcome Trust Case Control Consortium (WTCCC). The established risk loci were fine mapped using SNPs from dbSNP and 1000 Genomes Project pilot European population variants (Sept 2009 release) which allows for cost effective fine mapping of these regions. Included on the chip were ancestry informative markers to help with the

identification of population structure. In addition, SNPs associated with non-immunological conditions from the WTCCC2 studies such as Barrett's oesophagus, oesophageal adenocarcinoma, glaucoma, ischaemic stroke, pre-eclampsia, Parkinson's disease, psychosis endophenotypes, schizophrenia, and visceral leishmaniasis were also included. Also included were SNPs related to the genetics of reading and mathematics abilities in children. The Immunochip is also cheaper than GWAS chips ($39/sample) mainly because they are produced in bulk.

Table 2.1 The number of regions per autoimmune disease included in the Immunochip (www.ImmunoBase.org)

Organism	Disease	Region Type	Region Count
Human	Ankylosing Spondylitis	Spondylitis susceptibility regions identified through association studies	7
Human	Celiac Disease	Celiac susceptibility regions identified through association studies	39
Human	Crohn Disease	Crohn susceptibility regions identified through association studies	71
Human	Multiple Sclerosis	MS susceptibility regions identified through association studies	85
Human	Psoriasis	Psoriasis susceptibility regions identified through association studies	37
Human	Rheumatoid Arthritis	RA susceptibility regions identified through association studies	20
Human	Systemic Lupus Erythematosus	SLE susceptibility regions identified through association studies	26
Human	Type 1 Diabetes	T1D susceptibility regions identified through association studies	56

Cytogenetic bands are defined using data from UCSC genome browser (genome.ucsc.edu); protein coding and non-coding genes were extracted from ENSEMBL v67 (www.ensembl.org). Regions were defined by taking the index variant as described in the publication. Starting at the variant of interest, the region is extended out ±0.1cM and this new region was investigated further for variants of genome-wide significance (Table 2.2).

Table 2.2 Description of the 20 rheumatoid arthritis regions included in the Immunochip (www.ImmunoBase.org)

	Region	Location	Markers	Causal Gene Candidate	Immune Diseases
1	1q21.3	GRCh37:chr1:154370938-154666691 (0.30Mb)	rs7543174	IL6R	RA
2	1q24.2	GRCh37:chr1:167161573-167443423 (0.28Mb)	rs840016	CD247	Celiac, RA, SS
3	2p14	GRCh37:chr2:65397595-65717094 (0.32Mb)	rs934734	SPRED2	Celiac, RA
4	2q11.2	GRCh37:chr2:100544267-101041684 (0.50Mb)	rs10865035, rs1160542, rs11676922, rs9653442	AFF3	JRA, RA, T1D
5	3p14.3	GRCh37:chr3:58549647-58601047 (0.05Mb)	rs13315591	PXK	RA, SLE
6	4p15.2	GRCh37:chr4:26028508-26132152 (0.10Mb)	rs10517086, rs874040	RBPJ	RA, T1D
7	4q27	GRCh37:chr4:122938634-123565302 (0.63Mb)	rs13119723, rs13132308, rs6822844	IL2, IL21	Celiac, RA, T1D, UC
8	5q11.2	GRCh37:chr5:55414956-55448425 (0.03Mb)	rs6859219	IL6ST, ANKRD55	Celiac, RA
9	5q21.1	GRCh37:chr5:102034963-102749231 (0.71Mb)	rs26232	C5orf30	RA
10	6q27	GRCh37:chr6:167348507-167547954 (0.20Mb)	rs3093023	CCR6	RA, Vitiligo
11	7q32.1	GRCh37:chr7:128551740-128777520 (0.23Mb)	rs10488631, rs2070197	IRF5	PBC, RA, SLE, SS, UC
12	9p13.3	GRCh37:chr9:34680853-34996014 (0.32Mb)	rs951005	CCL21	RA
13	10p15.1	GRCh37:chr10:6030243-6171924 (0.14Mb)	rs11594656, rs706778, rs7090512	IL2RA	MS, RA, T1D, Vitiligo
14	10p11.22	GRCh37:chr10:31101541-31480324 (0.38Mb)	rs2793108	ZEB1	RA
15	12q24.12	GRCh37:chr12:111716376-113030487 (1.31Mb)	rs3184504, rs653178	SH2B3	Celiac, MS, RA, T1D, Vitiligo
16	14q24.3	GRCh37:chr14:75942923-76030642 (0.09Mb)	rs7155603	BATF	MS, RA
17	16p11.2	GRCh37:chr16:28295306-29032802 (0.74Mb)	rs8045689	CD19, NFATC2IP	RA
18	17q12	GRCh37:chr17:37382674-38240216 (0.86Mb)	rs2290400, rs2872507	IKZF3	Crohn, PBC, RA, T1D, UC
19	21q22.3	GRCh37:chr21:43809418-43878660 (0.07Mb)	rs11203203, rs11209026, rs9988642	UBASH3A	Celiac, Crohn, Psoriasis, RA, Spondylitis, T1D, Vitiligo
20	22q11.21	GRCh37:chr22:21809185-22003928 (0.19Mb)	rs5754217	UBE2L3	Celiac, RA, SLE

The Immunochip has been very successful in identifying new loci for RA and confirming previous associations. Thirteen novel loci have been identified in celiac disease (Trynka et al., 2011), 1 in primary biliary cirrhosis (Juran et al., 2012), 5 in Parkinson's disease (2011), 7 in autoimmune thyroid disease (Cooper et al., 2012), 2 in bipolar mood disorder (Green et al., 2012), and recently an impressive 14 in RA (Eyre et al., 2012). The findings of Eyre et al. bring the total number of non-HLA loci associated with RA to 45 (Fig 2.2) and adds an additional 4% to the estimate of heritability explained (Eyre et al., 2012).

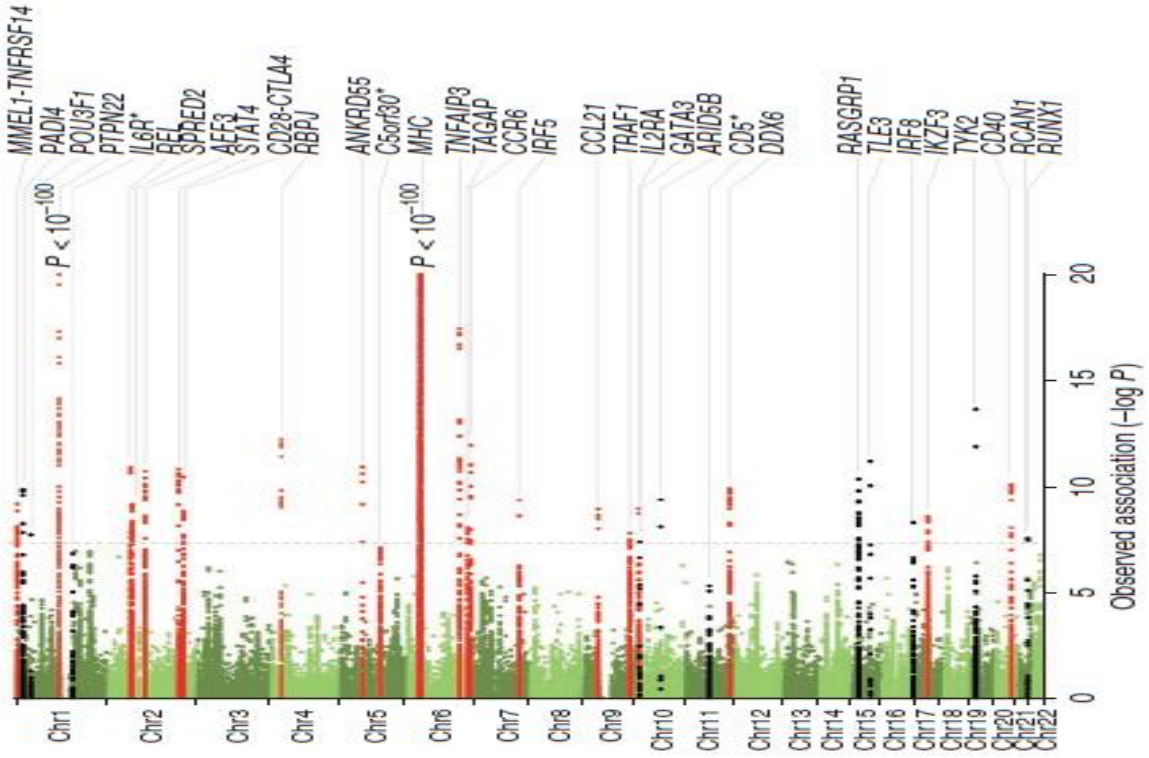

Figure 2.2 New rheumatoid arthritis loci identified using the Immunochip as a genotyping platform (Eyre et al., 2012).

A disadvantage to using the Immunochip is that it was not designed to cover genetic variations such as CNVs and SNPs that are common in non-European (i.e. Africa, in this case) populations since they were chosen based on data from populations of European origin. It is therefore, limited by the power of the original studies from which it was designed. The Immunochip does not have good

coverage of rare variants. Despite these limitations, the advantage of custom made chips such as the Immunochip is that it densely covers common variants of the loci of interest. Immunochips are cheap because they are produced in bulk at $39 per sample compared to $250 per sample for a GWAS genotyping. There are unfortunately no custom chips designed specifically for African populations.

Aims:

1) To provide descriptive statistics (demographic and disease characteristics) for the black South African RA cohort
2) To identify new loci from among the candidate regions on the Immunochip, associated with RA in black South Africans
3) To compare the minor allele frequencies of SNPs previously associated with RA in Caucasians to black South Africans in order to assess their relative roles

2.2 Subjects and Methods

2.2.1 Study participants

2.2.1.1 Cases

Consenting, unrelated black South African RA patients fulfilling the American College of Rheumatology (ACR) 1987 criteria for RA (Arnett et al., 1988) and ≥18 years at disease onset were studied. The cases were consecutively selected and considered ‘black’ African if they self reported all 4 grandparents as being black South Africans. All cases were recruited from a single centre, the Rheumatology Clinic of the Chris Hani Baragwanath Academic Hospital (CHBAH) in Johannesburg, South Africa. The CHBAH is a state sector tertiary hospital serving the urban and peri-urban population of the south western townships (Soweto). Apart from the demographic data, other clinical data such as history of joint surgery, baseline swollen and tender joint count, presence of extra-articular manifestations such as nodulosis were recorded. In addition, activity markers such as ESR, CRP, DAS and SDAI scores were recorded.

Disease activity markers available were swollen joint count (SJC), tender joint count (TJC), DAS28 ESR and the Simplified Disease Activity Index (SDAI) (Aletaha and Smolen, 2005). Severity parameters were the Larsen score, joint replacement surgery (a surrogate for severe joint disease) and the modified health assessment questionnaire (mHAQ) (Pincus et al., 1983).

The DAS28 ESR is a composite score of patient global health (patient self-assessment), tender joint-counts and swollen joint-counts (up to 28), and the ESR (erythrocyte sedimentation rate). It is scored as follows:

DAS28 >5.1 = high disease activity

DAS28 <3.2 = low disease activity

DAS28 <2.6 = remission

The SDAI is the numerical sum of five outcome parameters: tender and swollen joint count (based on a 28-joint assessment), patient and physician global assessment of disease activity [visual analogue scale (VAS) 0–10 cm] and level of C-reactive protein. It is scored as follows:

0.0 – 3.3 = Remission
3.4 – 11.0 = Low activity
11.1 – 26.0 = Moderate activity
26.1- 86.0 = High activity
On a variable number of patients, radiographic scoring for erosions was performed.

2.2.1.2 Control subjects

Control subjects were ethnically and geographically matched to the cases. They were mainly recruited from the staff working at CHBAH and patients attending the Casualty Clinic for minor trauma.
Written consent was attained from each study participant. The study was approved by the Human Research Ethics Committee of the University of the Witwatersrand (M10707) (Appendix 1).

2.2.2 Serological phenotypes

Antibodies and acute phase reactant

Rheumatoid factor (RF) (composite IgM, IgG, IgA) was assayed by nephelometry (Siemens Healthcare Diagnostics, BN Prospec Nephelometer, Newark, USA). The

AntiCCP2 antibodies (ACPA) were measured using a second-generation immunofluorimetric assay with the Immunocap 250 system and reagents and controls provided by the manufacturer (Phadia AB, Uppsala, Sweden). Rheumatoid factor and ACPA levels were considered positive when the concentrations were greater than 15 IU/ml and 10 U/ml, respectively. Levels of CRP >5µg/ml were considered positive as per manufacturers' supplier controls.

2.2.3 DNA extraction and normalisation

The DNA was extracted using the salting-out method that was first described by (Miller et al., 1988). Patient's blood was collected in EDTA tubes to prevent clotting and was stored at -20ºC until it was required for extraction. The blood was thawed prior to proceeding with DNA extraction and thereafter approximately 10ml of blood was transferred to NUNC tubes. The tube was filled with Sucrose-Triton X(S-T X) Lysing buffer, which lyses the blood cells. The tubes were inverted a few times to assist mixing. The tubes were centrifuged for 10 minutes at 2300rpm, thereafter the supernatant was removed and a red pellet remained. The pellet was resuspended in S-T X lysing buffer and placed on ice for 5 minutes. The mixture was then centrifuged again for 5 minutes at 2300rpm. The supernatant was again discarded and the pellet mixed with 1.5ml of T20E5 solution. Thereafter 100µl of 10% SDS and 250µl of Proteinase K was added to the solution. The mixture was incubated for 24 hours, allowing the Proteinase K to digest the remaining protein. After 24 hours 1ml of saturated NaCl solution (approximately 6M) was added. The mixture was vigorously shaken for 15 seconds and then placed on ice for 5 minutes. Thereafter the DNA was transferred to a NUNC tube. Two volumes of absolute ethanol were added to the solution to allow the DNA to precipitate out. DNA was visible as a stringy-like substance, was fished out using a pipette tip, and

then washed in 500µl of 70% ice cold ethanol to remove excess salt. The DNA was air dried and resuspended in 1xTRIS-EDTA buffer. The DNA was stored at 4ºC.

The DNA was quantified using the Nanodrop (ND-1000, Thermo Fischer Scientific Inc.) and diluted using theTecan (Freedom Evo, Tecan Group Ltd.). The DNA samples were normalised to a concentration of 50ng/µl. The quantification of DNA was performed at the Division of Human Genetics, National Health Laboratory Service, Johannesburg, South Africa.

The DNA samples studied belonged to RA patients from an existing registry. It had been stored with the intent for future genetic studies. I extracted and quantified the DNA in preparation for the genotyping process, which was then outsourced to the Feinstein Institute for Medical Research in New York, permitting use of the Immunochip array.

2.2.4 Selection of array and genotyping

An Illumina iSelect HD custom array BeadChip, the Immunochip was the genotyping platform used in this study. Genotyping using the Immunochip is a 3 day long process which was performed at the Feinstein Institute for Medical Research in New York, USA and in accordance with the manufacturer's manual, Infinium®Assay Ultra Protocol Guide (Fig 2.3).

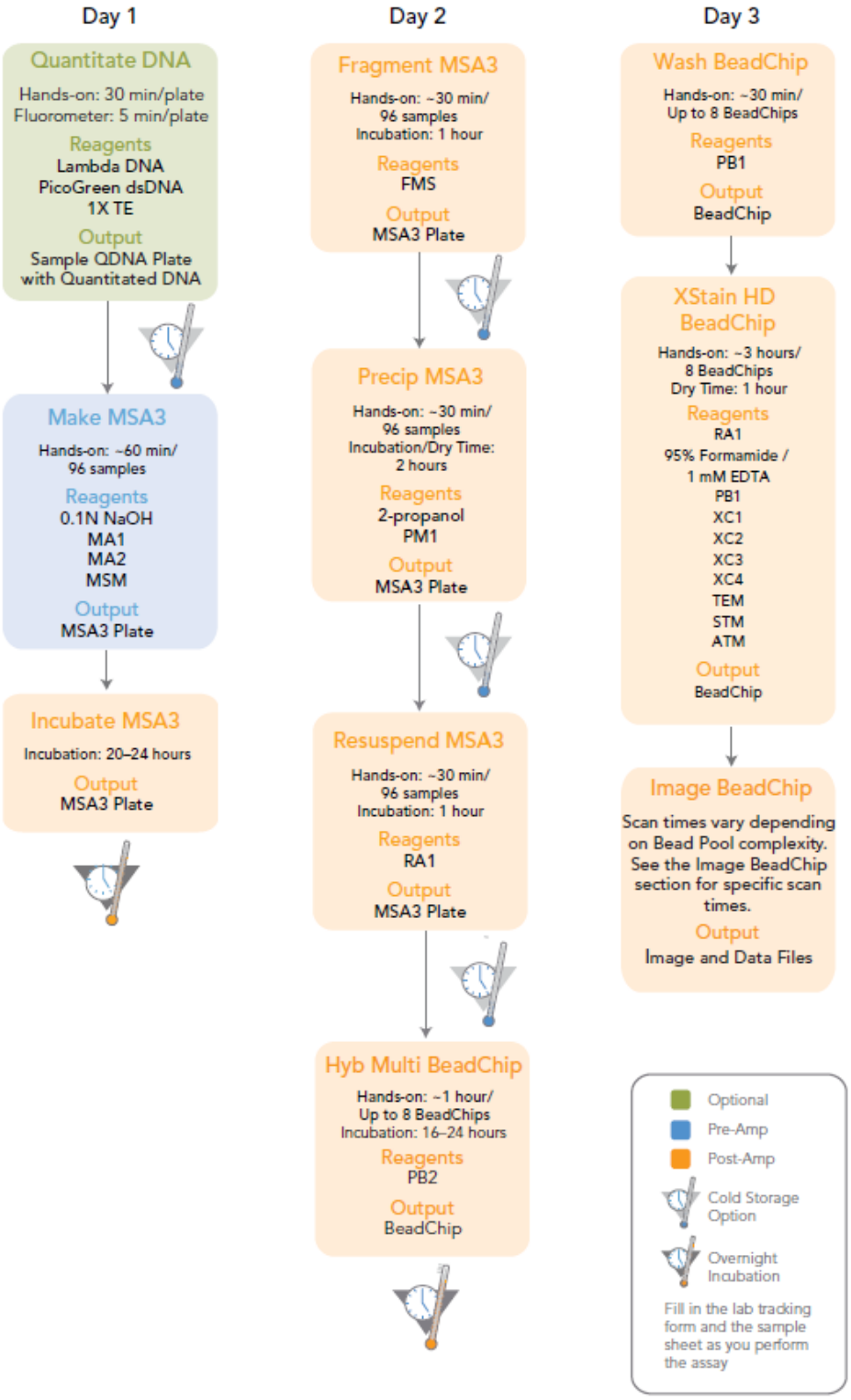

Figure 2.3 Genotyping process (Infinium®Assay Ultra Protocol Guide) using the Immunochip.

The genotyping process involves DNA amplification, fragmentation, precipitation, resuspension, and hybridization to the Immunochips. Single base extension of the primers with labelled nucleotides was followed by imaging of the primers. All reagents were either supplied by Illumina or by the Feinstein Institute. Using a laser the Illumina iScan reader excites the fluor on the single base pairs that extended the primers. The raw optical images created by the emitted fluorescence were obtained. The pixel value of the optical images was used to calculate the signal intensity. Once all the signal intensities were calculated, they were normalised and converted into genotype information. Assigning genotypes was performed using the Genotyping Module (v1.8.4) of the GenomeStudio Data Analysis Software package. Genotype clustering was performed using the default Illumina cluster file Immunochip_Gentrain_June2010.egt and manifest file Immuno_BeadChip_11419691.bpm (NCBI build 36) using the GenTrain2 clustering algorithm. Genotype calling was done using the Genotyping Module of the GenomeStudio Data Analysis Software package. Only the genotype clustering of the significantly associated SNPs were visually inspected. I spent time at the Feinstein Institute for Medical Research in New York and observed the laboratory processes.

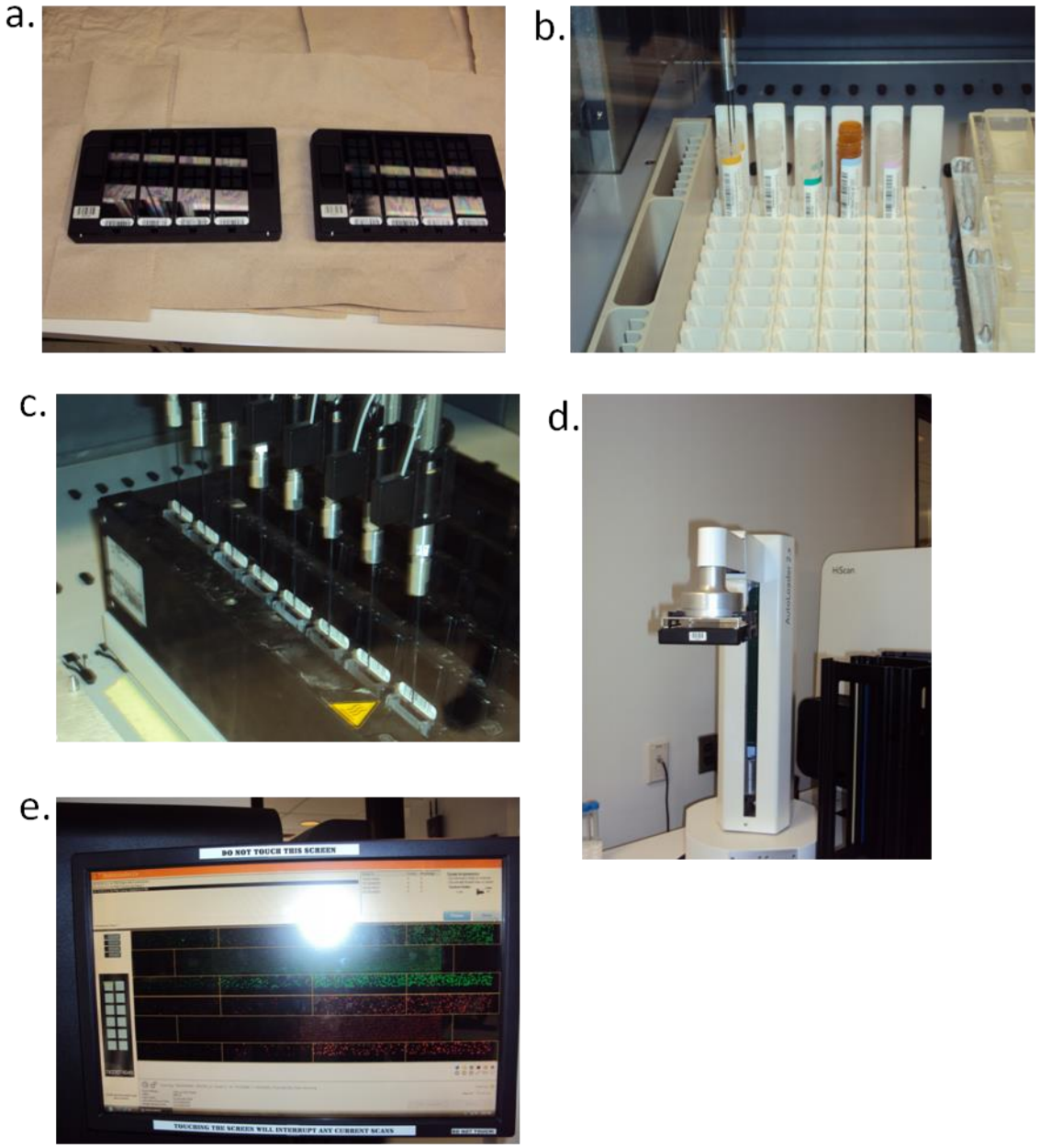

Figure 2.4 Pictures at different stages of the genotyping process**: a)** Immunochips in a Hybridisation chamber **b)** Reagents used during single base pair extension and staining **c)** Chamber rack with reagents being added to Immunochips **d)** Autoloader loading the Immunochips into the iScan system **e)** iScan imaging of Immunochip

2.2.5 Data quality control

Extensive and stringent data quality control (QC) was implemented.

2.2.5.1 SNP quality control

Single nucleotide polymorphisms were excluded if:

1) the per SNP missing genotype call rate was ≥2% in cases and controls
2) they were monomorphic or with a minor allele frequency (MAF) <0.05
3) the GenCall (GC) score <0.15. The GC score is the distance the sample is from the centre of the nearest cluster.
4) they were sex chromosome markers or duplicated markers
5) there was extensive deviation from Hardy-Weinberg equilibrium (HWE) ($p < 5 \times 10^{-7}$) in the controls (Anderson et al., 2010).

2.2.5.2 Sample quality control

Pre-processing of data was performed in which poor performing samples (genotype call rate < 90%) were removed.

Samples were excluded if:

1) The genotype missingness was ≥5%
2) The recorded sex differed from the genotype inferred sex
3) They were duplicated and related. Cryptic relatedness was assessed by estimating identity by state (IBS) statistic. An IBS >0.95 PIHAT <0.05 using PLINK v1.07. One from each pair of related individuals was removed.

2.2.6 Population structure

Using Eigenstat (Price et al., 2006) structure and principal component analyses, plots were constructed based on a comparison with HapMap 3 data (http://hapmap.ncbi.nlm.nih.gov/). Individuals who met either of the two criteria below were excluded from the analysis.

1. Any individuals with a greater than 25% European descent (CEU) ancestry or greater than 30% San ancestry.
2. From the principal component analysis (using the first 5 principal components, weighted by Eigenvalue) we computed the centre of the study group and computed the average distance of each individual to the centre (average 0.0113). Any individual with a distance greater than 0.017 was removed. This cut-off was chosen by plotting the distribution and picking the inflection point.

2.2.7 Allele frequency comparisons between ethnic groups

For comparative purposes, the MAF of SNPs previously identified as risk variants from GWAS of Caucasians with RA (Stahl et al., 2010) was correlated to the MAF in black South Africans and other ethnic groups such Asians and the Yoruba of West Africa. Of the 34 SNPs studied in the meta-analysis, the MAF of 20 SNPs were correlated to the MAF of various ethnic groups. The MAF of these SNPs in other ethnic groups were derived from the 1000 Genomes Project (http://hapmap.ncbi.nlm.nih.gov; www.1000genomes.org). In order to calculate the sample size required in black South Africans to detect the effect size of known RA risk variants in Caucasians, the sample size calculator, QUANTO (http:/hydra.usc.edu/gxe) was used.

2.2.8 Statistical Methods

Once the genotypes were scored and allele frequencies calculated, a Chi-squared test was performed to determine if the difference in allele and genotype frequencies between patients and controls was statistically significant. A corrected value of $p < 5x10^{-8}$ for multiple testing was considered significant for novel RA-associated SNPs and a $p < 0.05$ for SNPs previously associated with RA.

As described previously (Viatte et al., 2012), the Pearson correlation test was applied to determine the correlation of the MAF between the major ethnic groups.

2.3 Results

2.3.1 Demographic and clinical features of rheumatoid arthritis patients

The majority of RA patients were female with established disease. A high percentage of patients were seropositive RF+ (95%) and ACPA+ (89%). Between the cases and controls there was a statistically significant difference in the mean age (p=0.0001) and the female to male ratio (p=0.001) (Table 2.3).

Table 2.3 Demographic and clinical features of the patient and control subjects

Variable	**Cases (n=263)**	**Controls (n=374)**	**P value**
Age at enrolment, mean ±SD years	54.39 (10.84)	41.89 (10.32)	0.0001
Female (%)	235/263 (90)	228/374 (61)	0.001
Disease duration at enrolment, mean ±SD years	8.86 (8.66)		
Smokers (%)	47/230 (20)		
RF+, n (%)	234/248 (94.7)		
ACPA+, n (%)	195/219 (89)		

Although the objective of this study was not to explore associations of genetic variants with severe RA, Table 2.4 has been included to describe the clinical data recorded of the study participants. A high proportion of cases had a high baseline swollen and tender joint count. Less than 1% of cases had joint surgery, nodulosis or serious extra-articular manifestations. The mean modified health assessment questionnaire (mHAQ) was high. The disease activity markers at baseline were very high. The mean disease activity scores (SDAI and DAS ESR) were in the high disease activity range. The Larsen score showed high radiological damage scores. Fifty two percent of cases with reported radiographs had erosive disease.

Table 2.4 Clinical and laboratory variables related to severe rheumatoid arthritis

Variable	
Joint surgery (%)	16/259 (6.18)
Swollen joint count, mean (SD), n=145	9.4 (6.09)
Tender joint count, mean (SD), n=145	12.7 (7.58)
Nodulosis (%)	40/233 (17.16)
Serious extra-articular manifestations (%)	22/256 (8.5)
mHAQ, mean (SD), n=227	1.8 (0.69)
ESR, mean (SD), n=131	47.02 (32.28)
CRP, mean (SD), n=128	29.7 (31.75)
SDAI, mean (SD), n=52	41.8 (15.45)
DAS ESR, mean (SD), n=59	6.3 (1.2)
Larsen score, median (SD), n=60	22 (26.67)
Erosive disease (%)	25/48 (52)

2.3.2 Genotyping and quality control

2.3.2.1 SNP quality control (QC)

A total of 5422 SNPs were excluded from the analysis using a per SNP genotype missingness of ≥2% (Fig 2.5). This followed the initial basic minor allele frequency (MAF) pruning at (MAF<0.01), where 51 556 SNPs were removed; therefore the analysis for missingness started with approximately 140 000 SNPs rather than the initial 196 000. Further stringent QC was performed.

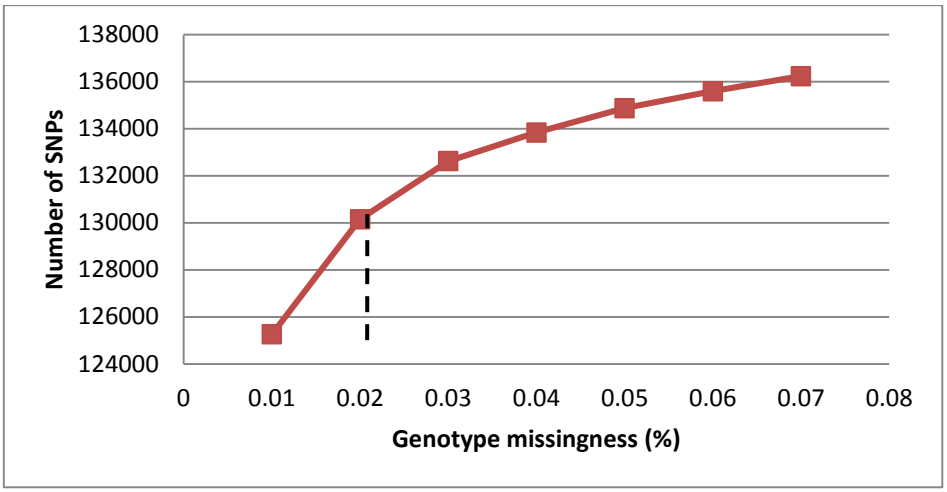

Figure 2.5 Plot of the genotype missingness and the number of SNPs at each missingness threshold.

A total of 6759 SNPs were found to be monomorphic in the data and removed before further analysis. Therefore the analysis for MAF began with a total of 185 918 not the initial 196 000 SNPs. The majority of the SNPs excluded from the analysis were due to low MAF. A total of 75 082 SNPs had a MAF ≤ 0.05 and were excluded from the analysis (Fig 2.6). A MAF ≤0.05 was used because genotype calling of rare variants is not robust in a small sample as in this study.

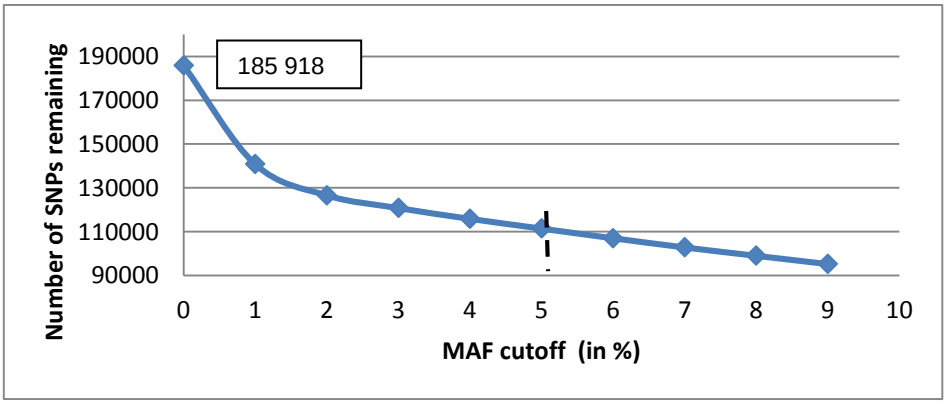

Figure 2.6 Plot of the minor allele frequency (MAF) and the number of SNPs at each MAF threshold.

Only autosomal SNP loci were included in the final analysis, the sex chromosome markers were excluded. After QC, 103 770 SNPs remained in the final analysis.

2.3.2.2 Sample quality control

Twenty one samples were excluded with a genotype missingness of ≥5% (Fig 2.7).

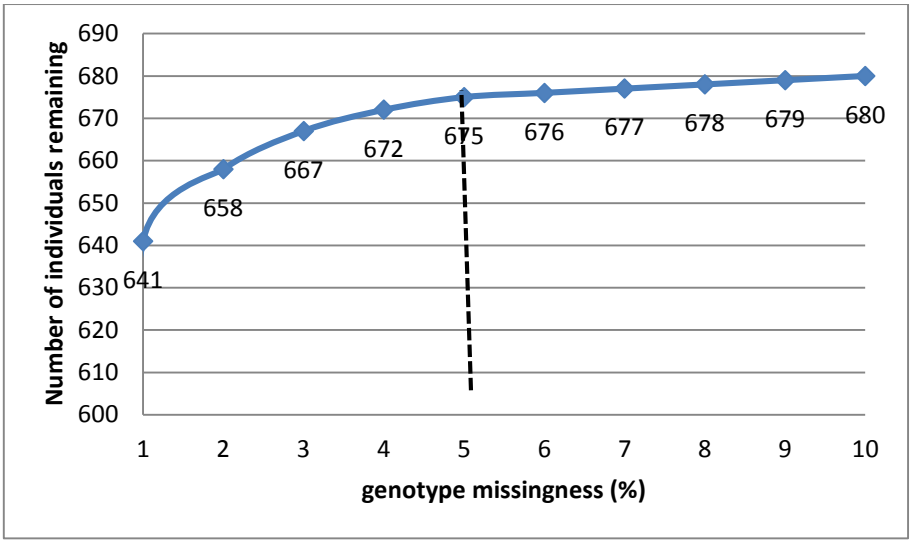

Figure 2.7 Plot of the number of individuals that are retained at each genotype missingness threshold.

There was a higher number of males than females in the control group. We addressed whether gender differences between cases and controls were a confounder in the analysis by testing the minor allele frequency of the significantly associated SNPs in 2 groups, males and females controls. No statistically significant difference was found, suggesting that sex was not a confounder in this study.

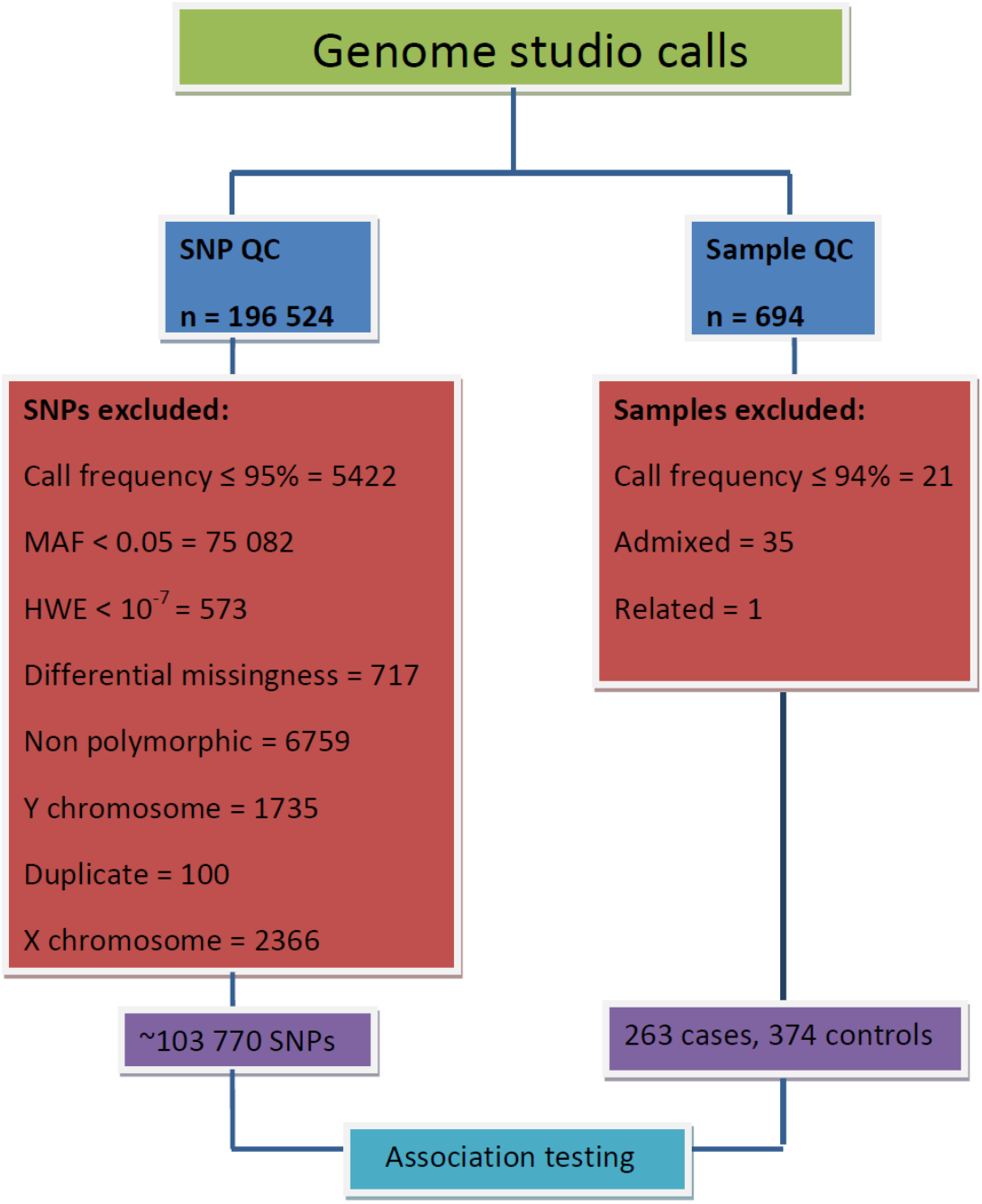

Figure 2.8 Flow diagram of quality control summarising the SNPs and samples that survived the QC procedures.

2.3.3 Population structure

The ancestry informative markers and principle component analyses (PCA) showed a distinction of the black South African RA cases and controls from Caucasians (CEU), West African populations i.e. Yoruba of Nigeria (YRI) and the East Africans, the Luhya (LWK) and Maasia (MKK) tribes of Kenya (Fig 2.9). The

majority of the samples formed a homogenous cluster however, some of the cases and controls showed admixture with 2 other populations encountered in South Africa, namely representative populations for Caucasians and Gujarati Indians, and where excluded from the analysis. Thirty seven samples, 5 controls and 30 cases were excluded based on admixture using the structure analysis.

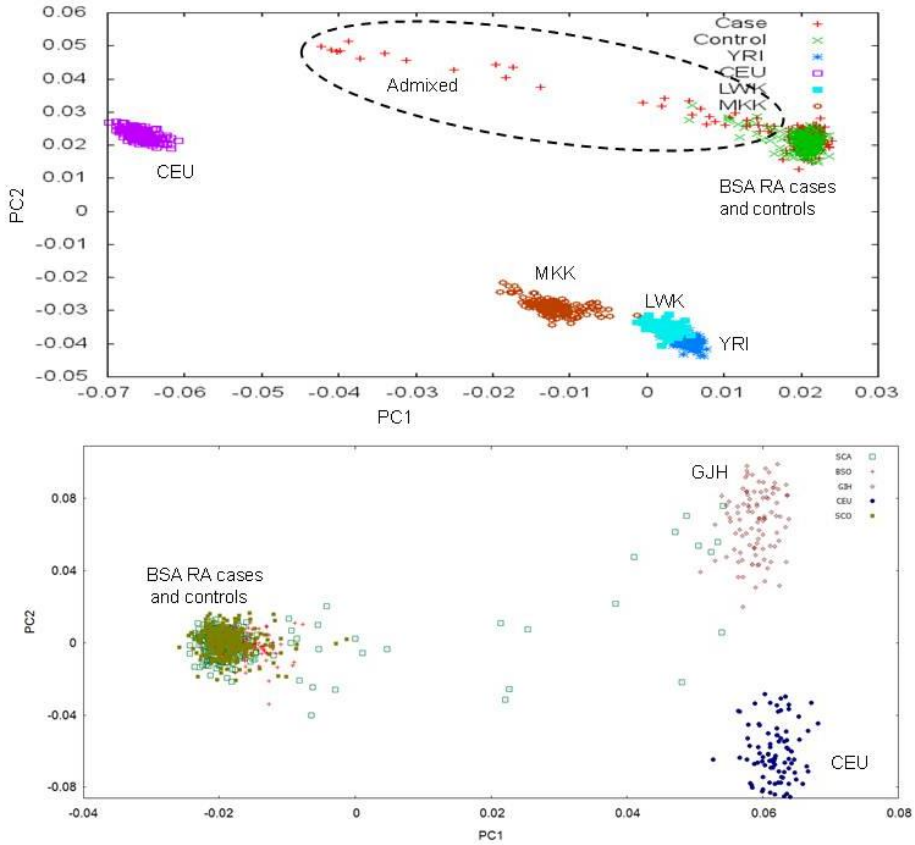

Figure 2.9 Principle component (PC) analyses plot showing (a) the structure of black South African RA cases and controls in relation to CEU, YRI, LWK, MKK populations and (b) admixture with CEU and GJH. The individuals with significant European admixture are circled. A total of 13 355 SNPs were used for the analysis.

In the final analysis 103 770 SNPs were tested for association in 263 cases and 374 controls.

Case control association testing

Quantile plot

The Q-Q (quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each

other. It is used during quality control because a large inflation from the expected p values could point to poor QC. The Q-Q plot shows a strong correlation between the expected and observed p values once the extreme p values of the HLA region were accounted for (Fig 2.10) and only a few SNPs deviate from the expected p values.

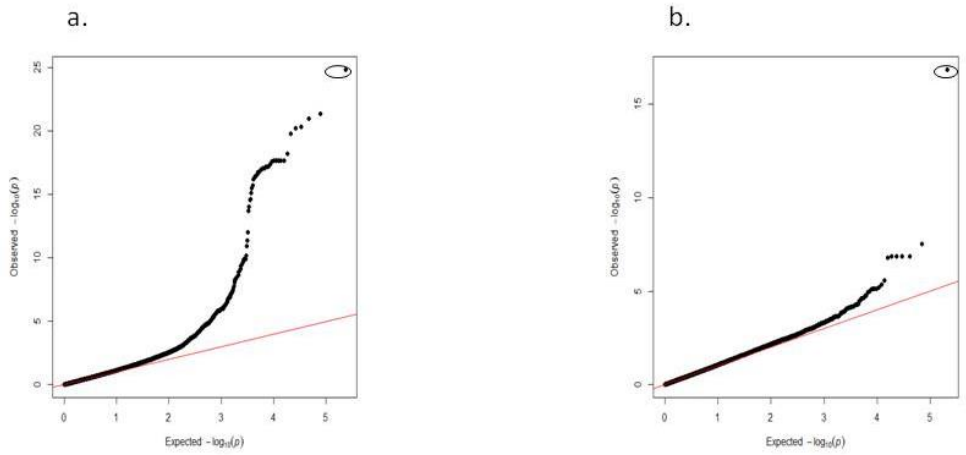

Figure 2.10 A Q-Q plot showing the expected versus the observed p values with (a) and without (b) the p values of the HLA SNPs. The circled SNP did not show robust genotype clustering and was discarded.

2.3.4 Association testing of the overall cohort

In the overall cohort, 88 SNPs were significantly associated with RA. The strongest SNPs associations were found on chromosome 6, most of which localised to the HLA region. A total of 72 SNPs in the HLA region, a further 10 non-SNPs on chromosome 6 and 6 non-HLA SNPs outside of chromosome 6 reached genome wide significance. The significant associations of these regions will be described in more detail in the following sections grouping them into these three categories: HLA, non-HLA on chromosome 6, and non-HLA outside of chromosome 6.

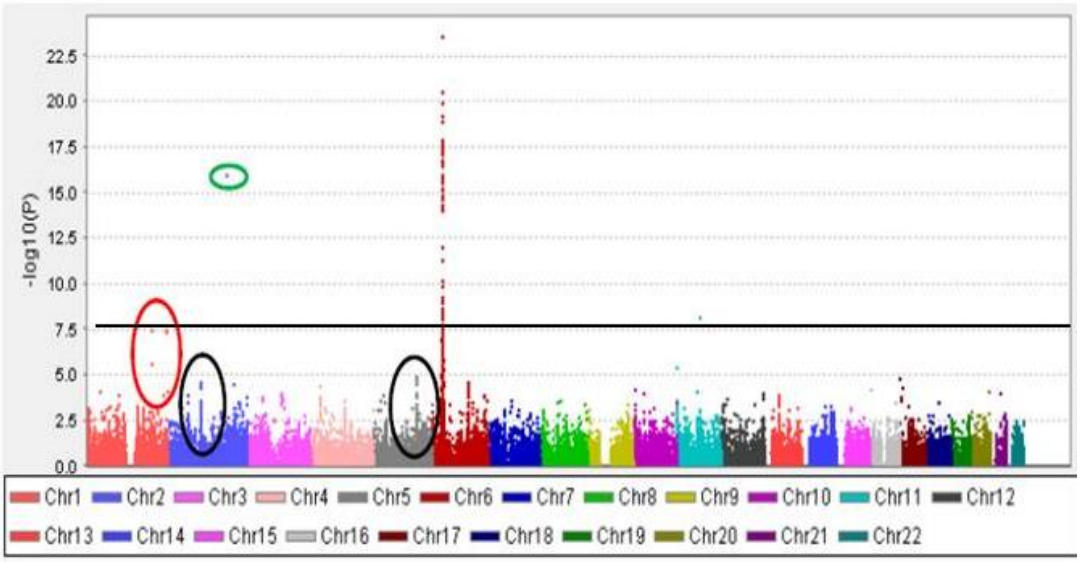

Figure 2.11 A Manhattan plot showing the strength of association between the cases and controls for each SNP in each of the chromosomes.

As shown by the Manhattan plot of the association study (Fig 2.11), a few SNPs on chromosome 1 reached genome wide significance (red circle). A SNP on chromosome 2 did not show robust genotype clustering and was discarded from the analysis (green circle). A few SNPs on chromosome 2 and 5 formed a 'tower' of SNPs but did not reach genome wide significance (black circles).

2.3.4.1 HLA region

A total of 72 HLA SNPs reached statistical significance at the $5x10^{-8}$ level. The strongest association was with a SNP in the intron of *HLA-DR B5* (rs34083746, OR=5.9, $p=2.86x10^{-24}$). Of the SNPs that reached statistical significance, 71 SNPs locate to the *HLA-DR* or *DQ* genes or to the intergenic region between these two genes. Of the above, 4 SNPs in the intergenic region between *HLA-DRB1* and *HLA-DQA1* (rs3104413, OR=3.88, $p=5.49x10^{-21}$; rs3129769, OR=3.91, $p=4.60x10^{-21}$; rs6931277, OR=3.97, $p=1.03x10^{-21}$; rs9272353, OR=4.1, $p=3.01x10^{-21}$) were highly associated. Table 2.5 lists 20 of the most strongly associated SNPs in the HLA region, the rest are listed in supplementary Table 1.

Association testing was also performed on two homogenous subgroups, the RF+ and ACPA+ patients. In the RF+ subgroup the strength of association of the SNPs that were significantly associated with the overall cohort were similar. Of note is the strong association of a SNP in the intron of *HLA-DR B5,* rs34083746, in the subgroups. Cases have 6-fold increased odds of developing RA in the RF+ subgroup and a 7-fold increased odds in the ACPA+ subgroup compared to the odds in healthy controls. Furthermore, SNPs were more strongly associated with the RF+ subgroup than the ACPA+ subgroup.

Table 2.5 Significantly associated SNPs ($p < 5 \times 10^{-8}$) in the HLA region on chromosome 6

SNP	BP position	A2	A1	Overall cohort				RF + subgroup (n-263)			ACPA+ subgroup (n=172)			Gene	Region
				Cases n=263	Controls n=374	P value	OR	MAF cases	P value	OR	MAF cases	P value	OR		
rs34083746	32493589	A	G	0.26	0.06	2.86E-24	5.942	0.27	6.832E-26	6.35	0.30	3.59E-24	7.15	*HLA-DRB5*	intron
rs9272353	32604456	G	C	0.31	0.10	3.02E-21	4.094	0.32	4.414E-22	4.28	0.35	2.62E-20	4.84	*HLA-DRB1|HLA-DQA1*	intergenic
rs6931277	32583357	T	A	0.32	0.11	1.30E-20	3.869	0.33	1.249E-21	4.06	0.35	9.44E-20	4.56	*HLA-DRB1|HLA-DQA1*	intergenic
rs3129769	32597022	G	A	0.31	0.11	5.92E-20	3.809	0.32	6.598E-21	3.99	0.35	2.99E-19	4.51	*HLA-DRB1|HLA-DQA1*	intergenic
rs3104413	32582650	C	G	0.31	0.11	6.95E-20	3.78	0.32	7.669E-21	3.96	0.35	4.69E-19	4.46	*HLA-DRB1|HLA-DQA1*	intergenic
rs1964995	32449411	A	G	0.48	0.23	1.29E-19	2.999	0.48	2.982E-19	3.02	0.35	1.99E-16	3.36	*HLA-DRA|HLA-DRB5*	intergenic
rs3998158	32681992	A	G	0.39	0.17	1.23E-18	3.118	0.40	1.992E-18	3.14	0.45	3.52E-19	3.91	*HLA-DQB1|HLA-DQA2*	intergenic
rs9268838	32428715	G	A	0.40	0.17	1.75E-18	3.094	0.39	6.294E-18	3.09	0.41	2.02E-14	3.26	*HLA-DRA|HLA-DRB5*	intergenic
rs9275428	32670978	A	G	0.53	0.29	2.27E-18	2.8	0.53	1.21E-21	2.7	0.57	2.28E-15	3.18	*HLA-DQB1|HLA-DQA2*	intergenic
rs7764856	32680640	A	T	0.54	0.30	3.48E-18	2.82	0.53	1.59E-16	2.72	0.57	4.29E-15	3.18	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275393	32669439	G	A	0.53	0.29	3.67E-18	2.781	0.52	2.01E-16	2.68	0.56	3.17E-15	3.15	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275407	32670037	C	A	0.53	0.29	3.89E-18	2.777	0.52	2.07E-16	2.68	0.57	2.28E-15	3.18	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275371	32668296	A	G	0.53	0.29	5.00E-18	2.77	0.52	2.68E-16	2.67	0.56	3.99E-15	3.14	*HLA-DQB1|HLA-DQA2*	intergenic
rs9268853	32429643	A	G	0.39	0.17	5.75E-18	3.045	0.40	2.18E-17	3.04	0.41	2.02E-14	3.26	*HLA-DRA|HLA-DRB5*	intergenic
rs9268923	32432835	G	A	0.39	0.17	5.75E-18	3.045	0.40	2.18E-17	3.04	0.41	2.02E-14	3.26	*HLA-DRA|HLA-DRB5*	intergenic
rs2395185	32433167	C	A	0.39	0.17	5.75E-18	3.045	0.39	2.18E-17	3.04	0.41	2.02E-14	3.26	*HLA-DRA|HLA-DRB5*	intergenic
rs9268969	32434349	G	A	0.39	0.17	5.75E-18	3.045	0.39	2.18E-17	3.04	0.41	2.02E-14	3.26	*HLA-DRA|HLA-DRB5*	intergenic

SNP	BP position	A2	A1	Overall cohort				RF + subgroup (n-263)			ACPA+ subgroup (n=172)			Gene	Region
				Cases n=263	Controls n=374	P value	OR	MAF cases	P value	OR	MAF cases	P value	OR		
rs9405108	32438648	G	A	0.39	0.17	5.75E-18	3.045	0.39	2.18E-17	3.04	0.41	2.02E-14	3.26	*HLA-DRA|HLA-DRB5*	intergenic
rs9275390	32669156	A	G	0.53	0.29	6.92E-18	2.756	0.52	3.68E-16	2.66	0.29	5.36E-15	3.13	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275482	32672932	C	A	0.52	0.28	8.70E-18	2.776	0.51	5.83E-16	2.66	0.55	8.61E-15	3.13	*HLA-DQB1|HLA-DQA2*	intergenic

BP= base pair

2.3.4.2 Significantly associated non-HLA related SNPs on chromosome 6

Ten SNPs located to non-HLA genes on chromosome 6 reached genome wide significance. Single nucleotide polymorphisms in the intergenic region between *LOC442175 and ZNF165* (rs149974, OR=2.2, $p=1.64x10^{-08}$), the coding region of *CCHCR1* (rs130071, OR=1.9, $p=8.69x10^{-08}$), the intergenic region between *PSMB9* and *HLA-DMB (*rs241406, OR=7.8, $p=7.9 \times 10^{-10}$) and 7 SNPs locating to the intergenic region between *BTNL2* and *HLA-DRA* were significantly associated (Table 2.6). All 7 SNPs locating to the intergenic regions between *BTNL2* and *HLA-DRA* showed a strong trend toward protection with odds ratios of ~0.5.

Table 2.6 Significantly associated non-HLA SNPs on chromosome 6

Variant	A1 allele	Overall cohort				RF+ subgroup (n=234)			ACPA+ subgroup (n=172)			Gene	Region
		A1 Allele freq cases (n=263)	A1 Allele freq controls (n=374)	P value	OR	A1 Allele freq	OR	P value	A1 Allele freq	OR	P value		
rs241406	C	0.08	0.01	7.98E-10	7.79	0.08	8.21	2.15E-10	0.08	7.69	2.30E-08	*PSMB9* | *HLA-DMB*	intergenic
rs149974	A	0.26	0.14	1.64 E-08	2.22	0.27	2.30	7.91E-09	0.25	2.1	2.23E-05	*LOC442175* | *ZNF165*	intergenic
rs6926737	G	0.34	0.49	4.28E-08	0.53	0.34	0.56	8.75E-07	0.33	0.52	1.35E-05	*BTNL2* | *HLA-DRA*	intergenic
rs5007263	A	0.35	0.51	6.23E-09	0.51	0.36	0.53	6.82E-08	0.34	0.48	1.01E-06	*BTNL2*| *HLA-DRA*	intergenic
rs5007259	A	0.35	0.51	6.23E-09	0.51	0.36	0.53	6.82E-08	0.34	0.48	1.01E-06	*BTNL2*| *HLA-DRA*	intergenic
rs9268507	A	0.35	0.51	8.28E-09	0.51	0.36	0.53	8.82E-08	0.34	0.48	1.21E-06	*BTNL2*| *HLA-DRA*	intergenic
rs6932542	A	0.35	0.51	8.28E-09	0.51	0.36	0.52	5.50E-08	0.34	0.49	1.21E-06	*BTNL2*| *HLA-DRA*	intergenic
rs5007265	A	0.36	0.52	1.01E-08	0.51	0.36	0.53	1.09E-07	0.35	0.49	1.91E-06	*BTNL2*| *HLA-DRA*	intergenic
rs4502931	T	0.35	0.51	3.58E-08	0.53	0.36	0.55	3.43E-07	0.35	0.50	4.24E-06	*BTNL2*| *HLA-DRA*	Intergenic
rs130071	A	0.41	0.27	8.69E-08	1.89	0.41	1.93	8.55E-08	0.41	1.91	1.52E-05	*CCHCR1*	coding

2.3.4.3 Significantly associated non-HLA SNPs outside chromosome 6

Six novel non-HLA SNPs outside of chromosome 6 reached genome wide significance. Five SNPs located to chromosome 1 and one SNP to chromosome 11 (Table 2.7).

The four SNPs on chromosome 1 localising to the intergenic region between *PLD5* and *LOC400*723 were significantly associated with RA in this study (rs2809867, OR=0.33, $p=5.01x10^{-08}$; rs12738883, OR=0.33, $p=3.98x10^{-08}$; rs78352781, OR=0.33, $p=3.98x10^{-08}$; rs12739315, OR=0.33, $p=3.98x10^{-08}$). Three of the 4 SNPs have the same minor allele frequencies in the cases and controls and the same p value suggesting that the SNPs are in near perfect linkage disequilibrium. Furthermore, the two additional significantly associated SNPs locating to chromosome 1(rs2809867, OR=0.34, $p= 4.67x10^{-08}$) and 11(rs35198051, OR=0.31, $p=7.06x10^{-08}$) have similar allele frequencies compared to the 4 earlier described SNPs. In dbSNP there is some ambiguity regarding the SNP on chromosome 11, rs35198051and when the DNA sequence surrounding the “SNP” is aligned to the reference sequence using BLAT, the map locations are not to a single region of the genome, one of the flanking sequences map to chromosome 1. It is therefore highly possible that the SNP on chromosome 11, in view of the same MAF and clustering (Fig 2.13), also locates to chromosome 1.

The SNPs that showed significance in the overall cohort reached similar strengths of association in the RF+ subgroup. However, the association of these SNPs in the ACPA+ subgroup was weaker.

Table 2.7 Significantly associated non-HLA loci identified outside chromosome 6

Chr	Variant	A1	A2	Overall cohort				RF+ subgroup (n= 234)			ACCP+ subgroup (n= 172)			Gene	Region
				A1 frequency control (n=374)	A1 frequency cases (n=263)	P value	OR	A1 frequency	P value	OR	A1 frequency	P	OR		
1	rs12739262	A	G	0.17	0.06	3.73E-08	0.34	0.126	9.10E-08	0.34	0.05	5.75E-06	0.29	*PPP1R12B*	Intron
1	rs2809867	A	G	0.17	0.06	4.67E-08	0.34	0.125	1.12E-07	0.34	0.05	6.55E-06	0.29	*PLD5|LOC400723*	Intergenic
1	rs12738883	A	G	0.17	0.06	3.73E-08	0.34	0.126	9.10E-08	0.34	0.05	5.75E-06	0.29	*PLD5|LOC400723*	Intergenic
1	rs78352781	A	G	0.17	0.06	3.73E-08	0.34	0.126	9.10E-08	0.34	0.05	5.75E-06	0.29	*PLD5|LOC400723*	Intergenic
1	rs12739315	A	G	0.17	0.06	3.73E-08	0.34	0.126	9.10E-08	0.34	0.05	5.75E-06	0.29	*PLD5|LOC400723*	Intergenic
11	rs35198051	A	G	0.17	0.06	7.06E-08	0.31	0.124	1.71E-08	0.31	0.05	2.98E-06	0.27	*C11orf76|LOC100133306*	Intergenic

Chr=chromosome

In order to assess whether the SNPs on chromosome 1 were in LD, a LocusZoom plot was designed to visually inspect the intergenic region between *PLD5* and *LOC400723* using LocusZoom Version 1.1 (http://csg.sph.umich.edu/locuszoom). The software utilizes LD information from HapMap Phase II (CEU, YRI and JPT+CHB) or 1000 Genomes (CEU) and gene information from the UCSC browser. LocusZoom visually displays regional information such as the strength and extent of the association signal relative to genomic position, local linkage disequilibrium and recombination patterns and the positions of genes in the region. The plot (Fig 2.12) showed dense SNP coverage of the region and the 4 significantly associated SNPs. The region has no recombination hot spots.

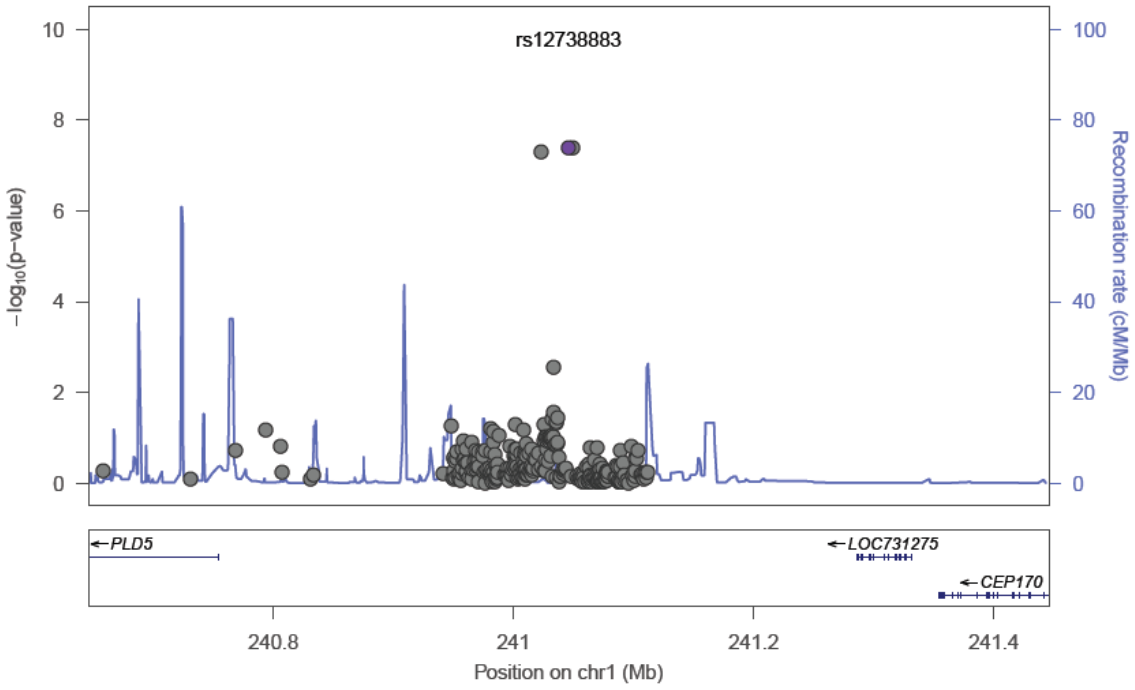

Figure 2.12 LocusZoom plot of the intergenic region between PLD5 and LOC400723 on chromosome 1.

Genotype cluster plots

It is good practice to manually inspect the genotype clusters plots of the significantly associated SNPs. The “SNP genotype” clusters for these 6 SNPs are

not typical, with just 2 parallel clusters representing “genotypes” (Fig 2.13). Similar clustering of these SNPs was observed in African Americans (unpublished data, communications with Dr R. Reynolds, University of Alabama at Birmingham) and in Caucasians (unpublished data, communications with Professor A. Bowcock, Washington University School of Medicine). The SNPs passed the HWE QC of $p<10^{-5}$ in the controls in this study, but not in the two studies by Reynolds and Bowcock. The reason is likely the smaller sample size of the present study, tolerating the absence of a “third genotype cluster”. Interestingly, when examining the allocation of individuals to specific “genotype” clusters, there is complete concordance between the cluster designations across all 6 “SNPs".

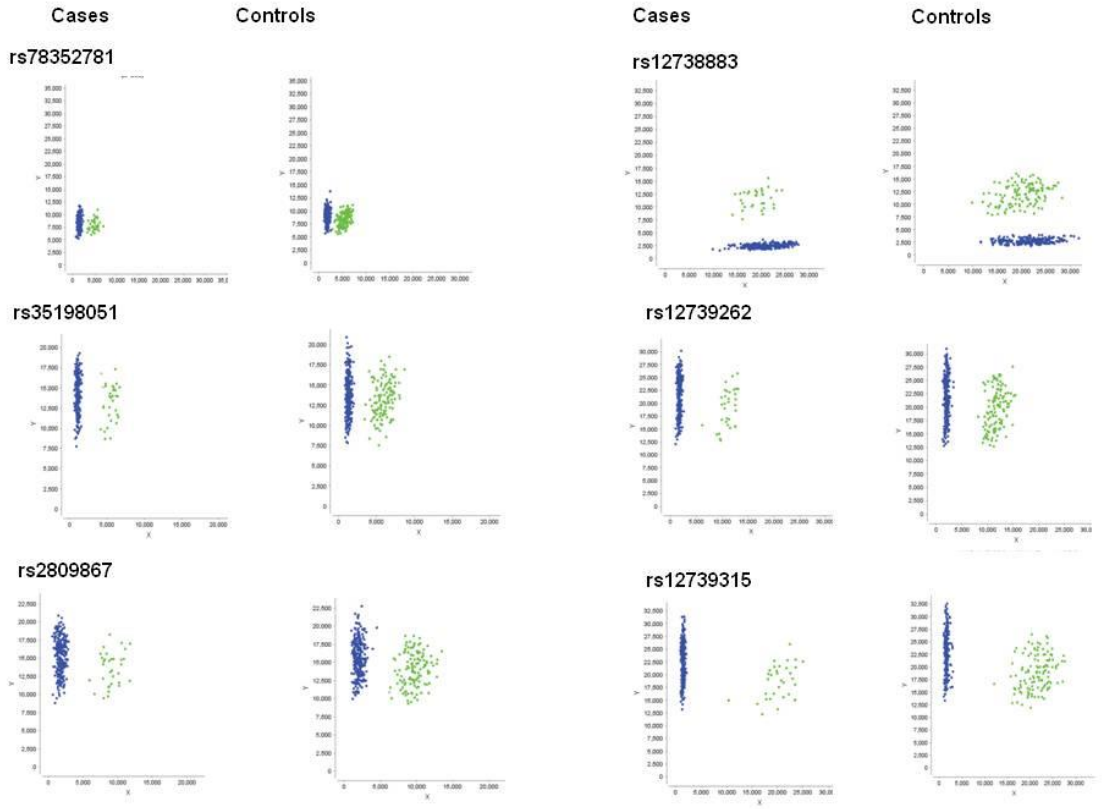

Figure 2.13 Genotype cluster plots of the significantly associated non-HLA SNPs outside of chromosome 6.

Copy number variants

In order to account for the atypical genotype clustering, we considered alternate genetic variants as an explanation. Copy number variants (CNV) are known to produce 'odd' clustering patterns. The Database of Genomic Variants (http://projects.tcag.ca/variation/) was used to identify possible known genetic variants, such as CNVs which could explain the peculiar clustering of the "genotypes". The 4 SNPs in the intergenic region between *PLD5* and *LOC400723* locate to 2 previously described CNVs 107685 and 71544 (Fig 2.14). The coordinates of the CNV 107685 is chr1:242,954,573.242,984,788 (Genome Browsers: UCSC , Ensembl) on chromosome 1q43. The size is 30kb. It was found in a control, J. Craig Venter (Pang et al. (2010) in whom there was an observed gain of the CNV region. The second variation in this region is also a CNV, 71544. The coordinates and the size are similar to the first CNV, chr1:242,954,573..242,984,842 (Genome Browsers: UCSC , Ensembl). This CNV was detected in 7 controls (Conrad et al. (2009)). There was an observed loss of the CNV region.

Together with the similar MAF, it is possible that the "genotype" clusters of the 4 SNPs on chromosome 1 can be explained by a CNV previously described in this region.

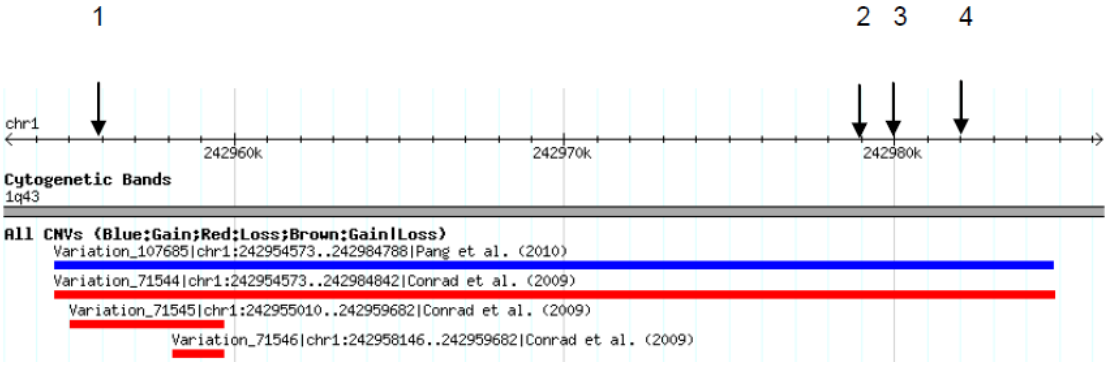

Figure 2.14 The documented copy number variants in the intergenic region between PLD5 and LOC400723 (Database of Genomic Variants [human genome build 37]). The arrows point to the positions of the 4 SNPs that are significantly associated with RA. (SNP,co-ordinates), (1=rs2809867, 2429 56205; 2=rs12738883, 242979398; 3=rs78352781, 242980991; 4=rs12739315, 242982941).

2.3.4.4 Analysis of previously RA associated loci in Caucasians

None of the previous SNPs associated with RA in Caucasians was found to be associated with risk for RA in black South Africans (Fig 2.15).

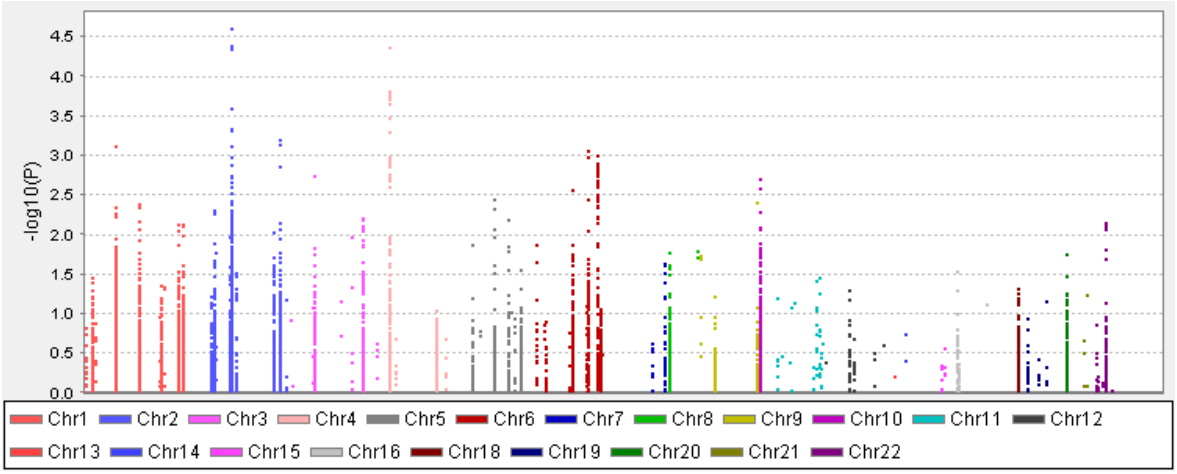

Figure 2.15 Manhattan plot showing the strength of association of SNPs in the RA case-control by only examining SNPs from previously identified RA risk loci in Caucasians.

Of the SNPs previously associated with RA only 1 SNP rs874040 locating to the intergenic region between *LOC389203* and *RBPJ* reached a statistically significant level for replication (p=0.001248, OR=1.446). This SNP locates to chromosome 4p15 and has been associated with RA and type 1diabetes in Caucasians. The MAF in black South African RA cases is 0.53 and 0.43 in controls, whereas in

Caucasians the MAF is 0.34 and 0.33 in cases and controls, respectively. This SNP has a modest effect in Caucasians with RA conferring an OR of 1.18. In this study nine other SNPs in the intergenic region between *LOC389203* and *RBPJ* region reach p values between $p < 5 \times 10^{-5}$ and $p < 5 \times 10^{-4}$ (Table 2.11).

Power calculation

After QC, 263 cases were available for association testing. With this sample size the study was 80% powered to detect an odds ratio of 1.8 if the minor allele frequency is 0.20. The study was not designed on power calculation; the samples used were those available at the time the study was initiated.

An online genetics power calculator QUANTO (http:/hydra.usc.edu/gxe) was used to estimate the power of the study.

Allele frequency comparisons between ethnic groups

The MAF of SNPs previously associated with RA in Caucasians show poor correlation with the black South African controls (Fig 2.17), with a number of SNPs showing significant differences in MAF. In particular, two SNPs in the *ILR2B* gene, rs3218258 and rs3218253, have marked differences between these two populations, 39% versus 6% and 26% versus 8% in Caucasians and black South Africans, respectively. In addition, a SNP in the *CCR6* gene, rs3093023 has a low MAF in black South Africans compared to Caucasians, 6% and 39%, respectively (Table 2.8).

Moreover, based on the effect size observed in Caucasians, this study was sufficiently powered to detect only 2 SNP, rs74377 in the intergenic region between *IL2RB* and *C1QTNF6* and rs2240335 in *PADI4.* For the rest of the SNPs the present study was in most cases underpowered, with samples sizes required

to detect a similar effect size ranging from as little as 447 to as much as 17000 cases.

Table 2.8 Some SNPs previously identified through GWAS as RA risk variants in Caucasians and the minor allele frequency in black South Africans (BSA), Caucasians (CEU), Chinese (CHB) and the Yoruba (YRI)

Chr	dbSNP131	A1 Allele	OR in Caucasians	MAF in CEU	MAF in CHB	MAF in YRI	MAF in BSA Controls	MAF in BSA RA cases	*Sample size required	A2 allele	Gene
4	rs874040	C	1.18 (1.12- 1.24)	0.28	0.05	0.38	0.43	0.52	2645	G	*LOC389203#RBPJ*
9	rs2812378	A	1.10 (1.05-1.16)	0.35	0.06	0.49	0.46	0.44	8354	G	*CCL21#LOC259308*
9	rs3761847	A	1.13 (1.08-1.18)	0.45	0.46	0.61	0.40	0.38	4577	G	*TRAF1#C5*
6	rs3093023	A	1.11 (1.06-1.16)	0.39	0.38	0.08	0.06	0.05	15 714	G	*CCR6*
5	rs6859219	A	0.85 (0.78-0.93)	0.22	0.005	0.21	0.21	0.20	2647	C	*LOC727984#ANKRD55*
2	rs934734	G	1.13 (1.06-1.21)	0.45	0.21	0.38	0.29	0.31	4208	A	*SPRED2#LOC100129850*
6	rs3093023	A	1.11 (1.06-1.16)	0.40	0.38	0.08	0.27	0.28	5797	G	*CCR6*
2	rs3087243	A	0.87 (0.83-0.91)	0.49	0.19	0.20	0.16	0.16	4172	G	*CTLA4#ICOS*
22	rs743777	A	1.72 (1.40-1.11)	0.28	0.11	0.47	0.38	0.39	244	G	*IL2RB#C1QTNF6*
12	rs1678542	C	0.91 (0.87-0.96)	0.62	0.22	0.55	0.44	0.44	7922	G	*KIF5A*
1	rs10919563	A	0.88 (0.82-0.94)	0.13	0.28	0.44	0.39	0.37	4008	G	*PTPRC*
1	rs11586238	G	1.13 (1.07-1.19)	0.24			0.05	0.09	13 822	C	*CD2, CD58*
2	rs1980422	G	1.12 (1.06-1.18)	0.62	0.99	0.99	0.21	0.23	5229	A	*CD28/CTLA4*
9	rs2812378	G	1.10 (1.05-1.16)	0.34	0.28		0.54	0.56	10 319	A	*CCL21*
9	rs3761847	G	1.13 (1.08-1.18)	0.43	0.46	0.59	0.60	0.62	7834	A	*TRAF, C5*
12	rs1678542	G	0.91 (0.87-0.96)	0.38	0.61	0.59	0.56	0.56	10 634	C	*KIF5A*
22	rs3218253	A	1.09 (1.03-1.15)	0.26			0.08	0.07	17 537	G	*IL2RB*
2	rs934734	G	1.13 (1.06-1.21)	0.49			0.29	0.31	4208	A	*SPRED2*
5	rs6859219	A	0.85 (0.78-0.93)	0.21	0.01	1.00	0.21	0.20	2647	C	*ANKRD55/IL6ST*
1	rs2240335	A	1.50 (1.37-1.54)	0.26	0.65	0.37	0.41	0.40	447	C	*PADI4*

*sample size required in black South Africans to detect known RA risk variants, using the effect size observed in Caucasians and the MAF in black South Africans controls.

In contrast to the Caucasians, after excluding just 2 extremely deviated SNPs from the analysis, the MAF of the SNPs studied correlated well between the black South Africans and another African population, notably the Yoruba (Fig 2.18).

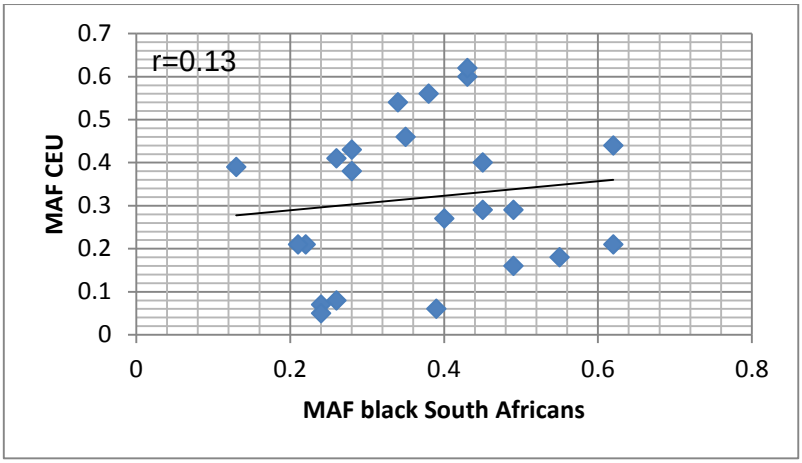

Figure 2.17 Correlation plot of the minor allele frequency (MAF) of SNPs associated with RA in Caucasians and black South Africans.

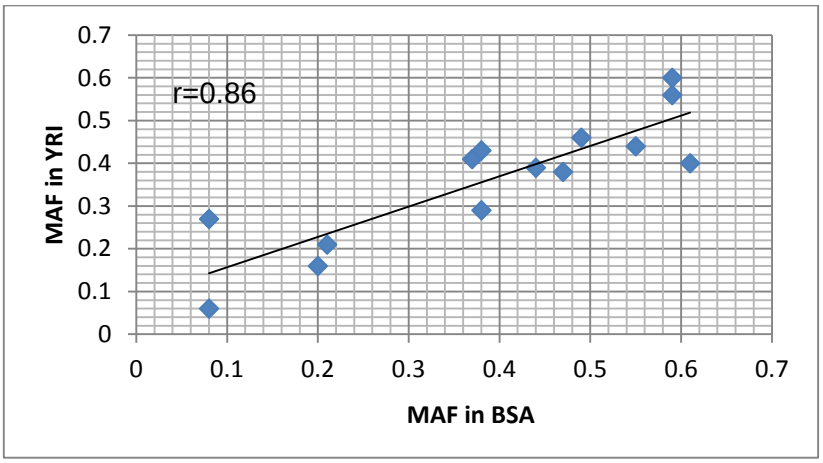

Figure 2.18 Correlation plot of the minor allele frequency (MAF) of SNPs associated with RA in Caucasians in the West African Yuroba populations and black South Africans.

PTPN22

Some of the SNPs in the *PTPN22* gene were monomorphic in black South Africans. For the other SNPs with reasonable allele frequencies in black South Africans, this study was at least 80% powered to detect a number of the known variants associated with RA in this gene. None of these SNPs reached statistical significance (Table 2.9).

Table 2.9 SNPs in the *PTPN22* gene with the odds ratio described in Caucasians and the power to detect significance at a minor allele frequency observed in Black South Africans

dbSNP131	A1	MAF Cases	MAF controls	A2	P	OR	Region	Power
rs3789607	G	0.02	0.02	A	0.7696	0.89	Intron	<0.60
rs12730735	G	0.02	0.02	A	0.7696	0.89	Intron	<0.60
rs76853721	A	0.04	0.02	G	0.008491	2.47	Intron	<0.60
rs12069381	A	0.03	0.04	G	0.3876	0.76	Intron	0.61
rs2358816	T	0.04	0.05	A	0.2136	0.70	Intergenic	0.71
rs17031885	G	0.06	0.06	A	0.9894	1.00	Intron	0.79
rs2257036	C	0.06	0.08	A	0.2546	0.78	Intron	0.88
rs1217412	G	0.06	0.08	A	0.2265	0.77	UTR	0.88
rs1217388	G	0.07	0.08	A	0.2265	0.77	Intron	0.88
rs1217395	G	0.07	0.08	A	0.2265	0.77	Intron	0.88
rs1217407	A	0.07	0.08	G	0.2265	0.77	Intron	0.88
rs1599971	A	0.07	0.08	G	0.2337	0.77	Intron	0.88
rs1217421	G	0.07	0.08	C	0.2164	0.76	Intron	0.88
rs1217389	G	0.07	0.09	A	0.1212	0.72	Intron	0.88
rs3811021	G	0.11	0.10	A	0.6538	1.09	UTR	>0.90
rs6700802	A	0.11	0.10	G	0.8949	1.03	Intron	>0.90
rs3761935	C	0.11	0.10	A	0.5947	1.10	Intron	>0.90
rs3765598	A	0.11	0.10	G	0.5785	1.11	Intron	>0.90
rs1217415	A	0.11	0.12	C	0.3633	0.85	Intron	>0.90
rs2488458	A	0.26	0.24	G	0.4879	1.10	Intron	>0.90
rs6665194	A	0.34	0.32	G	0.4831	1.09	Intergenic	>0.90
rs1235005	G	0.34	0.32	C	0.3872	1.11	Intergenic	>0.90
rs1217385	A	0.35	0.34	C	0.6554	1.06	Intergenic	>0.90
rs7555634	G	0.37	0.35	A	0.3716	1.11	Intron	>0.90
rs2476600	A	0.37	0.35	G	0.3716	1.11	Intron	>0.90
rs2797416	A	0.37	0.35	G	0.3198	1.13	Intron	>0.90
rs1217414	G	0.40	0.38	A	0.3492	1.12	Intron	>0.90
rs1217406	A	0.42	0.38	C	0.102	1.21	Intron	>0.90
rs1310182	A	0.44	0.40	G	0.1848	1.17	Intron	>0.90
rs1217419	A	0.44	0.40	C	0.2043	1.16	Intron	>0.90
rs2476604	A	0.44	0.41	T	0.2255	1.15	Intron	>0.90
rs1970559	A	0.46	0.43	G	0.2955	1.13	Intron	>0.90
rs2476602	G	0.46	0.43	A	0.2863	1.13	Intron	>0.90
rs2476599	G	0.46	0.43	A	0.3056	1.12	Intron	>0.90
rs1217403	A	0.50	0.45	G	0.0815	1.22	Intron	>0.90

*MAF=minor allele frequency, UTR=untranslated

PADI4

This study was underpowered to detect associations with common SNPs in the *PADI4* gene. To detect association with minor allele frequencies of between 0.2

and 0.3 and an OR=1.5, the study was only 70% powered (Fig 2.19). A meta-analysis of RA and *PADI4* studies showed that variants of this gene exert a very modest effect with an OR=1.14 (Iwamoto et al., 2006), therefore this study was underpowered to detect these effects.

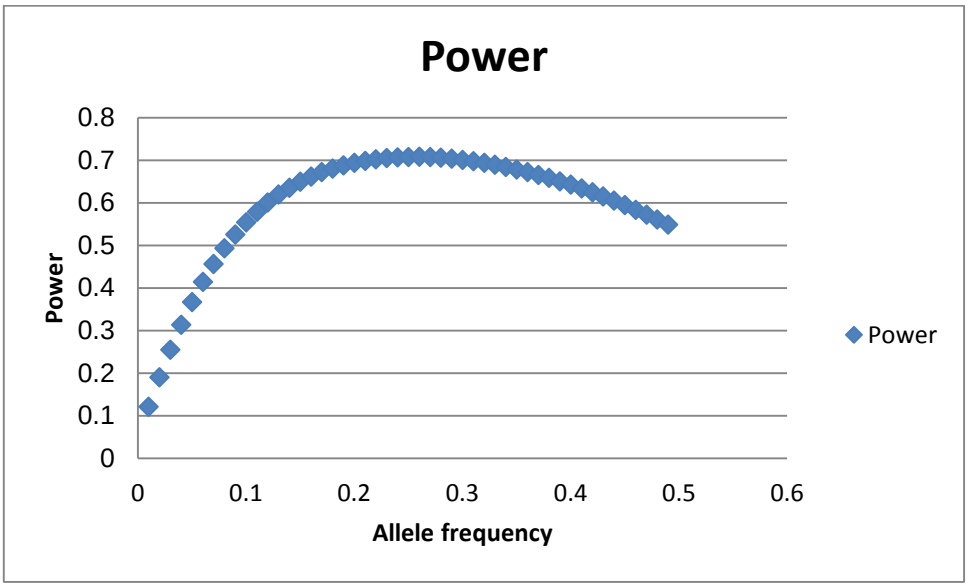

Figure 2.19 A plot showing the power of this study to detect minor allele frequencies in the PADI4 gene given an OR=1.5

2.4 Discussion

This was the first study of RA in Africans to use a high-throughput genotyping technique to identify new loci among candidate regions for autoimmune traits. The Immunochip was able to validate the previously known association with the HLA DRB1 and RA in black South Africans. Four SNPs on chromosome 1 showed significant association with RA and are located within a known CNV region. A potential intergenic CNV on chromosome 1 may be associated with RA in black South Africans and may have a regulatory function. *PTPN22* variants confer the second strongest risk, after the HLA locus, for RA in Europeans, but variants at this locus did not confer risk in black South Africans, if the effect size is similar. In addition, black South Africans were found to be genetically distinct from other African populations.

2.4.1 HLA region

Previous studies in black South Africans with RA showed a strong association with *HLA-DR4* haplotypes, *DR1* and *DR4* (Martell et al., 1989, Mody et al., 1989, Pile et al., 1992). This study validates those findings. As expected, the strongest associations in this study were with SNPs in the HLA class II genes. These SNPs were either located in the introns of the HLA DR or DQ regions or in the intergenic regions between these genes. Whilst the association between HLA DRB1 and RA is well established, the association with HLA DQ region is thought to probably be due to strong LD with the HLA DRB1 gene region and not due to an independent association (Newton et al., 2004). In contrast, few studies in Caucasians have demonstrated an association between HLA DP and RA (Lee et al., 2008, Raychaudhuri et al., 2012, Ding et al., 2009) independent of HLA DRB1. This

study showed no significantly associated SNPs in HLA DP and black South Africans with RA.

2.4.2 Significantly associated non-HLA SNPs on chromosome 6

Ten SNPs reached genome wide significance on chromosome 6 outside of the HLA region. Seven of these SNPs locate to the intergenic region between *BTNL2* and *HLA DRA*. Other significantly associated SNPs locate to the intergenic regions between *LOC442175* and *ZNF165*, *PSMB9* and *HLA-DMB* and to the coding region of *CCHCR1.* Many of these genes have a plausible biological role in RA pathogenesis because of their role in immunity. However, the SNPs locate to the intergenic regions and therefore the true functional roles in association are not known.

2.4.3 Significantly associated non-HLA SNPs outside of chromosome 6

Six novel non-HLA SNPs were found to be associated in this study. Although these SNPs localize to 2 different chromosomes, 1 and 11, in view of the similar MAFs and “genotype” clustering, they are likely poorly annotated and all localize to chromosome 1. Four of these SNPs locate to a CNV (30kb insertion/deletion) region. These are likely pseudo SNPs, but the fact that they are so strongly associated makes them interesting in terms of the association of this region with RA.

Four of the significantly associated SNPs locate to the intergenic region between *LOC400723* and *PLD5.* Interestingly, the *PLD4* gene is associated with RA in Japanese (Okada et al., 2012).

The CNVs, 107685 and 71544, have been implicated in a number of mental retardation and developmental defect phenotypes

(http://genome.ucsc.edu/ENCODE/pilot.html), however there has not been an association with the RA phenotype.

Copy number variants have been rarely described as risk factors for RA and other complex diseases. A large study conducted by the Wellcome Trust Case Control Consortium on 8 common diseases only identified 3 associated CNVs (Craddock et al., 2010). However in the first large analysis of CNVs in a genome wide study of RA, there was a 2-fold higher burden of CNVs in cases than controls (Uddin et al., 2011). Furthermore, 11 rare CNVs were found to be associated with RA, suggesting that CNVs may play a more significant role in RA susceptibility than previously thought. The HapMap data show that the Yuroba population have more CNVs than do Caucasians. This may imply that similar to the Yuroba, black South Africans too may have more CNVs accounting for common complex diseases.

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. Analysis using ENCODE (http://encodeproject.org/ENCODE/) database showed that these SNPs are within regions of predicted repressed or low gene activity (Table 2.10).

From the ENCODE data, there are a number of possible transcription factor binding sites and promoter associated histone marks identified in the CNV region suggesting a possible role of the region in regulation with both repressive and transcription signals. Chromatin immuno-precipitation-sequence data has validated the binding of a number of transcription factors.

Table 2.10 Known ENCODE regions covering the intergenic region between *PLD5* and *LOC400723*

Feature Set	Start (bp)	End (bp)	Feature Type	Feature Type Description
SegmentationState - GM12878 Enriched Sites	24297 9118	24298 2054	Predicted Repressed/Low Activity	Predicted repressed or low activity region
SegmentationState - GM12878 Enriched Sites	24298 2055	24298 2242	Predicted Transcribed Region	Predicted transcribed region
SegmentationState - HeLa-S3 Enriched Sites	24297 9801	24298 0000	Predicted Transcribed Region	Predicted transcribed region
SegmentationState - HeLa-S3 Enriched Sites	24298 0001	24298 1855	Predicted Repressed/Low Activity	Predicted repressed or low activity region
SegmentationState - HeLa-S3 Enriched Sites	24298 1856	24298 1947	Predicted Transcribed Region	Predicted transcribed region
SegmentationState - HeLa-S3 Enriched Sites	24298 1948	24298 2965	Predicted Repressed/Low Activity	Predicted repressed or low activity region
SegmentationState - HUVEC Enriched Sites	24298 1636	24298 1727	Predicted Transcribed Region	Predicted transcribed region

2.4.4 Rheumatoid arthritis risk loci previously identified in Caucasians

Genetic studies have identified numerous RA risk loci in Caucasians bringing the total number of associated loci to 46 (Eyre et al., 2012). The Immunochip array was designed from Caucasian GWAS data. We therefore attempted to determine if loci previously associated with RA and other autoimmune diseases in Caucasians confer risk for RA in black South Africans. It would have been interesting to assess whether loci associated with risk for RA in Asians are associated with RA in this study population; however, Asian specific loci would not have been included on the Immunochip array.

Numerous loci have been associated with RA in Caucasians and Asians. A few of these loci are shared risk loci between these two populations, however many are ethnic specific as shown in the Manhattan plots for 3 different ethnic groups, the Caucasians, Asians and the black South Africans (Fig 2.20).

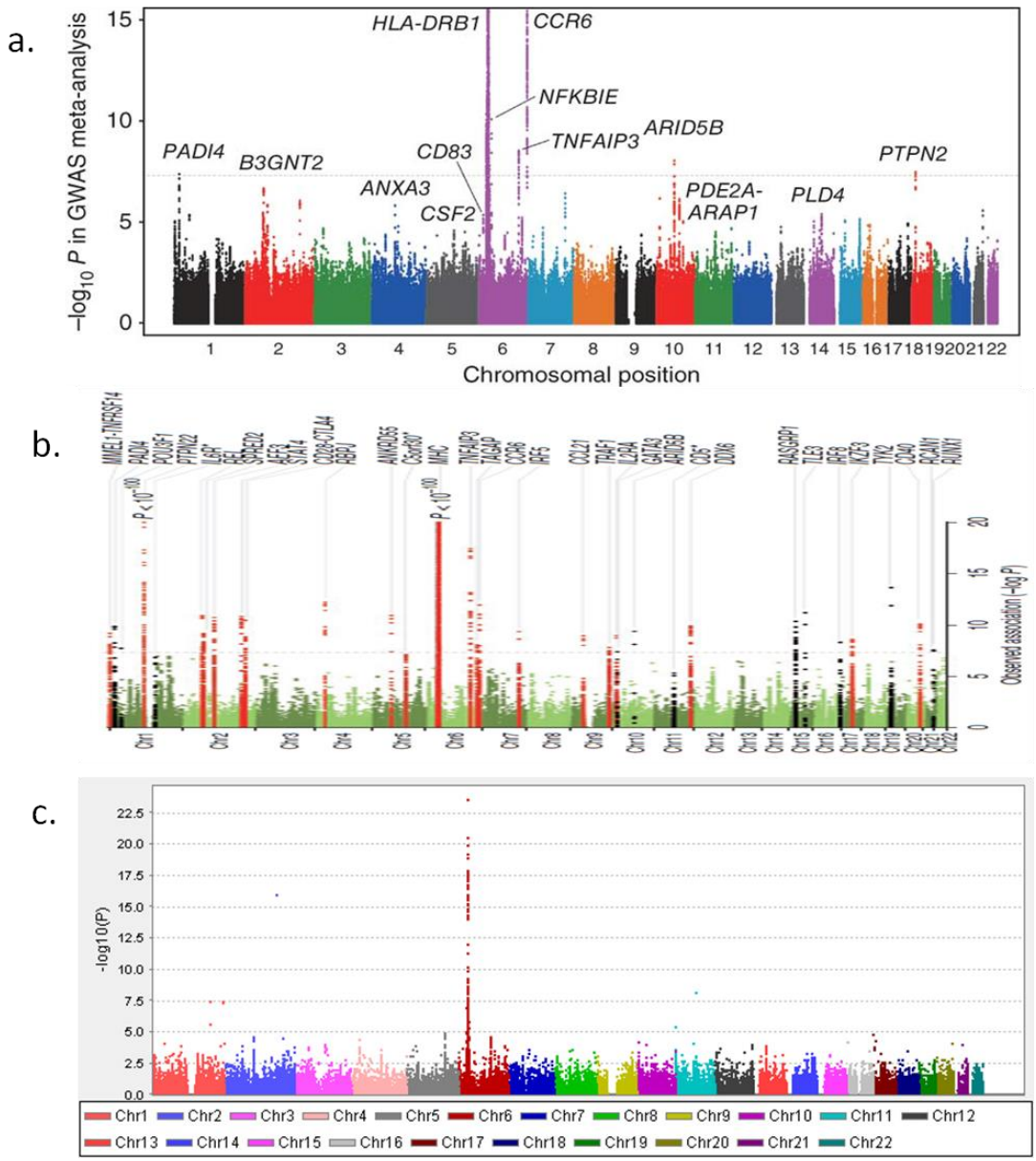

Figure 2.20 Manhattan plot of RA risk loci identified in a) Japanese (Okada et al., 2012) b) Caucasian (Eyre et al., 2012) and c) black South Africans

None of the previously associated non-HLA RA risk loci in Caucasians were associated with RA in black South Africans. The lack of association could be due to these loci being ethnic specific and that these are genetically distinct

populations. This is evidenced by the structure analysis and by the striking differences in allele frequencies of risk variants between the two populations. Many Caucasian RA risk variants are rare or non-polymorphic in black South Africans.

However, the lack of association is more likely due to the study's lack of statistical power to detect the moderate effects that the non-HLA loci confer in complex diseases such as RA.

PTPN22

PTPN22 variants confer the second strongest risk, after the HLA locus, for RA in Caucasians and increase the risk for multiple autoimmune diseases. However, these variants do not confer risk for RA in Asians. The R620W polymorphism was previously shown to be non-polymorphic in black South Africans (Tikly et al., 2010). None of the SNPs associated with the *PTPN22* gene was significantly associated in this study, even though this study was at least 80% powered to detect significance, given the effect sizes observed in Caucasians. There is now sufficient evidence to conclude that the genetic variants of the *PTPN22* gene are not associated with a major effect on RA in black South Africans.

PADI4

A haplotype of peptidyl arginine deiminase 4 (*PADI4*) has been associated with rheumatoid arthritis (RA) in several Asian populations, including Japanese (Suzuki et al., 2003, Ikari et al., 2005) and Korean (Kang et al., 2006), but not with the Han Chinese (Chen et al., 2011) population. Similarly, results are contradictory in Caucasian populations. Most studies in Caucasians show lack of association of genetic variants with the *PADI4* gene and RA (Gandjbakhch et al., 2009, Barton et al., 2004a, Martinez et al., 2005). However, recently an association with RA in

Caucasians was found using the Immunochip (Eyre et al., 2012). We therefore specifically tested for association of this gene in black South Africans. Given the allele frequencies and expected effect size, this study was underpowered to detect associations with common SNPs in the *PADI4* gene.

Replication of Caucasian SNPs in black South Africans with rheumatoid arthritis

Of the SNPs previously associated with RA in Caucasians, 1 SNP on chromosome 4, rs874040 in the intergenic region of *LOC389203* and *RBPJ* reached significance ($p = 0.001248$, OR = 1.44) as defined for replication studies ($p < 0.05$). Although this study was not powered to detect the very modest effect that this SNP confers for RA in Caucasians (OR = 1.14) (Stahl et al., 2010), the strong association indicates that this may be a risk SNP in black South Africans with RA and requires validation. This SNP is close to the *RBPJ* gene which is essential for the Notch pathway. Notch signalling controls numerous cell-fate specification events. The protein encoded for by the *RBPJ* gene is a transcriptional regulator that plays a central role in Notch signalling. It binds specifically to the immunoglobulin kappa-type J segment recombination signal sequence and acts as both a transcriptional repressor and activator (Nakazawa et al., 2001).

Differences of allele frequencies between black South Africans and Caucasians of SNPs associated with rheumatoid arthritis in Caucasians

Of the SNPs associated with RA in Caucasians, the minor allele frequencies of 20 SNPs were compared to that of Asians and different African populations. Interestingly, no correlation was observed between the Asian, Caucasian and black South Africans populations. Many of the studied SNP variants had a low

allele frequency in black South Africans. This suggest that possibly these may not be major risk variants for RA in black South Africans.

Despite the PC plots showing that the Yuroba of West Africa and black South African populations are genetically distinct, there is a tighter correlation of the minor allele frequencies in these 2 populations. The population structure and minor allele frequency analysis findings highlight that African populations are distinct, but more closely related than the Caucasians and Asians who are much more distantly related to Africans as shown in Fig 2.21.

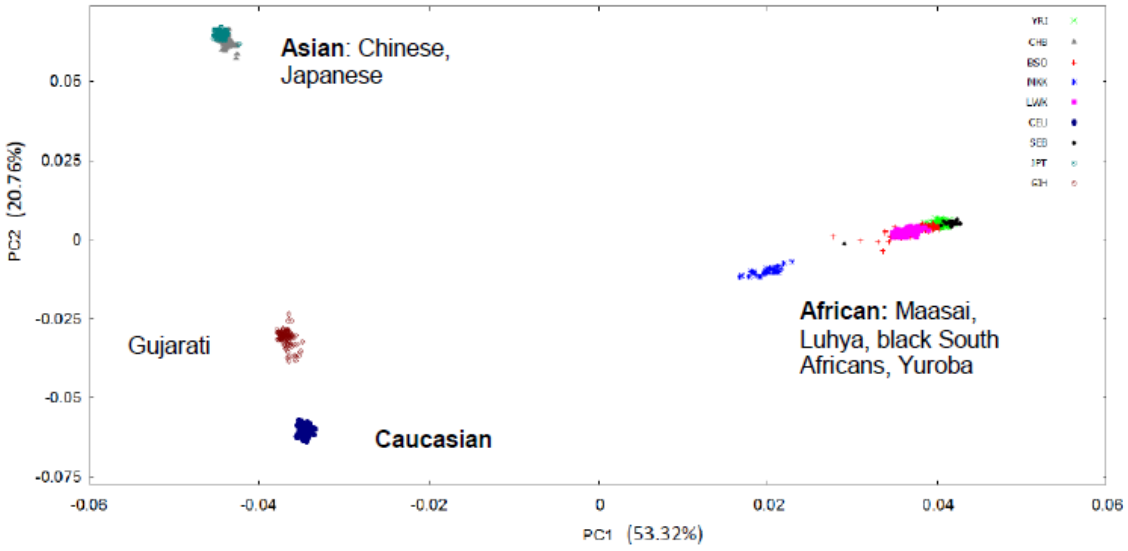

Figure 2.21 Principal component plot showing the genetic relationships between major ethnic groups based on data from 460 568 SNPs (May et al. unpublished).

Further evidence of genetic diversity between Africans is provided by the frequency of carriers of the SE alleles. In black South Africans at least 80% of RA cases carry at least one copy of the SE alleles (Meyer et al., 2011) compared to a much lower frequency of 30% in a Cameroonian population in West Africa (Singwe-Ngandeu et al., 2010).

A significantly high proportion of alleles on the Immunochip were uncommon in black South Africans. A total of 74 558 SNPs had MAF< 5% and a further 6759

SNPs were non-polymorphic, emphasising the differences in genetic structure between black South Africans and other populations. The genetic diversity between Africans and the relatively constant allele frequency amongst different European and Asian populations support the 'out of Africa' hypothesis that led to a severe population bottle-neck and reduced population diversity in the migrating populations (Tishkoff and Kidd, 2004).

2.4.5 Additional non-HLA SNPs of interest but not reaching genome wide significance

On chromosome 2 and 4 of the Manhattan plot of the overall cohort there are 'towers' of SNPs, however none of the individual SNPs reaches genome wide significance but the regions may be of interest for testing in future studies.

The 'tower' on chromosome 2 has six SNPs locating to the intergenic region between *LOC100131131* and *IL1R1 and* 2 SNPs locating to the intron of *IL1R1* on chromosome 2. Nine SNPs on chromosome 4 also form a 'tower' on the Manhattan plot and locate to the intergenic region between *LOC389203* and *RBPJ.* These SNPs reached statistical significance of $p<5x10^{-5}$ and $p<5x10^{-4}$ (Table 2.11).

The protein encoded for by the *IL1R1* gene is a cytokine that belongs to the IL1 receptor family. IL1 receptor is an important mediator of immune and inflammatory responses (Dinarello, 2011). As described earlier, *RBPJ* is a transcription factor that is vital to the Notch signalling pathway that is involved with many cell-cell interactions and cell fate pathways. The *RBPJ* genes have been associated with RA in Caucasians (Stahl et al., 2010).

Table 2.11 SNPs with $p < 5 \times 10^{-5}$ and $p < 5 \times 10^{-4}$ for association with RA in black South Africans

Chr	SNP	A1	A1 frequency Cases	A1 frequency Controls	A2	P value	OR	Gene	Region
2	ccc-2-102110447-A-T	A	0.34	0.46	T	2.51E-05	0.6096	*LOC100131131|IL1R1*	intergenic
2	ccc-2-102113137-G-A	A	0.34	0.46	G	2.51E-05	0.6096	*LOC100131131|IL1R1*	intergenic
2	ccc-2-102129642-G-A	A	0.38	0.50	G	4.14E-05	0.6222	*LOC100131131|IL1R1*	intergenic
4	imm_4_25729576	G	0.48	0.37	A	4.25E-05	1.604	*LOC389203|RBPJ*	intergenic
2	ccc-2-102127970-C-T	A	0.38	0.49	G	4.61E-05	0.6239	*LOC100131131|IL1R1*	intergenic
4	imm_4_25731956	A	0.45	0.35	C	0.000157	1.553	*LOC389203|RBPJ*	intergenic
4	imm_4_25694578	G	0.48	0.37	A	0.000158	1.548	*LOC389203|RBPJ*	intergenic
4	rs10517086	A	0.48	0.38	G	0.000172	1.544	*LOC389203|RBPJ*	intergenic
4	imm_4_25707908	A	0.52	0.41	G	0.000176	1.537	*LOC389203|RBPJ*	intergenic
4	imm_4_25732417	C	0.53	0.42	A	0.000194	1.534	*LOC389203|RBPJ*	intergenic
4	imm_4_25704412	G	0.46	0.36	A	0.000223	1.535	*LOC389203|RBPJ*	intergenic
2	ccc-2-102130269-T-C	G	0.37	0.48	A	0.000261	0.6546	*LOC100131131|IL1R1*	intergenic
2	ccc-2-102131210-C-G	G	0.37	0.48	C	0.000261	0.6546	*LOC100131131|IL1R1*	intergenic
4	imm_4_25695667	G	0.467	0.37	A	0.000342	1.514	*LOC389203|RBPJ*	intergenic
2	ccc-2-102138297-G-C	G	0.38	0.47	C	0.000461	0.6658	*IL1R1*	Intron
2	ccc-2-102102452-A-G	G	0.22	0.32	A	0.000484	0.6275	*IL1R2|LOC100131131*	intergenic
4	imm_4_25727853	A	0.42	0.52	G	0.000512	0.6713	*LOC389203|RBPJ*	intergenic
2	imm_2_204484099	A	0.19	0.27	G	0.000623	0.6207	*CTLA4|ICOS*	intergenic
2	imm_2_204466940	A	0.09	0.15	G	0.000725	0.5357	*CTLA4|ICOS*	intergenic
1	imm_1_67441558	A	0.15	0.22	G	0.00075	0.6019	*IL23R*	Intron
2	imm_2_102446420	G	0.27	0.37	A	0.000766	0.6564	*IL18RAP|SLC9A4*	intergenic
6	imm_6_138316386	A	0.51	0.42	G	0.000866	1.464	*TNFAIP3|PERP*	intergenic
6	imm_6_138316386	A	0.51	0.42	G	0.000866	1.464	*TNFAIP3|PERP*	intergenic
6	imm_6_159371045	A	0.24	0.17	G	0.000995	1.594	*RSPH3|TAGAP*	intergenic
6	imm_6_138314497	A	0.09	0.15	G	0.001034	0.5533	*TNFAIP3|PERP*	intergenic
6	imm_6_138314497	A	0.09	0.15	G	0.001034	0.5533	*TNFAIP3|PERP*	intergenic
2	ccc-2-102145251-G-A	A	0.13	0.08	G	0.001066	1.828	*IL1R1*	Intron

*Chr=chromosome

2.4.6 Subgroup analysis

Specific homogeneous subgroups with complex traits like RA may be useful in establishing specific genetic predisposition. Distinct subgroups in RA are RA+ and ACPA+. In this study there was a slight increase in the strength of association in the RF+ subgroup but not with the ACPA+ subgroup. The stronger association with the RF+ subgroup in black South Africans is supported by a study by Hodkinson et al. (2010) that showed that the diagnostic utility of the ACPA test is no better than RF. It follows that perhaps auto-antigens that trigger RA differ between ethnic groups and that there is a stronger association of ACPA with Caucasians than black South Africans.

2.4.7 Population structure

African populations are the most genetically diverse. Within Africa there are 2000 distinct languages which correlate well with different ethnic groups (http://www.ethnologue.com/region/Africa/***EDITION***). Initiatives such as the HapMap and 1000 Genomes Projects (http://hapmap.ncbi.nlm.nih.gov; www.1000genomes.org) have concentrated on the population structure of the western and central African populations. Much data are therefore available the Yoruba of Nigeria and the Luhya and Maasia populations of Kenya. However, very little is known about the population structure of southern Africans because only a few studies with small sample sizes and few genetic markers have been performed which may not reflect the population structure of black southern Africans well.

One of the contributions of this study is insight into the population structure of a black southern African population which has been largely understudied.

The CHBA Hospital from where the participants of the study were recruited is located in an urban area in Soweto, south of Johannesburg, South Africa. The southern Africa region is defined as the collection of Botswana, Lesotho, Swaziland, Namibia and South Africa (according to the United Nations Geoscheme; http://unstats.un.org/unsd/methods/m49/m49regin.htm). Southern Africa is home to a predominant population of Bantu-speakers; a sub-group of the Niger-Kordofanian (NK) linguistic group that expanded into the region approximately five thousand years (Campbell and Tishkoff, 2010). Specifically, speakers belong to the "S" group of Bantu language classification, consisting of mostly Sotho-Tswana, Venda and Nguni languages (Lane et al., 2002).The genetic architecture of NK-speakers, in general, has been described as fairly homogenous (Tishkoff et al., 2009, Veeramah et al., 2012).

Soweto-Johannesburg is a metropolitan area of the Gauteng province. As one of the urban centres most densely populated by south eastern Bantu-speakers, the area is a suitable microcosm for the larger southern Africa region as a whole. Historically, Soweto attracted migrant labourers from various rural areas that worked in the mines and has therefore been considered the 'melting pot' of ethnic groups. Even today, Soweto remains a major contributor of urbanisation. Soweto is a major contributor to South Africa's leading rates of urbanization (Richter et al., 2009) retaining a regular influx of migrant workers (and refugees) since the gold-mining era, who intermix with local inhabitants.

In this study, participants were recruited if they self reported all 4 grandparents as being 'black' South Africans. The majority of the participants of this study clustered together suggesting that those that migrated to this area are of the same genetic ancestry. The SNPs on the Immunochip were used to examine population

structure, even though they are biased towards SNPs in genes and regions previously associated with autoimmune diseases. There was loose clustering with the cases, suggesting more admixture within the cases. The admixture was found to be with two of the other larger populations in South Africa, the Caucasians and the Indians. The comparative data were from north American Caucasians and from Gujarati Indians living in the USA, so these were used as proxies for the SA equivalents. The tighter clustering in controls compared to cases can be explained by the fact that the controls were predominately the staff of the hospital who live in and around Soweto, however since CHBAH is a tertiary referral centre the cases often live a distance away from the hospital and occasionally even come from neighbouring countries seeking specialised health care in South Africa.

One other major finding is that southern Africans are distinct from other African groups such as the Yoruba, Maasia, and the Luyha. It is clear that tag SNP data from the other African populations cannot be used effectively as a proxy for south African populations. It is now clear that a reference set of markers for southern Africans are necessary. This has implications for future medical and genetic research.

Genetic studies assessing ancestry informative markers, such as in the present study is an advancement from previous studies. A candidate gene approach in which PCA plots can identify admixed samples is thus a valuable quality control step.

2.4.8 Limitations of the study

One of the limitations of the study was that the Immunochip was designed on the basis of Caucasian GWAS data and therefore not ideally suited for genotyping other ethnic groups. Africans differ from Caucasians and Asians in terms of their

low LD structure and therefore many more tagging SNPs are required to cover the African genome. The flip side is that once association is detected, it is likely that the causal variant is relatively close to the associated tag SNP, because of the lower LD. Whereas current arrays cover 94% of the common variation of the European populations, it only covers 81% in Africans (Teo et al., 2010). Using the 1000 Genomes Project data from 90 Yoruba individuals, it is estimated that 1.5 million SNPs will have the same power as 0.6 million SNPs in Caucasians (Teo et al., 2010). Current genotyping platforms have inadequate cover of variation in African genomes.

This study did not address other causes of missing hereditability such as gene-gene interaction, gene-environmental interaction and epigenetic factors. Gene-environmental interactions may account for the differences in the genetic risk for RA between Caucasians and black South Africans. There is evidence that urban black South Africans are exposed to some environmental risk factors that rural blacks are not. This may account for the differences in prevalence of RA between rural and urban black South Africans (Beighton et al., 1975). Gene-gene and gene-environmental interaction studies are still in their infancy and there is a need for future research in this area.

The statistically significant difference in ages between the cases and controls (controls being younger) could have led to potential cases being recruited as controls.

Another limitation of this study was the relatively small sample size and the inadequate power to detect the modest effects of genes that confer susceptibility for RA in black South Africans.

Chapter 3. ASSOCIATION OF SPECIFIC AMINO ACIDS IN VARIOUS POSITIONS IN THE HLA DRB1 PROTEIN WITH RHEUMATOID ARTHRITIS IN BLACK SOUTH AFRICANS

3.1 Introduction

3.1.1 The HLA region

The HLA region has evolved over millions of years to become the master co-ordinator of the immune system. The key role of the HLA complex in regulating the inflammatory response, complement cascade and the immune system is well known (Choo, 2007). The genes in the HLA region encode proteins involved in immune regulatory functions such as ligands, receptors, signalling factors and transcription regulators that are involved in antigen processing and presentation (Shiina et al., 2009). One of the primary functions of the HLA region is to discriminate 'self' from 'non-self' and although highly efficient, breakdown in self tolerance occurs resulting in autoimmune and auto-inflammatory diseases.

The HLA is the human major histocompatibilty complex (MHC); the terms are usually used interchangeably. The HLA region is situated on the short arm of chromosome 6 (6p21.3). The region spans over 3.6Mb and is highly gene-dense comprising over 160 protein-coding genes (Fig 3.1) (Newton et al., 2004). Although the HLA region occupies only 0.1% of the human genome, it contains approximately 0.5% of the known protein coding genes. The HLA region is highly polymorphic and has some of the most dense linkage disequilibrium (LD) regions in the genome. The classical HLA region is divided into 3 sub-regions: the telomeric class I, class III and the centromeric class II region. The class I region has the HLA A, B and C genes, the class II contains the class DQ, DR, DP genes

and the class III is made up of many genes, some of which are involved in the complement cascade and cytokine production.

The classical HLA region was extended to include genes in close LD. This region is now known as the extended MHC (xMHC) and is subdivided into 5 regions, the extended class I, classical class I, class III and classical class II and the extended class II. The xMHC spans 7.6 Mb and consists of 421 genetic loci of which approximately 60% are expressed and 22% have a role in immune function (Fernando et al., 2008).

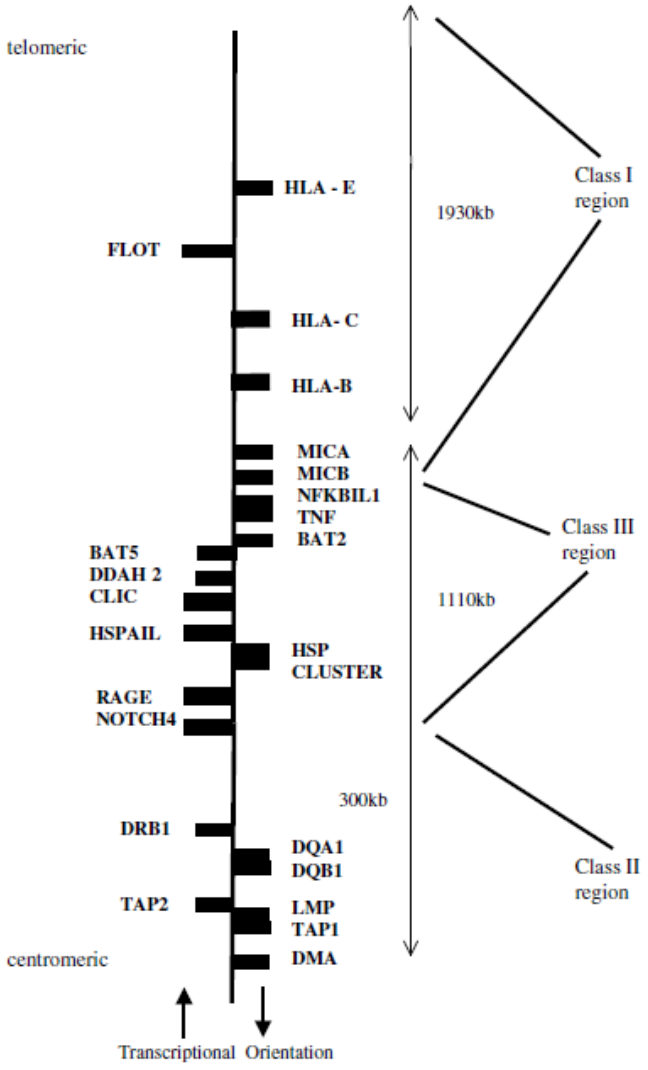

Figure 3.1 Some of the genes within the HLA including the HLA class I, II, III regions (Newton et al., 2004)

3.1.2 The HLA molecules: structure and function

The HLA class I molecule is encoded for by the class I genes. The molecule has 3 α chains and a β2 microglobulin chain encoded by a gene on chromosome 15. The extracellular β2 microglobulin chain is invariant, however the α1 and α2 chains have variable amino acid sequences which determine antigen specificities (Engelhard, 1994). The HLA class II genes have a series of A and B genes which encode the α and β chain, respectively. The HLA DR genes have been associated with numerous autoimmune diseases including RA. There is only one HLA DRA gene and nine HLA DRB genes which encode the invariable α chain and the highly polymorphic β chain, respectively. The HLA DR antigen specificity is determined by the highly polymorphic DRB1 chain which is encoded for by the HLA DRB1 alleles. The α and β chains form a heterodimer comprising α1, α2, β1 and β2 chains. These chains form an antigen binding groove which is highly polymorphic. The peptide binding specificity is determined by a limited number of amino acids in the peptide binding pockets (Strominger, 1987).

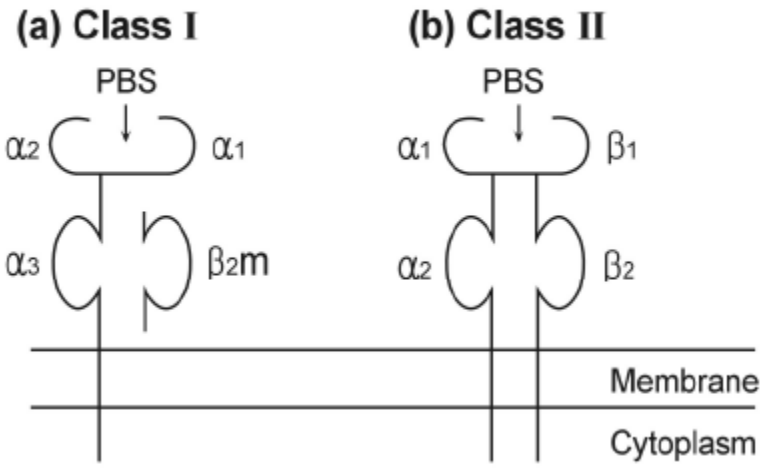

Figure 3.2 Structure of the HLA class I and II molecules with α and β chains and the peptide binding site (PBS) (Choo, 2007).

The HLA class I molecules present mainly endogenous peptides whereas the class II molecule present exogenous peptides (Fig 3.2). Class I molecules present antigen to CD8+ cytotoxic T-cells whereas the class II molecules which are present on APC, such as B-cells, monocytes, macrophages, dendritic cells and Langerhans cells, present antigen to CD4+ T cells.

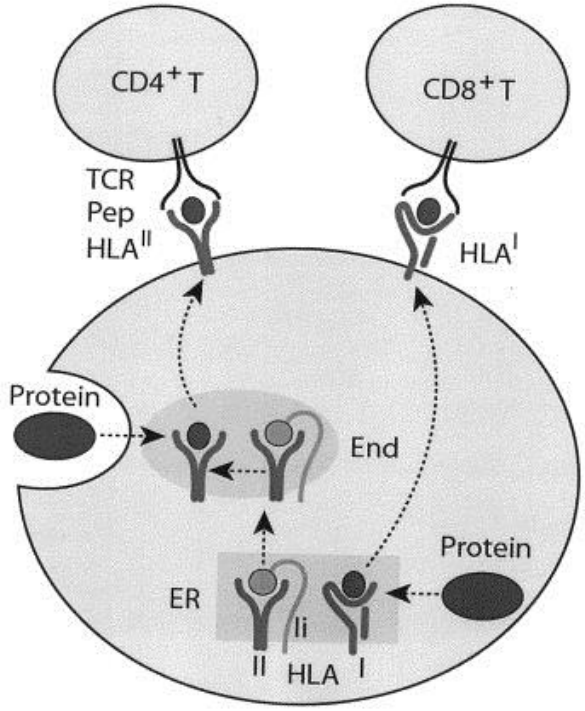

Figure 3.3 The pathways of processing endogenous and exogenous peptides (Pep) and presentation to T-cell receptors (TCR) by HLA molecules (Thorsby et al., 2004).

3.1.3 The role of HLA in rheumatoid arthritis

As discussed in Chapter 1, the heritability of RA is estimated to be 50-60% and the HLA contributes the strongest genetic risk for RA accounting for a third of the total risk. Within the HLA, the HLA DRB1 alleles confer the strongest association. The first association with HLA Dw4 was made by Stastny (1976) when it was found that the frequency of the HLA Dw4 serotype was increased in RA cases compared with healthy controls. With the improvement in resolution of HLA DRB1 genotyping it was discovered that not all HLA DR4 alleles conferred the same risk. In1986, using molecular typing, Gregersen et al. (1987) showed that the risk for RA was

associated with specific alleles in the HLA DRB1 that encode a conserved sequence of amino acids on the third hypervariable region (HVR3) of the HLA DRB1 chain at position 70-74 (^{70}QRRAA^{74}, ^{70}RRRAA^{74}, ^{70}QKRAAA^{74}). This sequence of amino acids is termed the shared epitope and the alleles that they encode are called the SE alleles. It was postulated that this portion of the HLA DRB1 molecule controls susceptibility to disease. The SE alleles are the HLA DRB1 *0401, *0404,* 0405, *0408, *0409, *0410, *1402, *1406, *0101, *0102, *09, *1001 alleles. The SE locates to pocket 4 of the peptide binding groove of the HLA class II molecule and is thought to play a role in antigen presentation (Fig 3.4). The mechanism by which the SE alleles predispose to RA is not fully understood. The hypothesis is that the SE allows for presentation of specific arthogenic peptides (Wucherpfennig and Strominger, 1995) or it selects a specific T-cell repertoire or acts as an immune stimulatory ligand that guides the differentiation of T-cells to Th1 cells (Bhayani and Hedrick, 1991).

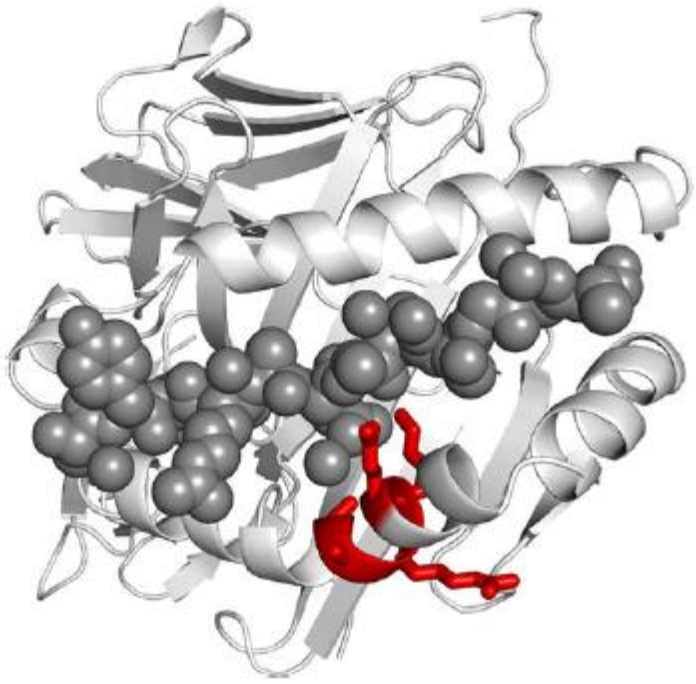

Figure 3.4 The HLA-DR4 molecule (Bax et al., 2011)

3.1.4 The shared epitope and rheumatoid arthritis in different ethnic groups

Different ethnic groups express different SE alleles (Taneja et al., 1992). DRB1 alleles show variable association with RA in different populations: DRB1*0404 in Caucasians, DRB1*0405 in Japanese, *0408 in Mediterranean Caucasians, *1402 in Native Americans, *0401 in North European Caucasians, *0401 in North American Caucasians, *0101 in Israeli Jews and *0404 and *0401 in black South Africans. However in some populations the frequency of the SE alleles are low in RA patients such as in African Americans and the Cameroonians in West Africa (Singwe-Ngandeu et al., 2010). The association of SE alleles is particularly with RF and ACPA seropositive RA.

3.1.5 The shared epitope and other autoimmune diseases

The SE alleles have also been associated with other autoimmune diseases including polymyalgia rheumatica (Weyand et al., 1994), giant cell arteritis (Weyand et al., 1994), Type 1 diabetes (Tait et al., 1995), and erosive joint disease in psoriatic arthritis (Korendowych et al., 2003).

3.1.6 Gene-gene and gene-environmental interaction with the shared epitope alleles

There is a gene-environmental interaction between the SE alleles and smoking and the risk of RA. As discussed in Chapter 1, smokers that are non-carriers of the SE alleles have a 1.5-fold increase in ACPA+ RA. However at the other extreme smokers who carry 2 copies of the SE alleles have a 21-fold increased risk in developing ACPA+ RA (Kallberg et al., 2011).

In addition to the gene-environmental effect, there is some evidence to suggest a gene-gene interaction between HLA DR SE alleles and the R620W polymorphism of the *PTPN22* gene. In individuals who do not carry any of these polymorphic

variants and had never smoked, there is no observed increased risk of RA, however in smokers who are carriers of the SE and the R620W polymorphism, the risk increased 23-fold (Kallberg et al., 2007).

3.1.7 The shared epitiope alleles and rheumatoid arthritis in Africans

Studies in black South Africans with RA showed a strong association with *HLA-DR4* haplotypes, *DR1* and *DR4* (Martell et al., 1989, Mody et al., 1989, Pile et al., 1992). High resolution genotyping has shown that specifically DRB*0401 and *0404 are susceptibility alleles and associated with severe disease in this population (Meyers et al., 2004). Ninety-two percent of black South Africans with RA carry at least one copy of the SE alleles (Meyer et al., 2011). This finding contrasts with that found in the Cameroon, where the SE allele carriers occur at a much lower frequency of approximately 30% (Singwe-Ngandeu et al., 2010).

3.1.8 The amino acid positions of HLA DRB1 and their association with rheumatoid arthritis

Since the landmark discovery of the association of the SE alleles with RA in 1987, it has been found that the SE does not entirely explain the genetic contribution of HLA region to RA. Recently, Raychaudhuri et al. (2012), using sophisticated genotyping and statistical analysis, demonstrated that just 5 amino acids in 3 genes account for most of the genetic risk for seropositive RA in the HLA region. The amino acids at positions 11, 71 and 74 of HLA DRB1, 9 of HLA B and 9 of HLA DPB1 account for most of the HLA region risk in RA. Interestingly, all these amino acids locate to peptide binding grooves (Fig 3.5).

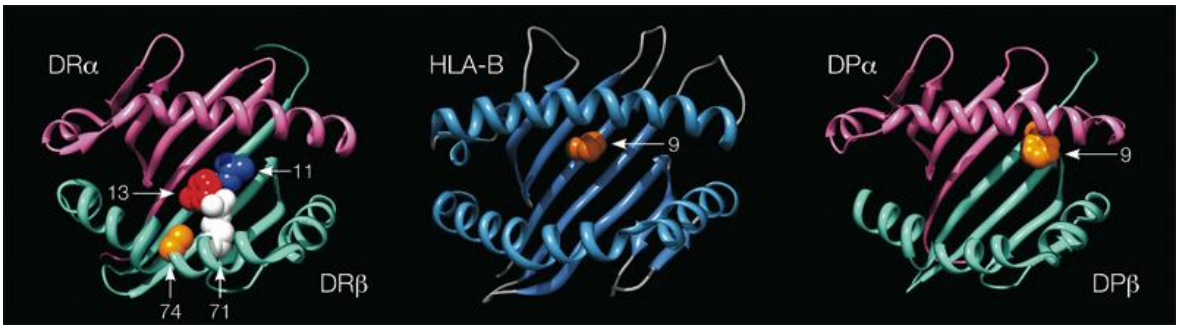

Figure 3.5 Amino acid positions that are significantly associated with RA in the HLA DRB, B and DPB molecules (Raychaudhuri et al., 2012).

The peptide binding groove of HLA DRB1 has 9 pockets that accommodate the side chains of peptides. Larger pockets at positions P1, P4, P6, P9 and smaller ones at P3 and P7 account for most of the polymorphism of the HLA class II molecules (Zavala-Ruiz et al., 2004). Most of the polymorphism of these molecules is in residues along the lining of the pockets of the peptide binding groove, which determine the peptide binding specificity. The SE residues are crucial in determining peptides specific for pocket 4. In addition, only peptides with negatively charged residues (aspartic acid and glutamic acid) at these positions increase the risk for RA.

Amino acids at position 11 of pocket 6 is of the HLA DRB1 molecule which is ~200 base pairs away from the SE was found to be the most strongly associated with RA in Caucasians. Of the 6 potential amino acids at this position, valine conferred the highest risk. Amino acid position 11 is the only variable amino acid in pocket 6 and is thought to be responsible for the antigen specificity. Conversely, serine was found to confer protection for RA (Raychaudhuri et al., 2012).

Aim

1) To investigate the association of specific amino acids at various positions in the HLA DRB1 chain with RA in black South Africans

3.2 Patients and Methods

3.2.1 Patients and controls

Controls and only the subgroup of RF positive and/or ACPA positive RA patients described in Chapter 2 were studied.

3.2.2 HLA DRB1 exon 2 sequencing

Four digit high resolution HLA typing was performed by DNA sequencing of exon 2, using the AlleleSEQR HLA DRB1 reagent kit and protocol (Celera Corporation, Alameda, CA). After polymerase chain reaction amplification of HLA DRB1 exon 2 from genomic DNA, forward and reverse cycle sequencing was performed, and the resulting fragments were collected and analyzed on an ABI 377 automated sequencer (Applied Biosystems, Foster City, CA). An additional sequence reaction was performed to analyse the GTG (valine) motif of codon 86 sequences, thus enabling resolution of ambiguous results for some exon 2 sequences. The sequences were analysed using Assign software (Conexio Genomics, Fremantle, Western Australia, Australia) (Fig 3.6), which enables assignment of genotypes at the four digit level of resolution based on a library file of HLA DRB1 alleles (Cano et al., 2007). This method detects all of the SE-positive alleles.

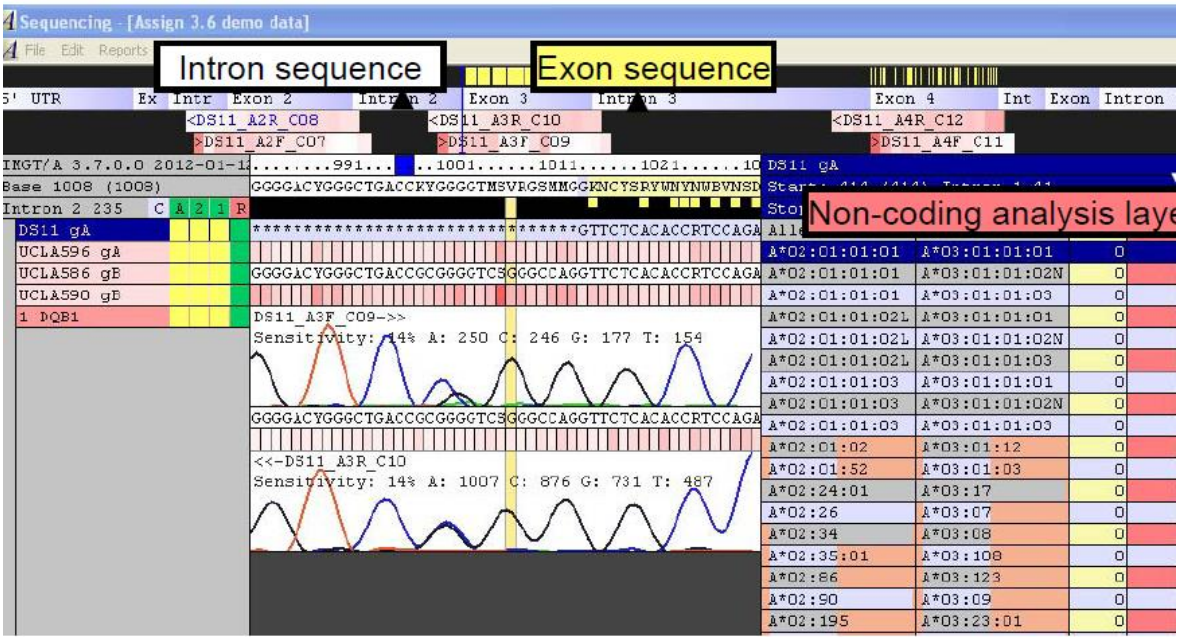

Figure 3.6 Analysis of the HLA DRB1 sequence using the Assign software (Assign 3.6 software manual)

Reference sequences of amino acids from the IMGT/HLA data base (URL: http://www.ebi.ac.uk/ipd/imgt/hla/index.html) were used to infer amino acid sequences corresponding to called four digit alleles (Table 3.1).

Table 3.1 Four digit HLA DRB1 sequence and inferred amino acid

HLA DRB1 allele	**Amino acid at position 11**
*0404,*0405, *0408, *0402, *0403, *0407,*1001	Valine
*1303, *1301, *1302, *1101, *1102, *1103, *1104,*1401	Serine
*0102, *0101, *0103	Leucine
*1601, *1501, *1502	Proline
*0701	Glycine
*0901	Aspartate

3.2.3 Statistics

All possible combinations of amino acid residues were tested for risk of disease. Alleles were converted to binary variables indicating the presence/ absence of a specific amino acid residue at a given position. A series of linear models were fitted to assess whether the joint effect of amino acid residues within each position was a significant source of variation for the presence of RA. The effect of each residue was determined using a logistic regression model and calculating the log-likelihood improvement in model fit over a null model.

Specifically, to determine whether a specific amino acid position, containing “n” residues in our sample, is significantly associated with RA, logistic regression models were fit with the sum of “n-1” indicator variables of the presence/absence of particular amino residues. This was repeated for all positions with polymorphic residues, e.g. valine, serine, leucine. To determine if additional variation was explained by other position after accounting for the most statistically significant positions (e.g., 11, 13), conditional logistic regression was used. In these models the p-value of the test that additional explanatory covariance is explained by a given position was obtained by comparing two models: 1) A base model with only the most significant positions’ residues, and 2) the base model plus additional residues from the position in question. The p-value was determined by considering the difference in residual deviance between two models as c^2 with DF = the difference in number residues between the two models.

3.3 Results

3.3.1 Demographic and autoantibody features of rheumatoid arthritis patients and control participants

As shown Table 3.2, the majority of patients were middle-aged-females and with established disease with a mean symptom duration of 7.2 years. The control participants were from the same ethnic and geographic background, but were younger and there was a higher proportion of males. Not all data were available for all participants, and where this was the case, the numbers are given.

Table 3.2 Demographic and clinical features of the seropositive rheumatoid arthritis patient and control subjects

Variable	**Cases (n=262)**	**Controls (n=362)**
Age at enrolment, mean ±SD years	55.0 (10.8)	41.89 (10.32)
Disease duration at enrolment, mean ±SD years	7.29 (8.9)	
Female (%)	234/262 (89.7)	194/322 (60)
Smokers (%)	47/230 (20)	
RF+, n (%)	234/248 (94.7)	
ACPA+, n (%)	195/217 (90.7)	

3.3.2 HLA DRB1 allele frequencies

Twenty-eight different 4-digit HLA DRB1 alleles in the controls and 26 in the cases were identified, of which 17 alleles occurred at low frequency (<0.05). Of the 9 SE alleles, 6 were identified in this study (Table 3.3). The frequencies of DRB1*0401, *0404, *0405 and *1001, all of which carry the SE motif, were significantly higher in RA patients compared to controls. Conversely, the frequencies of DRB *1101 *1301 and *1302 were significantly lower in the RA patients compared to controls.

Table 3.3 The allele frequencies of the different 4 digit HLA DRB1 alleles in black South Africans controls and RA cases

HLA DRB1 allele	Controls n (allele frequency) n=724	Cases n (allele frequency) n=524	OR (95% CI)	P value
0101	2 (0.0)	1 (0.0)		
0102	30 (0.04)	26 (0.04)		
0201	0 (0.0)	1 (0.0)		
0301	63 (0.08)	28 (0.05)		
0302	73 (0.10)	50 (0.09)		
0401	24 (0.03)	63 (0.12)	4.0 (2.5-6.5)	**<0.0001**
0402	0 (0.0)	1 (0.0)		
0404	15 (0.02)	67 (0.13)	6.9 (3.9-12.2)	**<0.0001**
0405	4 (0.01)	14 (0.02)	4.96 (1.6-15.2)	**0.0018**
0408	0 (0.0)	1 (0.0)		
0410	2 (0.0)	3 (0.0)		
0701	41 (0.06)	30 (0.06)		
0804	30 (0.04)	13 (0.02)		
0901	10 (0.01)	8 (0.02)		
1001	21 (0.02)	27 (0.05)	1.8 (1.0-3.3)	**0.039**
1101	101(0.14)	41 (0.08)	0.5 (0.4-0.8)	**0.0008**
1102	36 (0.05)	12 (0.02)		
1114	1 (0.0)	2 (0.0)		
1201	28 (0.04)	19 (0.04)		
1202	4 (0.0)	1 (0.0)		
1301	92 (0.13)	40 (0.08)	0.6 (0.4-0.8)	**0.004**
1302	47 (0.06)	21 (0.04)	0.6 (0.4-1.0)	**0.06**
1303	14 (0.02)	6 (0.01)		
1336	1 (0.0)	0 (0.0)		
1340	1 (0.0)	0 (0.0)		
1401	8 (0.01)	2 (0)		
1404	0 (0.0)	1 (0.0)		
1417	0 (0.0)	2 (0.0)		
1501	14 (0.02)	10 (0.02)		
1503	61 (0.08)	32 (0.06)		
1506	1 (0.0)	0 (0.0)		

3.3.3 Specific amino acid associations

Of the 29 amino acid positions examined amino acid position 11 (p = 3.4E-26), 13 (p = 1.2E-27), and 33 (p=2.1E-28) were the most highly associated with RA in black South Africans (Fig 3.7). The shared epitope position 71 was less

significantly associated (p=2.6E-05) whilst position 74 was not associated (p=0.15).

After conditioning on position 11, the effects at the other positions were diminished (pos 13: p = 0.01; pos 33: p= 0.004) (Table 3.4), suggesting that in this sample amino acid position 11 was the primary contributor to the association with HLA DRB1.

Table 3.4 The amino acid positions and p values of the strength of association before and after conditioning on the most highly associated amino acid positions 11

Amino acid position	**p value**	**p value after conditioning on position 11**
33	2.14E-28	0.0043
13	1.16E-27	0.015
11	3.39E-26	N/A
12	7.74E-17	N/A
10	7.88E-16	0.004
47	8.30E-14	0.35
67	3.11E-13	0.27
70	5.78E-10	0.007
37	2.30E-05	0.43
71	2.58E-05	0.7
58	3.84E-05	0.32
32	0.0003	0.96
40	0.04	0.004
26	0.08	0.48
31	0.08	0.004
73	0.1	0.16
38	0.12	0.009
74	0.15	0.53
16	0.18	0.68
86	0.19	0.32
57	0.33	0.73
60	0.56	0.58
30	0.57	0.69
28	0.72	0.46
9	0.73	N/A
85	0.86	0.32
14	0.95	N/A
25	0.95	N/A
78	0.96	N/A

N/A means that the effect at the position in question was completely collinear with the conditioned variable, i.e., there were no polymorphic residues to test with case/ control status after conditioning on a position

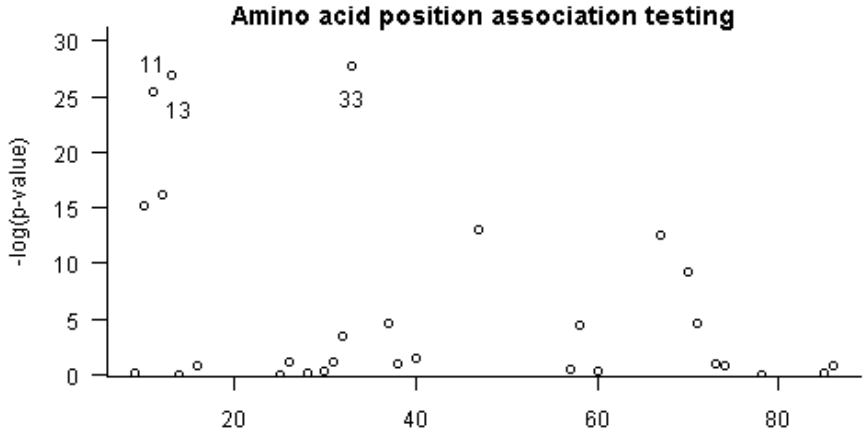

Figure 3.7 A plot of the strength of association between amino acid positions in the HLA DRB1 locus and RA in black South Africans before conditioning on position 11.

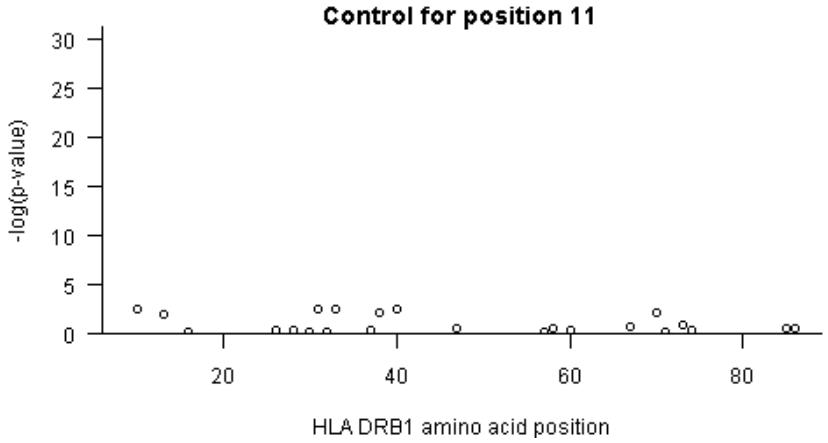

Figure 3.8 A plot of the strength of association between amino acid positions in the HLA DRB1 locus and RA in black South Africans after conditioning on position 11.

Within position 11 the odds ratios (+/- 95% CI) of association were: valine 5.1 (3.7, 6.97); leucine 1.2 (0.70, 2.00); aspartate 1.1 (0.4, 2.9); glycine 0.98 (0.6, 1.6); serine 0.38 (0.30, 0.48) with proline as the referent, depicted in Table 3.5.

Table 3.5 The frequency and strength of association of amino acid residues at HLA DRB1 position 11

AA	Frequency cases	Frequency controls	OR (95% CI)	P value
Valine	0.34	0.09	5.1 (3.74-6.97)	1.63E-27
Leucine	0.05	0.04	1.2 (0.70-2.00)	0.5273
Aspartate	0.02	0.01	1.1 (0.43-2.85)	0.8188
Glycine	0.06	0.06	0.98 (0.60-1.61)	0.9482
Proline	0.08	0.10	0.6 (0.50-1.11)	0.1507
Serine	0.46	0.69	0.3 (0.30-0.48)	2.46E-16

3.4 Discussion

The association of the HLA region and RA is well established and accounts for a third of the genetic risk. Despite the shared epitope hypothesis being the foundation for numerous genetic studies over the last 30 years, it does not fully explain the association at DRB1. Furthermore there has always been suspicion of the presence of independent associations in the HLA region, outside of DRB1. Moreover, although GWAS has identified numerous associations with RA, identifying the true causal variant in this highly polymorphic region has been challenging and identifying a functional variant with a plausible biological role has been difficult.

With the use of DNA sequencing and inference, this study was able to replicate an association observed in two other continental populations, Caucasians and South Koreans, in black South Africans. Amino acid position 11 of DRB1 which locates to the peptide binding groove of the molecule, and thus plays a role in antigen presentation, was identified as the most strongly associated amino acid position in black South Africans with seropositive RA.

The present study validates the previous known association with the *0404 and *0401 alleles (Meyer et al., 2004), however it is an advancement of studies done previously in that it examines all the allelic variation in HLA DRB1 to 4 digit resolution. Two other alleles apart from those already known to be associated with RA in black South Africans were found to be associated, *0405 and *1001. The *0401, *0404, *0405 and *1001 were identified as risk alleles and encode valine at position11 of DRB1. Conversely, *1301, *1302 and *1101 alleles were found to be protective and encode for serine at position 11. These risk and protective RA

alleles are very similar to that from Caucasian studies and highlight similarities in the HLA risk for RA across Caucasian and black South African populations.

The association of the SE alleles and RA has long been known, even in African populations. However identifying additional independent risk for RA in the HLA region has remained challenging until recently. Raychaudhuri et al., using a sophisticated imputation, technique examined the extended HLA region and found amino acid positions 11, 71 and 74 of DRB1 to be significantly associated with RA in Caucasians (Raychaudhuri et al., 2012). Amino acid position 11 was found the most significantly associated with RA in Caucasians.

This study demonstrated that similar to Caucasians, position 11 of DRB1 is the most highly associated with RA. Furthermore, valine was the most highly associated amino acid and leucine conferred a similar risk and serine was protective in black South Africans.

Unlike in Caucasians, amino acid position 71 and 74 was not found to confer risk in black South Africans. The difference in findings could be because these are two genetically distinct populations or because of this study being underpowered to detect the smaller effect exerted by position 71 and 74. Typically a study size of at least 1000 cases and controls to feel confident about the presence of a second amino acid effect. Approximately 90% of the DRB1 signal in Caucasians is from amino acid position 11.

Amino acid position 11 locates to the peptide binding groove of the HLA class II molecule emphasizing its role in antigen presentation. This position is also found to confer risk for ulcerative colitis (Achkar et al., 2012) and hydrophobic residues confer protection for sarcoidosis (Foley et al., 2001).

Raychaudhuri et al. used existing GWAS SNP data of 5,018 seropositive RA and 14,974 controls and imputed their allele and amino acid sequence from a large reference set. Their study spanned the HLA A, B, C, HLA DP, DQ and DR. The authors tested the accuracy of the imputation by comparing the imputed DRB1 classical alleles to genotyped alleles (Raychaudhuri et al., 2012). High accuracy of the imputation was observed. The authors' findings were that just 5 amino acid positions in 3 proteins accounted for most of the risk for RA in the HLA, positions 11, 71, 74 of DRB1, 9 in HLA B and 9 in HLA DPB1. In the DRB1 position 11 showed the strongest association and similarly positions 11 and 13 in the Koreans were the most significantly associated.

However due to the small sample size of this study, distinguishing the causal variant between the two positions, 11 and 13, just 2 codons apart was difficult. Nonetheless, position 11 has shown to be strongly associated with RA in 3 major ethnic groups, Caucasians, Asians and as shown in this study in black South Africans.

Amino acid position 11 locates to pocket 6 of the antigen binding groove of the HLA class II molecule. It is the only variable residue in pocket 6 thus accounting for all of the antigen variability of this pocket (Achkar et al., 2012). Valine confers the strongest risk for RA in European and the black South African populations however the effect size is higher in black South Africans. Similar to Caucasians, serine confers protection in both populations. Furthermore, position11 has also been identified as a major determinant in the risk for UC and sarcoidosis. However, unlike in RA, valine confers protection for UC which may highlight differences in antigen specificity of the two diseases. It is not known why valine

confers risk for RA; it is thought that it could be related to its non-polarity versus serine, which is highly polar.

Although the trend for risk and protection of these amino acids are the same in Caucasian and African populations, the allele frequencies are different. The frequency of valine in Caucasians is higher than in black South Africans and serine occurs at a much higher frequency in black South Africans (Fig 3.11). It is interesting that the frequency of the valine in controls is much lower in black South Africans than Caucasians. This can either be due to genetic drift or natural selection. Together, these results give the impression that Caucasians have more susceptibility and black South Africans more protection for RA.

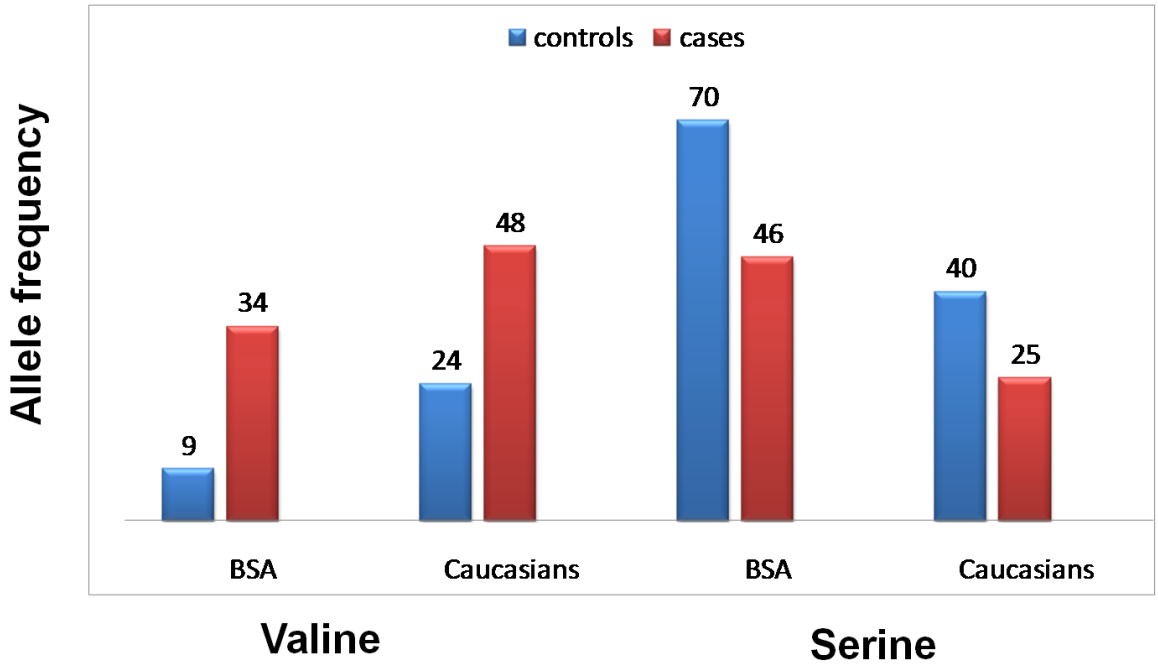

Figure 3.11 The frequency of the valine and serine in black South Africans and Caucasians.

One of the limitations of this study is the relatively small sample size. Although it was adequate to detect the large effect of amino acid position 11 it may not have been large enough to detect the smaller effects of other amino acid positions.

Now that the precise position at which the antigen that triggers the immune system in RA is known, the next step is to understand the antigen and the receptors that it interacts with, in the hope of finding new approaches to therapy.

CHAPTER 4: CONCLUSION

In this study, the first large scale genetic project performed on an African population with RA, the association of the HLA region and RA was validated. This association was further refined to a specific amino acid position in the HLA DRB1 molecule. Similar to Caucasians, valine at amino acid position 11 the HLA DRB1 chain is the most strongly associated with RA and serine is protective. Position 11 locates to the binding groove of the HLA DRB1 molecule and its role in antigen presentation makes its association a plausible and likely causal one. Furthermore by understanding the antigen that binds to this position, the receptors that it binds to and pathways that are triggered, a better understanding of the pathogenesis of RA will be achieved.

In addition, this study identified four novel non-HLA variants, pointing to copy number variation rather than SNPs that locate to the intergenic region of *LOC400723* and *PLD5* on chromosome 1 as being associated with RA in black South Africans. The possible significance of this finding is currently being investigated. The association of these 4 variants is stronger in the RF+ subgroup, suggesting this group is probably a more genetically homogenous.

None of the non-HLA loci previously associated with RA in Caucasians could be definitively shown to confer risk for RA in black South Africans. The possible reason for the lack of association in these genetically distinct populations may be that these loci are truly not risk loci in the studied population. However, this study was underpowered to detect the modest to small effects of many of these loci. In addition, mining of the data shows that some of the risk variants in Caucasians are non-polymorphic in black South Africans and in others the MAF is significantly lower and rare. To support this, it was shown that Caucasians and black South

Africans are genetically distinct in that the variants of the *PTPN22* gene which confer the strongest non-HLA risk for RA in Caucasians was found not to confer risk in black South Africans. This study was sufficiently powered to detect the effects that these variants exert in Caucasians. In particular, the well studied variant rs2476601, which encodes an amino acid change (R620W) in one of four SH3 domain binding sites in the PTPN22 molecule, is non-polymorphic in black South Africans. These findings therefore suggest that previously described non-HLA risk variants in Caucasians with RA probably do not confer significant risk in black South Africans.

Perhaps the subgroup of RA patients with the unique phenotype, 'the African variant' carries genetic variants that contribute to this phenotype. Another unanswered question is why the association of the non-HLA SNPs identified in this study have a stronger association with the RF+ than the ACCP+ subgroup.

Using a sophisticated high-throughput genotyping technique such as the Immunochip array, allowed for the identification of admixed individuals and thus for better quality control of the dataset where population structure as a confounder could be avoided. One of the interesting points of this study is that black South Africans are genetically distinct from populations resident in the East and West Africa. These findings emphasise the need for a unique reference set of data for southern Africans, which is currently lacking. Moreover, the currently available reference data for other African populations cannot necessarily be extrapolated to south Africans. The broader implications of these findings are that development of more effective vaccines, therapeutics and public health interventions elucidated in one African population cannot be simply applied to the rest of the African continent.

The overall significance of this study is that it has given great insights into a poorly studied field, the genetics of RA in black South Africans. There are risk loci that have been shown to be shared and others that are specific to this study population.

Importantly, despite the study's limitations the HLA region has been shown to contain shared risk loci between the major ethnic groups, but several of the non-HLA loci are probably population specific and there is a need for larger studies to be performed in different African populations with the hope of identifying pathogenic pathways that may be used in designing of individualized, targeted therapy.

4.1 The way forward

There is a clear need to conduct genetic studies in a larger sample size of black South Africans with RA to tests for risk variants with smaller effect sizes.

The significance of the non-HLA variants on chromosome 1 that locate to the intergenic region between *PLD5* and *LOC400723* needs to be determined.

Furthermore, there is a need to replicate the potential association of the *RBPJ* and *ILR1* loci in a larger sample size.

Identifying the antigen that binds to position11 of DRB1 will no doubt lead to a better understanding of the pathogenesis of RA. Another future project entails identifying non-HLA loci and amino acid position 11/13 of HLA DRB1 for association with severe RA in this population. Lastly, determining whether carriers of valine at amino acid postion11 have a unique cytokine profile compared to non-carriers is necessary and may lead to targets of therapy.

REFERENCES

(2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

(2011) A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet,* 7**,** e1002142.

Achkar, J. P., Klei, L., De Bakker, P. I., Bellone, G., Rebert, N., Scott, R.*, et al.* (2012) Amino acid position 11 of HLA-DRbeta1 is a major determinant of chromosome 6p association with ulcerative colitis. *Genes Immun,* 13**,** 245-52.

Adebajo, A. O. & Reid, D. M. (1991) The pattern of rheumatoid arthritis in West Africa and comparison with a cohort of British patients. *Q J Med,* 80**,** 633-40.

Aletaha, D. & Smolen, J. (2005) The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. *Clin Exp Rheumatol,* 23**,** S100-8.

Amos, C. I., Chen, W. V., Lee, A., Li, W., Kern, M., Lundsten, R.*, et al.* (2006) High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. *Genes Immun,* 7**,** 277-86.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. & Zondervan, K. T. (2010) Data quality control in genetic case-control association studies. *Nat Protoc,* 5**,** 1564-73.

Appelboom, T. (2005) Hypothesis: Rubens--one of the first victims of an epidemic of rheumatoid arthritis that started in the 16th-17th century? *Rheumatology (Oxford),* 44**,** 681-3.

Arnett, F. C., Edworthy, S. M., Bloch, D. A., Mcshane, D. J., Fries, J. F., Cooper, N. S.*, et al.* (1988) The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum,* 31**,** 315-24.

Baerwald, C. G., Mok, C. C., Tickly, M., Lau, C. S., Wordsworth, B. P., Ollier, B.*, et al.* (2000) Corticotropin releasing hormone (CRH) promoter polymorphisms in various ethnic groups of patients with rheumatoid arthritis. *Z Rheumatol,* 59**,** 29-34.

Bagg, L. R., Hansen, D. P., Lewis, C. & Houba, V. (1979) Rheumatoid arthritis in Kenya. I. Clinical observations. *Ann Rheum Dis,* 38**,** 23-5.

Baka, Z., Buzas, E. & Nagy, G. (2009) Rheumatoid arthritis and smoking: putting the pieces together. *Arthritis Res Ther,* 11**,** 238.

Barton, A., Bowes, J., Eyre, S., Spreckley, K., Hinks, A., John, S.*, et al.* (2004a) A functional haplotype of the PADI4 gene associated with rheumatoid arthritis in a Japanese population is not associated in a United Kingdom population. *Arthritis Rheum,* 50**,** 1117-21.

Barton, A., Jury, F., Eyre, S., Bowes, J., Hinks, A., Ward, D.*, et al.* (2004b) Haplotype analysis in simplex families and novel analytic approaches in a case-control cohort reveal no evidence of association of the CTLA-4 gene with rheumatoid arthritis. *Arthritis Rheum,* 50**,** 748-52.

Bax, M., Van Heemst, J., Huizinga, T. W. & Toes, R. E. (2011) Genetics of rheumatoid arthritis: what have we learned? *Immunogenetics,* 63**,** 459-66.

Begovich, A. B., Carlton, V. E., Honigberg, L. A., Schrodi, S. J., Chokkalingam, A. P., Alexander, H. C.*, et al.* (2004) A missense single-nucleotide

polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet,* 75**,** 330-7.
Beighton, P., Solomon, L. & Valkenburg, H. A. (1975) Rheumatoid arthritis in a rural South African Negro population. *Ann Rheum Dis,* 34**,** 136-41.
Bengtsson, C., Nordmark, B., Klareskog, L., Lundberg, I. & Alfredsson, L. (2005) Socioeconomic status and the risk of developing rheumatoid arthritis: results from the Swedish EIRA study. *Ann Rheum Dis,* 64**,** 1588-94.
Benitha, R. & Tikly, M. (2007) Functional disability and health-related quality of life in South Africans with rheumatoid arthritis and systemic lupus erythematosus. *Clin Rheumatol,* 26**,** 24-9.
Benito-Garcia, E., Feskanich, D., Hu, F. B., Mandl, L. A. & Karlson, E. W. (2007) Protein, iron, and meat consumption and risk for rheumatoid arthritis: a prospective cohort study. *Arthritis Res Ther,* 9**,** R16.
Berglin, E., Kokkonen, H., Einarsdottir, E., Agren, A. & Rantapaa Dahlqvist, S. (2010) Influence of female hormonal factors, in relation to autoantibodies and genetic markers, on the development of rheumatoid arthritis in northern Sweden: a case-control study. *Scand J Rheumatol,* 39**,** 454-60.
Bhatia, S. S., Majka, D. S., Kittelson, J. M., Parrish, L. A., Ferucci, E. D., Deane, K. D.*, et al.* (2007) Rheumatoid factor seropositivity is inversely associated with oral contraceptive use in women without rheumatoid arthritis. *Ann Rheum Dis,* 66**,** 267-9.
Bhayani, H. R. & Hedrick, S. M. (1991) The role of polymorphic amino acids of the MHC molecule in the selection of the T cell repertoire. *J Immunol,* 146**,** 1093-8.
Bileckot, R. & Malonga, A. C. (1998) Rheumatoid arthritis in Congo-Brazzaville. A study of thirty-six cases. *Rev Rhum Engl Ed,* 65**,** 308-12.
Boonen, A. & Mau, W. (2009) The economic burden of disease: comparison between rheumatoid arthritis and ankylosing spondylitis. *Clin Exp Rheumatol,* 27**,** S112-7.
Buchs, N., Di Giovine, F. S., Silvestri, T., Vannier, E., Duff, G. W. & Miossec, P. (2001) IL-1B and IL-1Ra gene polymorphisms and disease severity in rheumatoid arthritis: interaction with their plasma levels. *Genes Immun,* 2**,** 222-8.
Bukhari, M., Lunt, M., Barton, A., Bunn, D., Silman, A. & Symmons, D. (2007) Increasing age at symptom onset is associated with worse radiological damage at presentation in patients with early inflammatory polyarthritis. *Ann Rheum Dis,* 66**,** 389-93.
Calafell, F., Shuster, A., Speed, W. C., Kidd, J. R. & Kidd, K. K. (1998) Short tandem repeat polymorphism evolution in humans. *Eur J Hum Genet,* 6**,** 38-49.
Camacho, E. M., Verstappen, S. M. & Symmons, D. P. (2012) Association between socioeconomic status, learned helplessness, and disease outcome in patients with inflammatory polyarthritis. *Arthritis Care Res (Hoboken),* 64**,** 1225-32.
Campbell, M. C. & Tishkoff, S. A. (2010) The evolution of human genetic and phenotypic variation in Africa. *Curr Biol,* 20**,** R166-73.
Cano, P., Klitz, W., Mack, S. J., Maiers, M., Marsh, S. G., Noreen, H.*, et al.* (2007) Common and well-documented HLA alleles: report of the Ad-Hoc

committee of the american society for histocompatiblity and immunogenetics. *Hum Immunol,* 68**,** 392-417.
Cantagrel, A., Navaux, F., Loubet-Lescoulie, P., Nourhashemi, F., Enault, G., Abbal, M.*, et al.* (1999) Interleukin-1beta, interleukin-1 receptor antagonist, interleukin-4, and interleukin-10 gene polymorphisms: relationship to occurrence and severity of rheumatoid arthritis. *Arthritis Rheum,* 42**,** 1093-100.
Carrier, N., Cossette, P., Daniel, C., De Brum-Fernandes, A., Liang, P., Menard, H. A.*, et al.* (2009) The DERAA HLA-DR alleles in patients with early polyarthritis: protection against severe disease and lack of association with rheumatoid arthritis autoantibodies. *Arthritis Rheum,* 60**,** 698-707.
Chang, M. C., Chang, Y. T., Tien, Y. W., Liang, P. C., Jan, I. S., Wei, S. C.*, et al.* (2007) T-cell regulatory gene CTLA-4 polymorphism/haplotype association with autoimmune pancreatitis. *Clin Chem,* 53**,** 1700-5.
Chen, J. Y., Wang, C. M., Ma, C. C., Hsu, L. A., Ho, H. H., Wu, Y. J.*, et al.* (2008) A transmembrane polymorphism in FcgammaRIIb (FCGR2B) is associated with the production of anti-cyclic citrullinated peptide autoantibodies in Taiwanese RA. *Genes Immun,* 9**,** 680-8.
Chen, R., Wei, Y., Cai, Q., Duan, S., Ren, D., Shen, J.*, et al.* (2011) The PADI4 gene does not contribute to genetic susceptibility to rheumatoid arthritis in Chinese Han population. *Rheumatol Int,* 31**,** 1631-4.
Chikanza, I. C., Stein, M., Lutalo, S. & Gibson, T. (1994) The clinical, serologic and radiologic features of rheumatoid arthritis in ethnic black Zimbabwean and British Caucasian patients. *J Rheumatol,* 21**,** 2011-5.
Choo, S. Y. (2007) The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Med J,* 48**,** 11-23.
Cojocaru, M., Cojocaru, I. M., Silosi, I., Vrabie, C. D. & Tanasescu, R. (2010) Extra-articular Manifestations in Rheumatoid Arthritis. *Maedica (Buchar),* 5**,** 286-91.
Constantin, A., Lauwers-Cances, V., Navaux, F., Abbal, M., Van Meerwijk, J., Mazieres, B.*, et al.* (2002) Stromelysin 1 (matrix metalloproteinase 3) and HLA-DRB1 gene polymorphisms: Association with severity and progression of rheumatoid arthritis in a prospective study. *Arthritis Rheum,* 46**,** 1754-62.
Cooles, F. A. & Isaacs, J. D. (2011) Pathophysiology of rheumatoid arthritis. *Curr Opin Rheumatol,* 23**,** 233-40.
Cooper, J. D., Simmonds, M. J., Walker, N. M., Burren, O., Brand, O. J., Guo, H.*, et al.* (2012) Seven newly identified loci for autoimmune thyroid disease. *Hum Mol Genet,* 21**,** 5202-8.
Cornelis, F., Faure, S., Martinez, M., Prud'homme, J. F., Fritz, P., Dib, C.*, et al.* (1998) New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. *Proc Natl Acad Sci U S A,* 95**,** 10746-50.
Cortes, A. & Brown, M. A. (2011) Promise and pitfalls of the Immunochip. *Arthritis Res Ther,* 13**,** 101.
Costenbader, K. H., Feskanich, D., Holmes, M., Karlson, E. W. & Benito-Garcia, E. (2008) Vitamin D intake and risks of systemic lupus erythematosus and rheumatoid arthritis in women. *Ann Rheum Dis,* 67**,** 530-5.
Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S.*, et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature,* 464**,** 713-20.

Cutolo, M. (2004) Estrogen metabolites: increasing evidence for their role in rheumatoid arthritis and systemic lupus erythematosus. *J Rheumatol,* 31**,** 419-21.
Dieppe, P. & Rogers, J. (1986) Art, history and the antiquity of rheumatic diseases. *Clin Rheumatol,* 5**,** 179-80.
Dinarello, C. A. (2011) Interleukin-1 in the pathogenesis and treatment of inflammatory diseases. *Blood,* 117**,** 3720-32.
Ding, B., Padyukov, L., Lundstrom, E., Seielstad, M., Plenge, R. M., Oksenberg, J. R.*, et al.* (2009) Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region. *Arthritis Rheum,* 60**,** 30-8.
Doran, M. F., Crowson, C. S., O'fallon, W. M. & Gabriel, S. E. (2004) The effect of oral contraceptives and estrogen replacement therapy on the risk of rheumatoid arthritis: a population based study. *J Rheumatol,* 31**,** 207-13.
Dorr, S., Lechtenbohmer, N., Rau, R., Herborn, G., Wagner, U., Muller-Myhsok, B.*, et al.* (2004) Association of a specific haplotype across the genes MMP1 and MMP3 with radiographic joint destruction in rheumatoid arthritis. *Arthritis Res Ther,* 6**,** R199-207.
Douglas, K. M., Pace, A. V., Treharne, G. J., Saratzis, A., Nightingale, P., Erb, N.*, et al.* (2006) Excess recurrent cardiac events in rheumatoid arthritis patients with acute coronary syndrome. *Ann Rheum Dis,* 65**,** 348-53.
Dowman, B., Campbell, R. M., Zgaga, L., Adeloye, D. & Chan, K. Y. (2012) Estimating the burden of rheumatoid arthritis in Africa: A systematic analysis. *J Glob Health,* 2**,** 20406.
Emery, P. & Salmon, M. (1995) Early rheumatoid arthritis: time to aim for remission? *Ann Rheum Dis,* 54**,** 944-7.
Engelhard, V. H. (1994) Structure of peptides associated with class I and class II MHC molecules. *Annu Rev Immunol,* 12**,** 181-207.
Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P.*, et al.* (2012) High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet,* 44**,** 1336-40.
Feitsma, A. L., Van Der Helm-Van Mil, A. H., Huizinga, T. W., De Vries, R. R. & Toes, R. E. (2008) Protection against rheumatoid arthritis by HLA: nature and nurture. *Ann Rheum Dis,* 67 Suppl 3**,** iii61-3.
Fernando, M. M., Stevens, C. R., Walsh, E. C., De Jager, P. L., Goyette, P., Plenge, R. M.*, et al.* (2008) Defining the role of the MHC in autoimmunity: a review and pooled analysis. *PLoS Genet,* 4**,** e1000024.
Foley, P. J., Mcgrath, D. S., Puscinska, E., Petrek, M., Kolek, V., Drabek, J.*, et al.* (2001) Human leukocyte antigen-DRB1 position 11 residues are a common protective marker for sarcoidosis. *Am J Respir Cell Mol Biol,* 25**,** 272-7.
Fung, E. Y., Smyth, D. J., Howson, J. M., Cooper, J. D., Walker, N. M., Stevens, H.*, et al.* (2009) Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. *Genes Immun,* 10**,** 188-91.
Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B.*, et al.* (2002) The structure of haplotype blocks in the human genome. *Science,* 296**,** 2225-9.

Gabriel, S. E., Crowson, C. S. & O'fallon, W. M. (1999) The epidemiology of rheumatoid arthritis in Rochester, Minnesota, 1955-1985. *Arthritis Rheum,* 42**,** 415-20.

Gandjbakhch, F., Fajardy, I., Ferre, B., Dubucquoi, S., Flipo, R. M., Roger, N.*, et al.* (2009) A functional haplotype of PADI4 gene in rheumatoid arthritis: positive correlation in a French population. *J Rheumatol,* 36**,** 881-6.

Genevay, S., Di Giovine, F. S., Perneger, T. V., Silvestri, T., Stingelin, S., Duff, G.*, et al.* (2002) Association of interleukin-4 and interleukin-1B gene variants with Larsen score progression in rheumatoid arthritis. *Arthritis Rheum,* 47**,** 303-9.

Goodall, J. W. (1956) Joint swellings in Africans; a review of 90 cases. *Cent Afr J Med,* 2**,** 220-3.

Graham, R. R., Cotsapas, C., Davies, L., Hackett, R., Lessard, C. J., Leon, J. M.*, et al.* (2008) Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nat Genet,* 40**,** 1059-61.

Green, E. K., Hamshere, M., Forty, L., Gordon-Smith, K., Fraser, C., Russell, E.*, et al.* (2012) Replication of bipolar disorder susceptibility alleles and identification of two novel genome-wide significant associations in a new bipolar disorder case-control sample. *Mol Psychiatry*.

Greenwood, B. M. (1969) Polyarthritis in Western Nigeria. I. Rheumatoid arthritis. *Ann Rheum Dis,* 28**,** 488-96.

Gregersen, P. K. (2005) Gaining insight into PTPN22 and autoimmunity. *Nat Genet,* 37**,** 1300-2.

Gregersen, P. K., Amos, C. I., Lee, A. T., Lu, Y., Remmers, E. F., Kastner, D. L.*, et al.* (2009) REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat Genet,* 41**,** 820-3.

Gregersen, P. K., Silver, J. & Winchester, R. J. (1987) The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum,* 30**,** 1205-13.

Heggarty, S., Suppiah, V., Silversides, J., O'doherty, C., Droogan, A., Mcdonnell, G.*, et al.* (2007) CTLA4 gene polymorphisms and multiple sclerosis in Northern Ireland. *J Neuroimmunol,* 187**,** 187-91.

Hitchon, C. A., Chandad, F., Ferucci, E. D., Willemze, A., Ioan-Facsinay, A., Van Der Woude, D.*, et al.* (2010) Antibodies to porphyromonas gingivalis are associated with anticitrullinated protein antibodies in patients with rheumatoid arthritis and their relatives. *J Rheumatol,* 37**,** 1105-12.

Hodkinson, B., Meyer, P. W., Musenge, E., Ally, M. M., Wadee, A. A., Anderson, R.*, et al.* (2010) The diagnostic utility of the anti-CCP antibody test is no better than rheumatoid factor in South Africans with early rheumatoid arthritis. *Clin Rheumatol,* 29**,** 615-8.

Holoshitz, J. (2010) The rheumatoid arthritis HLA-DRB1 shared epitope. *Curr Opin Rheumatol,* 22**,** 293-8.

Huber, L. C., Brock, M., Hemmatazad, H., Giger, O. T., Moritz, F., Trenkmann, M.*, et al.* (2007) Histone deacetylase/acetylase activity in total synovial tissue derived from rheumatoid arthritis and osteoarthritis patients. *Arthritis Rheum,* 56**,** 1087-93.

Hughes, L. B., Reynolds, R. J., Brown, E. E., Kelley, J. M., Thomson, B., Conn, D. L.*, et al.* (2010) Most common single-nucleotide polymorphisms associated

with rheumatoid arthritis in persons of European ancestry confer risk of rheumatoid arthritis in African Americans. *Arthritis Rheum,* 62**,** 3547-53.
Huizinga, T. W., Keijsers, V., Yanni, G., Hall, M., Ramage, W., Lanchbury, J.*, et al.* (2000) Are differences in interleukin 10 production associated with joint damage? *Rheumatology (Oxford),* 39**,** 1180-8.
Huppa, J. B. & Davis, M. M. (2003) T-cell-antigen recognition and the immunological synapse. *Nat Rev Immunol,* 3**,** 973-83.
Ikari, K., Kuwahara, M., Nakamura, T., Momohara, S., Hara, M., Yamanaka, H.*, et al.* (2005) Association between PADI4 and rheumatoid arthritis: a replication study. *Arthritis Rheum,* 52**,** 3054-7.
Ikari, K., Momohara, S., Inoue, E., Tomatsu, T., Hara, M., Yamanaka, H.*, et al.* (2006) Haplotype analysis revealed no association between the PTPN22 gene and RA in a Japanese population. *Rheumatology (Oxford),* 45**,** 1345-8.
Iwamoto, T., Ikari, K., Nakamura, T., Kuwahara, M., Toyama, Y., Tomatsu, T.*, et al.* (2006) Association between PADI4 and rheumatoid arthritis: a meta-analysis. *Rheumatology (Oxford),* 45**,** 804-7.
Jacobsson, L. T., Jacobsson, M. E., Askling, J. & Knowler, W. C. (2003) Perinatal characteristics and risk of rheumatoid arthritis. *BMJ,* 326**,** 1068-9.
Jawaheer, D., Seldin, M. F., Amos, C. I., Chen, W. V., Shigeta, R., Monteiro, J.*, et al.* (2001) A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am J Hum Genet,* 68**,** 927-36.
Jouvenne, P., Chaudhary, A., Buchs, N., Giovine, F. S., Duff, G. W. & Miossec, P. (1999) Possible genetic association between interleukin-1alpha gene polymorphism and the severity of chronic polyarthritis. *Eur Cytokine Netw,* 10**,** 33-6.
Julia, A., Ballina, J., Canete, J. D., Balsa, A., Tornero-Molina, J., Naranjo, A.*, et al.* (2008) Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. *Arthritis Rheum,* 58**,** 2275-86.
Juran, B. D., Hirschfield, G. M., Invernizzi, P., Atkinson, E. J., Li, Y., Xie, G.*, et al.* (2012) Immunochip analyses identify a novel risk locus for primary biliary cirrhosis at 13q14, multiple independent associations at four established risk loci and epistasis between 1p31 and 7q32 risk variants. *Hum Mol Genet*.
Kallberg, H., Ding, B., Padyukov, L., Bengtsson, C., Ronnelid, J., Klareskog, L.*, et al.* (2011) Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke. *Ann Rheum Dis,* 70**,** 508-11.
Kallberg, H., Jacobsen, S., Bengtsson, C., Pedersen, M., Padyukov, L., Garred, P.*, et al.* (2009) Alcohol consumption is associated with decreased risk of rheumatoid arthritis: results from two Scandinavian case-control studies. *Ann Rheum Dis,* 68**,** 222-7.
Kallberg, H., Padyukov, L., Plenge, R. M., Ronnelid, J., Gregersen, P. K., Van Der Helm-Van Mil, A. H.*, et al.* (2007) Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. *Am J Hum Genet,* 80**,** 867-75.

Kang, C. P., Lee, H. S., Ju, H., Cho, H., Kang, C. & Bae, S. C. (2006) A functional haplotype of the PADI4 gene associated with increased rheumatoid arthritis susceptibility in Koreans. *Arthritis Rheum,* 54**,** 90-6.
Kanyerezi, B. R., Baddeley, H. & Kisumba, D. (1970) Rheumatoid arthritis in Ugandan Africans. *Ann Rheum Dis,* 29**,** 617-21.
Karlson, E. W., Chibnik, L. B., Cui, J., Plenge, R. M., Glass, R. J., Maher, N. E.*, et al.* (2008) Associations between human leukocyte antigen, PTPN22, CTLA4 genotypes and rheumatoid arthritis phenotypes of autoantibody status, age at diagnosis and erosions in a large cohort study. *Ann Rheum Dis,* 67**,** 358-63.
Karlson, E. W., Mandl, L. A., Hankinson, S. E. & Grodstein, F. (2004) Do breast-feeding and other reproductive factors influence future risk of rheumatoid arthritis? Results from the Nurses' Health Study. *Arthritis Rheum,* 50**,** 3458-67.
Karouzakis, E., Gay, R. E., Gay, S. & Neidhart, M. (2009a) Epigenetic control in rheumatoid arthritis synovial fibroblasts. *Nat Rev Rheumatol,* 5**,** 266-72.
Karouzakis, E., Gay, R. E., Michel, B. A., Gay, S. & Neidhart, M. (2009b) DNA hypomethylation in rheumatoid arthritis synovial fibroblasts. *Arthritis Rheum,* 60**,** 3613-22.
Keitel, W. (2009) [The High Priest of gout - Sir Alfred Baring Garrod (1819-1907)]. *Z Rheumatol,* 68**,** 851-6.
Kilding, R., Iles, M. M., Timms, J. M., Worthington, J. & Wilson, A. G. (2004) Additional genetic susceptibility for rheumatoid arthritis telomeric of the DRB1 locus. *Arthritis Rheum,* 50**,** 763-9.
Kim, S. Y., Han, S. W., Kim, G. W., Lee, J. M. & Kang, Y. M. (2004) TGF-beta1 polymorphism determines the progression of joint damage in rheumatoid arthritis. *Scand J Rheumatol,* 33**,** 389-94.
Klareskog, L., Malmstrom, V., Lundberg, K., Padyukov, L. & Alfredsson, L. (2011) Smoking, citrullination and genetic variability in the immunopathogenesis of rheumatoid arthritis. *Semin Immunol,* 23**,** 92-8.
Klareskog, L., Stolt, P., Lundberg, K., Kallberg, H., Bengtsson, C., Grunewald, J.*, et al.* (2006) A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. *Arthritis Rheum,* 54**,** 38-46.
Knight, J. C. (2013) Genomic modulators of the immune response. *Trends Genet,* 29**,** 74-83.
Korendowych, E., Dixey, J., Cox, B., Jones, S. & Mchugh, N. (2003) The Influence of the HLA-DRB1 rheumatoid arthritis shared epitope on the clinical characteristics and radiological outcome of psoriatic arthritis. *J Rheumatol,* 30**,** 96-101.
Kozyrev, S. V., Abelson, A. K., Wojcik, J., Zaghlool, A., Linga Reddy, M. V., Sanchez, E.*, et al.* (2008) Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat Genet,* 40**,** 211-6.
Kurreeman, F., Liao, K., Chibnik, L., Hickey, B., Stahl, E., Gainer, V.*, et al.* (2011) Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am J Hum Genet,* 88**,** 57-69.
Kurreeman, F. A., Padyukov, L., Marques, R. B., Schrodi, S. J., Seddighzadeh, M., Stoeken-Rijsbergen, G.*, et al.* (2007) A candidate gene approach identifies

the TRAF1/C5 region as a risk factor for rheumatoid arthritis. *PLoS Med,* 4**,** e278.
Kvien, T. K. (2004) Epidemiology and burden of illness of rheumatoid arthritis. *Pharmacoeconomics,* 22**,** 1-12.
Kyogoku, C., Langefeld, C. D., Ortmann, W. A., Lee, A., Selby, S., Carlton, V. E.*, et al.* (2004) Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *Am J Hum Genet,* 75**,** 504-7.
Landre-Beauvais, A. J. (2001) The first description of rheumatoid arthritis. Unabridged text of the doctoral dissertation presented in 1800. *Joint Bone Spine,* 68**,** 130-43.
Lane, A. B., Soodyall, H., Arndt, S., Ratshikhopha, M. E., Jonker, E., Freeman, C.*, et al.* (2002) Genetic substructure in South African Bantu-speakers: evidence from autosomal DNA and Y-chromosome studies. *Am J Phys Anthropol,* 119**,** 175-85.
Lee, H. S., Korman, B. D., Le, J. M., Kastner, D. L., Remmers, E. F., Gregersen, P. K.*, et al.* (2009) Genetic risk factors for rheumatoid arthritis differ in Caucasian and Korean populations. *Arthritis Rheum,* 60**,** 364-71.
Lee, H. S., Lee, A. T., Criswell, L. A., Seldin, M. F., Amos, C. I., Carulli, J. P.*, et al.* (2008) Several regions in the major histocompatibility complex confer risk for anti-CCP-antibody positive rheumatoid arthritis, independent of the DRB1 locus. *Mol Med,* 14**,** 293-300.
Lie, B. A., Viken, M. K., Odegard, S., Van Der Heijde, D., Landewe, R., Uhlig, T.*, et al.* (2007) Associations between the PTPN22 1858C->T polymorphism and radiographic joint destruction in patients with rheumatoid arthritis: results from a 10-year longitudinal study. *Ann Rheum Dis,* 66**,** 1604-9.
Linn-Rasker, S. P., Van Der Helm-Van Mil, A. H., Van Gaalen, F. A., Kloppenburg, M., De Vries, R. R., Le Cessie, S.*, et al.* (2006) Smoking is a risk factor for anti-CCP antibodies only in rheumatoid arthritis patients who carry HLA-DRB1 shared epitope alleles. *Ann Rheum Dis,* 65**,** 366-71.
Lubbe, S., Tikly, M., Van Der Merwe, L., Hodkinson, B. & Ramsay, M. (2008) Interleukin-1 receptor antagonist gene polymorphisms are associated with disease severity in Black South Africans with rheumatoid arthritis. *Joint Bone Spine,* 75**,** 422-5.
Lubeck, D. P. (2001) A review of the direct costs of rheumatoid arthritis: managed care versus fee-for-service settings. *Pharmacoeconomics,* 19**,** 811-8.
Macgregor, A. J., Snieder, H., Rigby, A. S., Koskenvuo, M., Kaprio, J., Aho, K.*, et al.* (2000) Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum,* 43**,** 30-7.
Mackay, K., Milicic, A., Lee, D., Tikly, M., Laval, S., Shatford, J.*, et al.* (2003) Rheumatoid arthritis susceptibility and interleukin 10: a study of two ethnically diverse populations. *Rheumatology (Oxford),* 42**,** 149-53.
Maehlen, M. T., Nordang, G. B., Syversen, S. W., Van Der Heijde, D. M., Kvien, T. K., Uhlig, T.*, et al.* (2011) FCRL3 -169C/C genotype is associated with anti-citrullinated protein antibody-positive rheumatoid arthritis and with radiographic progression. *J Rheumatol,* 38**,** 2329-35.
Mandl, L. A., Costenbader, K. H., Simard, J. F. & Karlson, E. W. (2009) Is birthweight associated with risk of rheumatoid arthritis? Data from a large cohort study. *Ann Rheum Dis,* 68**,** 514-8.

Marinou, I., Healy, J., Mewar, D., Moore, D. J., Dickson, M. C., Binks, M. H.*, et al.* (2007) Association of interleukin-6 and interleukin-10 genotypes with radiographic damage in rheumatoid arthritis is dependent on autoantibody status. *Arthritis Rheum,* 56**,** 2549-56.

Marinou, I., Maxwell, J. R. & Wilson, A. G. (2010) Genetic influences modulating the radiological severity of rheumatoid arthritis. *Ann Rheum Dis,* 69**,** 476-82.

Maritz, N. G., Gerber, A. J., Greyling, S. J. & Sandra, B. B. (2005) A radiological study of the rheumatoid hand in black South Africans. *S Afr Med J,* 95**,** 795-7.

Martell, R. W., Du Toit, E. D., Kalla, A. A. & Meyers, O. L. (1989) Association of rheumatoid arthritis with HLA in three South African populations--whites, blacks and a population of mixed ancestry. *S Afr Med J,* 76**,** 189-90.

Martinez, A., Fernandez-Arquero, M., Pascual-Salcedo, D., Conejero, L., Alves, H., Balsa, A.*, et al.* (2000) Primary association of tumor necrosis factor-region genetic markers with susceptibility to rheumatoid arthritis. *Arthritis Rheum,* 43**,** 1366-70.

Martinez, A., Valdivia, A., Pascual-Salcedo, D., Lamas, J. R., Fernandez-Arquero, M., Balsa, A.*, et al.* (2005) PADI4 polymorphisms are not associated with rheumatoid arthritis in the Spanish population. *Rheumatology (Oxford),* 44**,** 1263-6.

Mattey, D. L., Nixon, N., Dawes, P. T. & Kerr, J. (2005) Association of polymorphism in the transforming growth factor {beta}1 gene with disease outcome and mortality in rheumatoid arthritis. *Ann Rheum Dis,* 64**,** 1190-4.

Mattey, D. L., Nixon, N. B., Dawes, P. T., Ollier, W. E. & Hajeer, A. H. (2004) Association of matrix metalloproteinase 3 promoter genotype with disease outcome in rheumatoid arthritis. *Genes Immun,* 5**,** 147-9.

May, A., Hazelhurst, S., Li, Y., Norris, S. A., Govind, N., Tikly, M.*, et al.* (2013) Genetic diversity in black South Africans from Soweto. *BMC Genomics,* 14**,** 644.

Mcinnes, I. B. & Schett, G. (2007) Cytokines in the pathogenesis of rheumatoid arthritis. *Nat Rev Immunol,* 7**,** 429-42.

Mcinnes, I. B. & Schett, G. (2011) The pathogenesis of rheumatoid arthritis. *N Engl J Med,* 365**,** 2205-19.

Meinecke, I., Cinski, A., Baier, A., Peters, M. A., Dankbar, B., Wille, A.*, et al.* (2007) Modification of nuclear PML protein by SUMO-1 regulates Fas-induced apoptosis in rheumatoid arthritis synovial fibroblasts. *Proc Natl Acad Sci U S A,* 104**,** 5073-8.

Meune, C., Touze, E., Trinquart, L. & Allanore, Y. (2009) Trends in cardiovascular mortality in patients with rheumatoid arthritis over 50 years: a systematic review and meta-analysis of cohort studies. *Rheumatology (Oxford),* 48**,** 1309-13.

Meyer, P. W., Hodkinson, B., Ally, M., Musenge, E., Wadee, A. A., Fickl, H.*, et al.* (2011) HLA-DRB1 shared epitope genotyping using the revised classification and its association with circulating autoantibodies, acute phase reactants, cytokines and clinical indices of disease activity in a cohort of South African rheumatoid arthritis patients. *Arthritis Res Ther,* 13**,** R160.

Michaud, K. & Wolfe, F. (2007) Comorbidities in rheumatoid arthritis. *Best Pract Res Clin Rheumatol,* 21**,** 885-906.

Miller, S. A., Dyk, D. D. & Pelesky, H. F. (1988) A simple salting-out procedure for extracting DNA from human nucleated cells. *Nucleic Acid Res,* 16**,** 1215.

Mody, G. M., Hammond, M. G. & Naidoo, P. D. (1989) HLA associations with rheumatoid arthritis in African blacks. *J Rheumatol,* 16**,** 1326-8.
Mody, G. M. & Meyers, O. L. (1989) Rheumatoid arthritis in blacks in South Africa. *Ann Rheum Dis,* 48**,** 69-72.
Mody, G. M., Shaw, J. & Ramchurren, A. (1988) Rheumatoid arthritis impact survey. *S Afr Med J,* 74**,** 409-10.
Moodley, D., Mody, G. M. & Chuturgoon, A. A. (2010) Functional analysis of the p53 codon 72 polymorphism in black South Africans with rheumatoid arthritis--a pilot study. *Clin Rheumatol,* 29**,** 1099-105.
Moolenburgh, J. D., Moore, S., Valkenburg, H. A. & Erasmus, M. G. (1984) Rheumatoid arthritis in Lesotho. *Ann Rheum Dis,* 43**,** 40-3.
Mulcahy, B., Waldron-Lynch, F., Mcdermott, M. F., Adams, C., Amos, C. I., Zhu, D. K.*, et al.* (1996) Genetic variability in the tumor necrosis factor-lymphotoxin region influences susceptibility to rheumatoid arthritis. *Am J Hum Genet,* 59**,** 676-83.
Murai, C., Munakata, Y., Takahashi, Y., Ishii, T., Shibata, S., Muryoi, T.*, et al.* (1999) Rheumatoid arthritis after human parvovirus B19 infection. *Ann Rheum Dis,* 58**,** 130-2.
Nakazawa, M., Ishii, H., Nakamura, H., Yoshino, S. I., Fukamizu, A., Nishioka, K.*, et al.* (2001) NFkappaB2 (p52) promoter activation via Notch signaling pathway in rheumatoid synoviocytes. *Int J Mol Med,* 7**,** 31-5.
Newton, J. L., Harney, S. M., Wordsworth, B. P. & Brown, M. A. (2004) A review of the MHC genetics of rheumatoid arthritis. *Genes Immun,* 5**,** 151-7.
Nielen, M. M., Van Schaardenburg, D., Lems, W. F., Van De Stadt, R. J., De Koning, M. H., Reesink, H. W.*, et al.* (2006) Vitamin D deficiency does not increase the risk of rheumatoid arthritis: comment on the article by Merlino et al. *Arthritis Rheum,* 54**,** 3719-20.
Niewold, T. B., Harrison, M. J. & Paget, S. A. (2007) Anti-CCP antibody testing as a diagnostic and prognostic tool in rheumatoid arthritis. *QJM,* 100**,** 193-201.
Nyhall-Wahlin, B. M., Petersson, I. F., Nilsson, J. A., Jacobsson, L. T. & Turesson, C. (2009) High disease activity disability burden and smoking predict severe extra-articular manifestations in early rheumatoid arthritis. *Rheumatology (Oxford),* 48**,** 416-20.
Oen, K., Malleson, P. N., Cabral, D. A., Rosenberg, A. M., Petty, R. E., Nickerson, P.*, et al.* (2005) Cytokine genotypes correlate with pain and radiologically defined joint damage in patients with juvenile rheumatoid arthritis. *Rheumatology (Oxford),* 44**,** 1115-21.
Okada, Y., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A.*, et al.* (2012) Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nat Genet,* 44**,** 511-6.
Ollier, W. (2000) Rheumatoid arthritis and Epstein-Barr virus: a case of living with the enemy? *Ann Rheum Dis,* 59**,** 497-9.
Orozco, G., Abelson, A. K., Gonzalez-Gay, M. A., Balsa, A., Pascual-Salcedo, D., Garcia, A.*, et al.* (2009) Study of functional variants of the BANK1 gene in rheumatoid arthritis. *Arthritis Rheum,* 60**,** 372-9.
Orozco, G., Eyre, S., Hinks, A., Ke, X., Wilson, A. G., Bax, D. E.*, et al.* (2010) Association of CD40 with rheumatoid arthritis confirmed in a large UK case-control study. *Ann Rheum Dis,* 69**,** 813-6.
Orozco, G., Sanchez, E., Gonzalez-Gay, M. A., Lopez-Nevot, M. A., Torres, B., Caliz, R.*, et al.* (2005) Association of a functional single-nucleotide

polymorphism of PTPN22, encoding lymphoid protein phosphatase, with rheumatoid arthritis and systemic lupus erythematosus. *Arthritis Rheum,* 52**,** 219-24.

Ouedraogo, D. D., Singbo, J., Diallo, O., Sawadogo, S. A., Tieno, H. & Drabo, Y. J. (2011) Rheumatoid arthritis in Burkina Faso: clinical and serological profiles. *Clin Rheumatol,* 30**,** 1617-21.

Padyukov, L., Silva, C., Stolt, P., Alfredsson, L. & Klareskog, L. (2004) A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. *Arthritis Rheum,* 50**,** 3085-92.

Patsopoulos, N. A. & Ioannidis, J. P. (2010) Susceptibility variants for rheumatoid arthritis in the TRAF1-C5 and 6q23 loci: a meta-analysis. *Ann Rheum Dis,* 69**,** 561-6.

Pattison, D. J., Symmons, D. P., Lunt, M., Welch, A., Bingham, S. A., Day, N. E.*, et al.* (2005) Dietary beta-cryptoxanthin and inflammatory polyarthritis: results from a population-based prospective study. *Am J Clin Nutr,* 82**,** 451-5.

Pattison, D. J., Symmons, D. P., Lunt, M., Welch, A., Luben, R., Bingham, S. A.*, et al.* (2004) Dietary risk factors for the development of inflammatory polyarthritis: evidence for a role of high level of red meat consumption. *Arthritis Rheum,* 50**,** 3804-12.

Pawlik, A., Kurzawski, M., Florczak, M., Gawronska Szklarz, B. & Herczynska, M. (2005a) IL1beta+3953 exon 5 and IL-2 -330 promoter polymorphisms in patients with rheumatoid arthritis. *Clin Exp Rheumatol,* 23**,** 159-64.

Pawlik, A., Kurzawski, M., Szklarz, B. G., Herczynska, M. & Drozdzik, M. (2005b) Interleukin-10 promoter polymorphism in patients with rheumatoid arthritis. *Clin Rheumatol,* 24**,** 480-4.

Pedersen, M., Jacobsen, S., Klarlund, M. & Frisch, M. (2006) Socioeconomic status and risk of rheumatoid arthritis: a Danish case-control study. *J Rheumatol,* 33**,** 1069-74.

Peer, N., Bradshaw, D., Laubscher, R., Steyn, N. & Steyn, K. (2013) Urban-rural and gender differences in tobacco and alcohol use, diet and physical activity among young black South Africans between 1998 and 2003. *Glob Health Action,* 6**,** 19216.

Pierer, M., Kaltenhauser, S., Arnold, S., Wahle, M., Baerwald, C., Hantzschel, H.*, et al.* (2006) Association of PTPN22 1858 single-nucleotide polymorphism with rheumatoid arthritis in a German cohort: higher frequency of the risk allele in male compared to female patients. *Arthritis Res Ther,* 8**,** R75.

Pile, K. D., Tikly, M., Bell, J. I. & Wordsworth, B. P. (1992) HLA-DR antigens and rheumatoid arthritis in black South Africans: a study of ethnic groups. *Tissue Antigens,* 39**,** 138-40.

Pincus, T., Brooks, R. H. & Callahan, L. F. (1994) Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. *Ann Intern Med,* 120**,** 26-34.

Pincus, T., Callahan, L. F., Sale, W. G., Brooks, A. L., Payne, L. E. & Vaughn, W. K. (1984) Severe functional declines, work disability, and increased mortality in seventy-five rheumatoid arthritis patients studied over nine years. *Arthritis Rheum,* 27**,** 864-72.

Pincus, T., Summey, J. A., Soraci, S. A., Jr., Wallston, K. A. & Hummon, N. P. (1983) Assessment of patient satisfaction in activities of daily living using a

modified Stanford Health Assessment Questionnaire. *Arthritis Rheum,* 26**,** 1346-53.
Plenge, R. M., Cotsapas, C., Davies, L., Price, A. L., De Bakker, P. I., Maller, J.*, et al.* (2007a) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet,* 39**,** 1477-82.
Plenge, R. M., Padyukov, L., Remmers, E. F., Purcell, S., Lee, A. T., Karlson, E. W.*, 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**,** 1044-60.
Plenge, R. M., Seielstad, M., Padyukov, L., Lee, A. T., Remmers, E. F., Ding, B.*, et al.* (2007b) TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *N Engl J Med,* 357**,** 1199-209.
Pope, R. M. (2002) Apoptosis as a therapeutic tool in rheumatoid arthritis. *Nat Rev Immunol, 2***,** 527-35.
Prasad, P., Kumar, A., Gupta, R., Juyal, R. C. & Thelma, B. K. (2012) Caucasian and Asian specific rheumatoid arthritis risk loci reveal limited replication and apparent allelic heterogeneity in north Indians. *PLoS One,* 7**,** e31584.
Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38**,** 904-9.
Ramsay, M. (2012) Africa: continent of genome contrasts with implications for biomedical research and health. *FEBS Lett,* 586**,** 2813-9.
Ravindran, V., Seah, M. A., Elias, D. A., Choy, E. H., Scott, D. L. & Gordon, P. A. (2008) Clinical and radiological features of rheumatoid arthritis in British black Africans. *Clin Rheumatol,* 27**,** 97-100.
Raychaudhuri, S., Sandor, C., Stahl, E. A., Freudenberg, J., Lee, H. S., Jia, X.*, et al.* (2012) Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet,* 44**,** 291-6.
Raychaudhuri, S., Thomson, B. P., Remmers, E. F., Eyre, S., Hinks, A., Guiducci, C.*, et al.* (2009) Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nat Genet,* 41**,** 1313-8.
Remmers, E. F., Plenge, R. M., Lee, A. T., Graham, R. R., Hom, G., Behrens, T. W.*, et al.* (2007) STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med,* 357**,** 977-86.
Richter, L. M., Panday, S. & Norris, S. A. (2009) Factors influencing enrollment: a case study from Birth to Twenty, the 1990 birth cohort in Soweto-Johannesburg. *Eval Program Plann,* 32**,** 197-203.
Ruyssen-Witrand, A., Constantin, A., Cambon-Thomsen, A. & Thomsen, M. (2012) New insights into the genetics of immune responses in rheumatoid arthritis. *Tissue Antigens,* 80**,** 105-18.
Saal, J. G., Krimmel, M., Steidle, M., Gerneth, F., Wagner, S., Fritz, P.*, et al.* (1999) Synovial Epstein-Barr virus infection increases the risk of rheumatoid arthritis in individuals with the shared HLA-DR4 epitope. *Arthritis Rheum,* 42**,** 1485-96.
Schett, G., Zwerina, J. & Firestein, G. (2008) The p38 mitogen-activated protein kinase (MAPK) pathway in rheumatoid arthritis. *Ann Rheum Dis,* 67**,** 909-16.

Schneider, M., Manabile, E. & Tikly, M. (2008) Social aspects of living with rheumatoid arthritis: a qualitative descriptive study in Soweto, South Africa - a low resource context. *Health Qual Life Outcomes,* 6**,** 54.
Seldin, M. F., Amos, C. I., Ward, R. & Gregersen, P. K. (1999) The genetics revolution and the assault on rheumatoid arthritis. *Arthritis Rheum,* 42**,** 1071-9.
Shaper, A. G. & Shaper, L. (1958) Analysis of medical admissions to Mulago Hospital, 1957. *East Afr Med J,* 35**,** 647-78.
Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. (2009) The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet,* 54**,** 15-39.
Shiozawa, S., Hayashi, S., Tsukamoto, Y., Goko, H., Kawasaki, H., Wada, T.*, et al.* (1998) Identification of the gene loci that predispose to rheumatoid arthritis. *Int Immunol,* 10**,** 1891-5.
Shmerling, R. H. & Delbanco, T. L. (1992) How useful is the rheumatoid factor? An analysis of sensitivity, specificity, and predictive value. *Arch Intern Med,* 152**,** 2417-20.
Silman, A. J., Macgregor, A. J., Thomson, W., Holligan, S., Carthy, D., Farhan, A.*, et al.* (1993) Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Br J Rheumatol,* 32**,** 903-7.
Silman, A. J., Newman, J. & Macgregor, A. J. (1996) Cigarette smoking increases the risk of rheumatoid arthritis. Results from a nationwide study of disease-discordant twins. *Arthritis Rheum,* 39**,** 732-5.
Silman, A. J. & Pearson, J. E. (2002) Epidemiology and genetics of rheumatoid arthritis. *Arthritis Res,* 4 Suppl 3**,** S265-72.
Singwe-Ngandeu, M., Finckh, A., Bas, S., Tiercy, J. M. & Gabay, C. (2010) Diagnostic value of anti-cyclic citrullinated peptides and association with HLA-DRB1 shared epitope alleles in African rheumatoid arthritis patients. *Arthritis Res Ther,* 12**,** R36.
Skorka, A., Bednarczuk, T., Bar-Andziak, E., Nauman, J. & Ploski, R. (2005) Lymphoid tyrosine phosphatase (PTPN22/LYP) variant and Graves' disease in a Polish population: association and gene dose-dependent correlation with age of onset. *Clin Endocrinol (Oxf),* 62**,** 679-82.
Smyth, D., Cooper, J. D., Collins, J. E., Heward, J. M., Franklyn, J. A., Howson, J. M.*, et al.* (2004) Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes,* 53**,** 3020-3.
Solomon, A., Christian, B. F., Dessein, P. H. & Stanwix, A. E. (2005) The need for tighter rheumatoid arthritis control in a South African public health care center. *Semin Arthritis Rheum,* 35**,** 122-31.
Solomon, L., Beighton, P., Valkenburg, H. A., Robin, G. & Soskolne, C. L. (1975) Rheumatic disorders in the South African Negro. Part I. Rheumatoid arthritis and ankylosing spondylitis. *S Afr Med J,* 49**,** 1292-6.
Spielman, R. S., Mcginnis, R. E. & Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet,* 52**,** 506-16.
Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P.*, et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet,* 42**,** 508-14.

Stanczyk, J., Pedrioli, D. M., Brentano, F., Sanchez-Pernaute, O., Kolling, C., Gay, R. E.*, et al.* (2008) Altered expression of MicroRNA in synovial fibroblasts and synovial tissue in rheumatoid arthritis. *Arthritis Rheum,* 58**,** 1001-9.
Stastny, P. (1976) Mixed lymphocyte cultures in rheumatoid arthritis. *J Clin Invest,* 57**,** 1148-57.
Steer, S., Lad, B., Grumley, J. A., Kingsley, G. H. & Fisher, S. A. (2005) Association of R602W in a protein tyrosine phosphatase gene with a high risk of rheumatoid arthritis in a British population: evidence for an early onset/disease severity effect. *Arthritis Rheum,* 52**,** 358-60.
Stolt, P., Bengtsson, C., Nordmark, B., Lindblad, S., Lundberg, I., Klareskog, L.*, et al.* (2003) Quantification of the influence of cigarette smoking on rheumatoid arthritis: results from a population based case-control study, using incident cases. *Ann Rheum Dis,* 62**,** 835-41.
Strominger, J. L. (1987) Structure of class I and class II HLA antigens. *Br Med Bull,* 43**,** 81-93.
Sugiyama, D., Nishimura, K., Tamaki, K., Tsuji, G., Nakazawa, T., Morinobu, A.*, et al.* (2010) Impact of smoking as a risk factor for developing rheumatoid arthritis: a meta-analysis of observational studies. *Ann Rheum Dis,* 69**,** 70-81.
Suppiah, V., O'doherty, C., Heggarty, S., Patterson, C. C., Rooney, M. & Vandenbroeck, K. (2006) The CTLA4+49A/G and CT60 polymorphisms and chronic inflammatory arthropathies in Northern Ireland. *Exp Mol Pathol,* 80**,** 141-6.
Suzuki, A., Yamada, R., Chang, X., Tokuhiro, S., Sawada, T., Suzuki, M.*, et al.* (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat Genet,* 34**,** 395-402.
Suzuki, A., Yamada, R., Kochi, Y., Sawada, T., Okada, Y., Matsuda, K.*, et al.* (2008) Functional SNPs in CD244 increase the risk of rheumatoid arthritis in a Japanese population. *Nat Genet,* 40**,** 1224-9.
Sverdrup, B., Kallberg, H., Bengtsson, C., Lundberg, I., Padyukov, L., Alfredsson, L.*, et al.* (2005) Association between occupational exposure to mineral oil and rheumatoid arthritis: results from the Swedish EIRA case-control study. *Arthritis Res Ther,* 7**,** R1296-303.
Tait, B. D., Drummond, B. P., Varney, M. D. & Harrison, L. C. (1995) HLA-DRB1*0401 is associated with susceptibility to insulin-dependent diabetes mellitus independently of the DQB1 locus. *Eur J Immunogenet,* 22**,** 289-97.
Taneja, V., Mehra, N. K., Kailash, S., Anand, C. & Malaviya, A. N. (1992) Protective & risk DR phenotypes in Asian Indian patients with rheumatoid arthritis. *Indian J Med Res,* 96**,** 16-23.
Teo, Y. Y., Small, K. S. & Kwiatkowski, D. P. (2010) Methodological challenges of genome-wide association analysis in Africa. *Nat Rev Genet,* 11**,** 149-60.
Thabet, M. M., Huizinga, T. W., Marques, R. B., Stoeken-Rijsbergen, G., Bakker, A. M., Kurreeman, F. A.*, et al.* (2009) Contribution of Fcgamma receptor IIIA gene 158V/F polymorphism and copy number variation to the risk of ACPA-positive rheumatoid arthritis. *Ann Rheum Dis,* 68**,** 1775-80.
Thomson, W., Barton, A., Ke, X., Eyre, S., Hinks, A., Bowes, J.*, et al.* (2007) Rheumatoid arthritis association at 6q23. *Nat Genet,* 39**,** 1431-3.

Thorsby, E., Pfeffer, P. & Leivestad, T. (2004) Role of HLA molecules in the induction of alloimmune responses: clinical significance in the cyclosporine era. *Transplant Proc,* 36**,** 16S-21S.
Tikly, M., Govind, N., Frost, J. & Ramsay, M. (2010) The PTPN22 R620W polymorphism is not associated with systemic rheumatic diseases in South Africans. *Rheumatology (Oxford),* 49**,** 820-1.
Tikly, M., Zannettou, N. & Hopley, M. (2003) A longitudinal study of rheumatoid arthritis in South Africans. *MedGenMed,* 5**,** 2.
Tishkoff, S. A. & Kidd, K. K. (2004) Implications of biogeography of human populations for 'race' and medicine. *Nat Genet,* 36**,** S21-7.
Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A.*, et al.* (2009) The genetic structure and history of Africans and African Americans. *Science,* 324**,** 1035-44.
Tishkoff, S. A. & Verrelli, B. C. (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet,* 4**,** 293-340.
Tishkoff, S. A. & Williams, S. M. (2002) Genetic analysis of African populations: human evolution and complex disease. *Nat Rev Genet,* 3**,** 611-21.
Tokuhiro, S., Yamada, R., Chang, X., Suzuki, A., Kochi, Y., Sawada, T.*, et al.* (2003) An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat Genet,* 35**,** 341-8.
Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A.*, et al.* (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet,* 43**,** 1193-201.
Tsoulfa, G., Rook, G. A., Van-Embden, J. D., Young, D. B., Mehlert, A., Isenberg, D. A.*, et al.* (1989) Raised serum IgG and IgA antibodies to mycobacterial antigens in rheumatoid arthritis. *Ann Rheum Dis,* 48**,** 118-23.
Uddin, M., Sturge, M., Rahman, P. & Woods, M. O. (2011) Autosome-wide copy number variation association analysis for rheumatoid arthritis using the WTCCC high-density SNP genotype data. *J Rheumatol,* 38**,** 797-801.
Uhlig, T., Hagen, K. B. & Kvien, T. K. (1999) Current tobacco smoking, formal education, and the risk of rheumatoid arthritis. *J Rheumatol,* 26**,** 47-54.
Van Der Linden, M. P., Feitsma, A. L., Le Cessie, S., Kern, M., Olsson, L. M., Raychaudhuri, S.*, et al.* (2009) Association of a single-nucleotide polymorphism in CD40 with the rate of joint destruction in rheumatoid arthritis. *Arthritis Rheum,* 60**,** 2242-7.
Veeramah, K. R., Wegmann, D., Woerner, A., Mendez, F. L., Watkins, J. C., Destro-Bisol, G.*, et al.* (2012) An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol Biol Evol,* 29**,** 617-30.
Velaga, M. R., Wilson, V., Jennings, C. E., Owen, C. J., Herington, S., Donaldson, P. T.*, et al.* (2004) The codon 620 tryptophan allele of the lymphoid tyrosine phosphatase (LYP) gene is a major determinant of Graves' disease. *J Clin Endocrinol Metab,* 89**,** 5862-5.
Verrelli, B. C. & Tishkoff, S. A. (2004) Signatures of selection and gene conversion associated with human color vision variation. *Am J Hum Genet,* 75**,** 363-75.
Viatte, S., Flynn, E., Lunt, M., Barnes, J., Singwe-Ngandeu, M., Bas, S.*, et al.* (2012) Investigation of Caucasian rheumatoid arthritis susceptibility loci in African patients with the same disease. *Arthritis Res Ther,* 14**,** R239.

Weyand, C. M., Hunder, N. N., Hicok, K. C., Hunder, G. G. & Goronzy, J. J. (1994) HLA-DRB1 alleles in polymyalgia rheumatica, giant cell arteritis, and rheumatoid arthritis. *Arthritis Rheum,* 37**,** 514-20.

Wucherpfennig, K. W. & Strominger, J. L. (1995) Selective binding of self peptides to disease-associated major histocompatibility complex (MHC) molecules: a mechanism for MHC-linked susceptibility to human autoimmune diseases. *J Exp Med,* 181**,** 1597-601.

Zapico, I., Coto, E., Rodriguez, A., Alvarez, C., Torre, J. C. & Alvarez, V. (2000) CCR5 (chemokine receptor-5) DNA-polymorphism influences the severity of rheumatoid arthritis. *Genes Immun,* 1**,** 288-9.

Zavala-Ruiz, Z., Strug, I., Anderson, M. W., Gorski, J. & Stern, L. J. (2004) A polymorphic pocket at the P10 position contributes to peptide binding specificity in class II MHC proteins. *Chem Biol,* 11**,** 1395-402.

Zheng, W. & She, J. X. (2005) Genetic association between a lymphoid tyrosine phosphatase (PTPN22) and type 1 diabetes. *Diabetes,* 54**,** 906-8.

SUPPLEMENTARY INFORMATION

Supplementary Table 1 Significantly associated SNPs (p < 5 x 10-8) in the HLA region on chromosome 6

SNP	*BP	A1	**MAF Cases	**MAF Controls	A2	Chi sq	P	OR	Gene	region
rs9368726	32438542	G	0.3887	0.1743	A	72.27	1.88E-17	3.013	*HLA-DRA|HLA-DRB5*	intergenic
rs9268835	32428115	A	0.4077	0.1903	G	71.86	2.31E-17	2.928	*HLA-DRA|HLA-DRB5*	intergenic
rs9275332	32666943	A	0.5191	0.2855	G	71.13	3.34E-17	2.701	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275338	32667343	T	0.3697	0.1622	A	70.99	3.58E-17	3.03	*HLA-DQB1|HLA-DQA2*	intergenic
rs9271488	32589000	A	0.3569	0.1555	C	68.71	1.14E-16	3.014	*HLA-DRB1|HLA-DQA1*	intergenic
rs9271588	32590953	G	0.5115	0.2836	A	67.91	1.72E-16	2.644	*HLA-DRB1|HLA-DQA1*	intergenic
rs3104389	32595097	A	0.3544	0.1555	C	67.12	2.56E-16	2.981	*HLA-DRB1|HLA-DQA1*	intergenic
rs9275425	32670874	A	0.51	0.2832	C	65.55	5.67E-16	2.634	*HLA-DQB1|HLA-DQA2*	intergenic
rs477515	32569691	A	0.3582	0.1609	G	65.2	6.76E-16	2.912	*HLA-DRB1|HLA-DQA1*	intergenic
rs660895	32577380	G	0.3378	0.1475	A	63.75	1.41E-15	2.949	*HLA-DRB1|HLA-DQA1*	intergenic
rs3021061	32651839	G	0.4244	0.2158	C	63.07	1.99E-15	2.679	*HLA-DQB1|HLA-DQA2*	intergenic
rs521539	32581973	A	0.3352	0.1465	G	62.8	2.28E-15	2.938	*HLA-DRB1|HLA-DQA1*	intergenic
rs2856691	32653435	A	0.4804	0.2638	G	61.43	4.59E-15	2.58	*HLA-DQB1|HLA-DQA2*	intergenic
rs4947342	32653070	A	0.4195	0.2158	G	60.73	6.54E-15	2.626	*HLA-DQB1|HLA-DQA2*	intergenic
rs7774434	32657578	G	0.5595	0.3387	A	59.85	1.03E-14	2.48	*HLA-DQB1|HLA-DQA2*	intergenic

rs9275313	32665759	A	0.1609	0.04313	C	51.01	9.19E-13	4.255	*HLA-DQB1|HLA-DQA2*	intergenic
rs9272358	32604538	A	0.4748	0.2842	G	47.58	5.29E-12	2.278	*HLA-DRB1|HLA-DQA1*	intergenic
rs9275334	32667107	G	0.1456	0.04178	A	42.62	6.64E-11	3.908	*HLA-DQB1|HLA-DQA2*	intergenic
rs482044	32576064	G	0.5288	0.3446	C	42.54	6.91E-11	2.135	*HLA-DRB1|HLA-DQA1*	intergenic
rs9275495	32673574	T	0.145	0.0429	A	41.27	1.32E-10	3.785	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275530	32675523	C	0.145	0.0429	G	41.27	1.32E-10	3.785	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275532	32675634	G	0.145	0.0429	C	41.27	1.32E-10	3.785	*HLA-DQB1|HLA-DQA2*	intergenic
rs3957146	32681530	G	0.145	0.0429	A	41.27	1.32E-10	3.785	*HLA-DQB1|HLA-DQA2*	intergenic
rs7454108	32681483	G	0.145	0.04301	A	41.07	1.47E-10	3.775	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275184	32654714	G	0.1325	0.03641	A	38.64	5.10E-10	4.043	*HLA-DQB1|HLA-DQA2*	intergenic
rs3129882	32409530	G	0.2996	0.4732	A	38.59	5.24E-10	0.476	*HLA-DRA*	intron
rs3998159	32682019	C	0.1409	0.0429	A	38.57	5.28E-10	3.66	*HLA-DQB1|HLA-DQA2*	intergenic
rs9271850	32595060	G	0.281	0.4536	A	38.16	6.52E-10	0.470	*HLA-DRB1|HLA-DQA1*	intergenic
rs6457617	32663851	G	0.3552	0.5309	A	37.97	7.19E-10	0.486	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275224	32659878	A	0.3544	0.5282	G	37.36	9.80E-10	0.490	*HLA-DQB1|HLA-DQA2*	intergenic
rs6457620	32663999	G	0.3569	0.5296	C	36.94	1.22E-09	0.492	*HLA-DQB1|HLA-DQA2*	intergenic
rs2647012	32664458	A	0.3053	0.473	G	35.88	2.10E-09	0.489	*HLA-DQB1|HLA-DQA2*	intergenic
rs241406	32863143	C	0.07443	0.01072	A	35.05	3.22E-09	7.418	*PSMB9|HLA-DMB*	intergenic
rs6911777	32409996	G	0.3263	0.1823	A	34.84	3.59E-09	2.173	*HLA-DRA*	intron
rs2858324	32660375	A	0.3046	0.4688	G	34.27	4.80E-09	0.496	*HLA-DQB1|HLA-DQA2*	intergenic

rs5007263	32378982	A	0.3525	0.5174	G	33.76	6.23E-09	0.507	*BTNL2|HLA-DRA*	intergenic
rs5007259	32379101	A	0.3525	0.5174	G	33.76	6.23E-09	0.507	*BTNL2|HLA-DRA*	intergenic
rs35198051	70845204	A	0.05792	0.1662	G	33.52	7.06E-09	0.308	*C11orf76|LOC100133306*	intergenic
rs9275206	32657565	G	0.1284	0.04032	A	33.46	7.28E-09	3.506	*HLA-DQB1|HLA-DQA2*	intergenic
rs2856717	32670308	A	0.2977	0.4584	G	33.36	7.65E-09	0.500	*HLA-DQB1|HLA-DQA2*	intergenic
rs9268507	32377539	A	0.3525	0.5161	G	33.21	8.28E-09	0.510	*BTNL2|HLA-DRA*	intergenic
rs6932542	32380262	A	0.3525	0.5161	G	33.21	8.28E-09	0.510	*BTNL2|HLA-DRA*	intergenic
rs9275596	32681631	G	0.2958	0.4558	A	33.11	8.69E-09	0.501	*HLA-DQB1|HLA-DQA2*	intergenic
rs5007265	32378866	A	0.355	0.5174	C	32.82	1.01E-08	0.513	*BTNL2|HLA-DRA*	intergenic
rs7754768	32420179	A	0.5498	0.3892	G	31.84	1.68E-08	1.917	*HLA-DRA|HLA-DRB5*	intergenic
rs2239803	32411833	A	0.5496	0.3901	G	31.58	1.91E-08	1.908	*HLA-DRA*	intron
rs9275582	32680070	A	0.2977	0.1662	G	30.96	2.63E-08	2.126	*HLA-DQB1|HLA-DQA2*	intergenic
rs4502931	32380782	T	0.3544	0.5107	A	30.37	3.58E-08	0.525	*BTNL2|HLA-DRA*	intergenic
rs9275580	32679462	G	0.2977	0.1676	A	30.26	3.78E-08	2.106	*HLA-DQB1|HLA-DQA2*	intergenic
rs1612904	32669018	C	0.3036	0.4602	A	30.12	4.06E-08	0.511	*HLA-DQB1|HLA-DQA2*	intergenic
rs9275578	32679384	A	0.2946	0.1662	C	29.41	5.86E-08	2.095	*HLA-DQB1|HLA-DQA2*	intergenic
rs2097432	32590771	G	0.32	0.4732	A	28.48	9.46E-08	0.523	*HLA-DRB1|HLA-DQA1*	intergenic
rs9275766	32689362	G	0.1603	0.06702	A	28.38	9.98E-08	2.657	*HLA-DQB1|HLA-DQA2*	intergenic

.UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Dr Nimmisha H Govind

CLEARANCE CERTIFICATE	M10707
PROJECT	The Genetics of Rheumatoid Arthritis in Black South Africans (revised title)
INVESTIGATORS	Dr Nimmisha H Govind
DEPARTMENT	Department of Medicine/Rheumatology
DATE CONSIDERED	30/07/2010
DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 18/01/2013 **CHAIRPERSON** .....
(Professor P E Cleaton Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof M Tikly

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES