

https://dspace.org/

Rheumatoid arthritis in black South Africans: genomic susceptibility and methotrexate response

Mathebula, Evans Mantiri

2020
https://hdl.handle.net/10539/29951

Downloaded from a DSpace repository. https://dspace.org/

RHEUMATOID ARTHRITIS IN BLACK SOUTH AFRICANS: GENOMIC SUSCEPTIBILITY AND METHOTREXATE RESPONSE



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Evans Mantiri Mathebula

1701006

A Thesis Submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

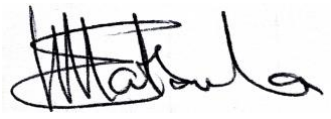
Johannesburg, 2020

DECLARATION

I, Evans Mantiri Mathebula (1701006), am a student registered for the degree of Doctor of Philosophy from the academic year 2016.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other University.
- The information used in the thesis has not been obtained by me while employed by or working under the aegis of, any person or organisation other than the University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: 

Date:__23/04/2020_____

DEDICATION

This Thesis is dedicated to my lovely wife Thembisile and kids Ndzalama and Xiluva, without whom this work would not have been possible. Thank you for giving me time to pursue my dreams and complete this work. I shall forever remain indebted to your sacrifices.

Thank you so much!

PRESENTATIONS AND PUBLICATIONS ARISING FROM THIS PROJECT

Presentations:

Evans M. Mathebula, Ananyo Choudhury, Nimmisha Govind, Tikly Mohammed, Michele Ramsay. “Rheumatoid Arthritis in Black South Africans: Genomic Susceptibility and Methotrexate Response” (Poster). 10th Conference of the African Society of Human Genetics. Cairo, Egypt, 2017 (16 - 18 November).

Evans M. Mathebula, Ananyo Choudhury, Nimmisha Govind, Tikly Mohammed, Michele Ramsay. ”Rheumatoid Arthritis in Black South Africans: Genomic Susceptibility and Methotrexate Response” (Oral). 1st WASLITBRe Postgraduate Bioinformatics Workshop/ 1st WASLITBRe Consortium and TAC Meeting. Ota state, Nigeria, 2018 (08 May - 22 June).

Evans M. Mathebula, Ananyo Choudhury, Nimmisha Govind, Tikly Mohammed, Michele Ramsay. “A Genome-Wide Association Study of Rheumatoid Arthritis in Black South Africans” (Oral). South African Rheumatism Arthritis Association Congress 2019, Durban, South Africa, 2019 (07 - 10 March).

Evans M. Mathebula, Ananyo Choudhury, Tikly Mohammed, Michele Ramsay, Nimmisha Govind. “Response to Methotrexate in the treatment of Rheumatoid Arthritis in influenced by genetic variants on Methotrexate Polyglutamation gene polymorphisms” (Poster first prize winner). South African Rheumatism Arthritis Association Congress 2019, Durban, South Africa, 2019 (07 – 10 March).

Publications:

Evans M. Mathebula, Ananyo Choudhury, Nimmisha Govind, Tikly Mohammed, Michele Ramsay. “A genome-wide association study identifies a novel risk locus for rheumatoid arthritis in Black South Africans”. (Advanced draft).

Evans M. Mathebula, Ananyo Choudhury, Tikly Mohammed, Michele Ramsay, Nimmisha Govind. “. A replication study in black South African rheumatoid arthritis patients highlights the role of methotrexate polyglutamation gene in methotrexate monotherapy treatment outcome’. (Advanced draft).

ABSTRACT

Rheumatoid arthritis (RA) is a complex chronic inflammatory joint disease with no cure. The heritability of RA is estimated to be 60% and no large genome-wide association studies have been performed in African populations. RA is treated using methotrexate (MTX), an antifolate disease-modifying antirheumatic drug. Previous studies have shown that variations in MTX polyglutamation genes play a role in MTX response in RA patients. However, very little is known about the RA aetiopathogenesis and pharmacogenetics of MTX use in people of African ancestry. This study used a case-control genome-wide association study (GWAS) approach to assess genetic susceptibility to RA in 577 Black South Africans (BSA) representing seropositive RA cases and 1,612 population controls. The cases and controls were genotyped on the H3Africa consortium genotyping SNP array and imputed using the African Genome Resources reference panel. The association was tested using a logistic regression analysis assuming an additive model and adjusted for sex, age, smoking, and the first three principal components for possible population structure. This study confirmed the strong association between increased risk and genetic variants in the Human Leucocyte Antigen (HLA) region. Furthermore, this study identified four novel single nucleotide polymorphism (SNP) associations in the intron regions of the *CPT1A* gene, which has been shown to be differentially expressed in RA patients, compared to unaffected controls. GWAS results were followed by integration with functional annotation of variants obtained from data related to tissues and cells relevant to RA pathogenesis, revealing that the associated SNPs are likely involved in the regulation of the expression of *CPT1A* and other nearby genes.

In addition to the GWAS, this study investigated 247 BSA seropositive RA patients on MTX monotherapy to evaluate the association of 36 previously identified index SNPs spanning 19 genes, and an additional 7812 SNPs in the surrounding regions, with MTX response. Patients were treatment naïve at baseline and were categorised after 6 months of MTX treatment into two groups of either EULAR good/moderate responders (n=186) or poor responders (n=61). A logistic regression analysis with an additive model adjusting for sex, age, smoking and disease duration was used for association testing. The regional replication of previously reported MTX response loci identified a region on chromosome 9 with 14 significantly associated SNPs mapping to the intronic region of the *CDK9* gene, of which four (rs1002095; rs2297214; rs3217740, rs3217751) are expression quantitative trait loci (eQTL) that potentially influence the expression of the folylpolyglutamate synthetase (*FPGS)* gene

involved in MTX polyglutamation and intracellular persistence of MTX polyglutamation. The regulation of the expression of the *FPGS* gene by eQTL variants mapping to the *CDK9* gene may influence response to MTX therapy in BSA with seropositive RA.

Although the sample size in this study had limited power to test many of the previously identified RA genetic associations with confidence, our results indicate that the pharmacokinetics of MTX might have a different genetic underpinning in African populations and underlines the necessity of larger studies in populations from Africa. Our study is the first and largest GWAS of RA and pharmacogenetics of MTX in sub-Saharan Africa.

ACKNOWLEDGEMENTS

"If I have seen further, it is by standing upon the shoulders of giants"

– *Isaac Newton*

This body of work is as a result of a great deal of help and support from the following people and institutions whom I would like to acknowledge:

- My sincere gratitude and appreciation to my supervisor collective for agreeing to work with me in this interesting and challenging project:
    - Professor Michele Ramsay, for accepting me into the SBIMB research group and giving me an opportunity to learn from her great wealth of experience in the field of Human Genetics. Thanks for the support and guidance during the research and writing up of this thesis.
    - Dr Nimmisha Govind for guidance and support throughout the research project and thesis write up. Her extensive experience and intimate knowledge of rheumatoid arthritis and methotrexate treatment have been an invaluable asset to provide me with advice throughout the project.
    - Professor Mohammed Tikly for agreeing to stay on after retirement to provide guidance and providing me with thoughtful insights into rheumatoid arthritis genetics and the pharmacogenetics of methotrexate response during the course of my research.
    - Dr Ananyo Choudhury for your continuous support in Bioinformatics, data analysis, guidance and support through the project and writing up of thesis.

- My sincere thanks to the Chris Hani Baragwanath clinic staff, particularly Audrey Twala, for assisting with sample and data collection, sometimes at odd times of the week.
- My sincerest thanks to the Biobank Staff, particularly Cassandra Soo, for assisting with DNA extraction training and assisting with sample shipping logistics.
- My great thanks to the AWI-Gen participants and data collection teams.

- My great thanks to the funding bodies National Research Foundation (NRF), Wits Connective Tissue Disease research fund, Prof Ramsay, Prof Tikly and Dr. Govind, WASLITBRe Consortium for giving me an opportunity to do the research and travel to present at conferences and learn from other researchers.
- Thank you Professor Scott Hazelhurst, Dr Jean-Tristan Brandenburg and Phelelani Mpangase for technical assistance with the computing cluster and bioinformatics pipeline.
- Many thanks to Dr Dhriti Sengupta for assistance with genotype data imputation.
- My deepest thanks to my good friend and fellow Ph.D. candidate Romuald Boua for the long nights and weekends spent discussing and encouraging each other to carry on. May God bless him richly.
- The project office for assisting with traveling logistics and many other things.

Finally, I would like to thank my fellow students for assistance, courage, and guidance throughout the project.

TABLE OF CONTENTS

DECLARATION ..... *i*
DEDICATION ..... *ii*
PRESENTATIONS AND PUBLICATIONS ARISING FROM THIS PROJECT ..... *iii*
ABSTRACT ..... *iv*
ACKNOWLEDGEMENTS ..... *vi*
TABLE OF CONTENTS ..... *viii*
LIST OF FIGURES ..... *xi*
LIST OF TABLES ..... *xiii*
ABBREVIATIONS ..... *xiv*
CHAPTER 1: LITERATURE REVIEW AND INTRODUCTION *1*
1.1 Rheumatoid arthritis **2**
1.2 Classification and Diagnosis of RA **3**
1.2.1 RA history 3
1.2.2 RA classification criteria 3
1.3 Epidemiology of RA **5**
1.3.1 Worldwide 5
1.3.2 Africa 6
1.4 Pathophysiology **7**
1.4.1 Pre-clinical phase 8
1.4.2 Disease onset and inflammation 9
1.4.3 Joint damage 9
1.4.4 Extra-articular manifestation 10
1.5 Risk Factors **10**
1.5.1 Environmental factors 11
1.5.2 Genetic factors 15
1.6 Genetic association studies of RA **16**
1.6.1 HLA association with RA 16
1.6.2 Non-HLA genes association 20
1.7 RA treatment **26**
1.7.1 Methotrexate 26
1.7.2 Biologics 28
1.7.3 Methotrexate pharmacogenetics 28
1.8 Function annotation of GWAS of RA **30**
1.9 Study Rationale **34**
1.10 Aim and objectives **35**
CHAPTER 2: GENOME-WIDE ASSOCIATION STUDY IN BLACK SOUTH AFRICANS WITH RHEUMATOID ARTHRITIS *36*
2.1. Abstract **37**
2.2. Introduction **38**

2.3. Materials and methods 41
2.3.1. RA Cases 41
2.3.2. Population controls 41
2.3.3. Patient data collection and quality controls 43
2.3.4. Laboratory methods 43
2.3.5. SNP genotype data processing 45
2.3.6. Genotype imputation and post imputation quality control 46
2.3.7. Statistical analysis 47
2.3.8. GWAS catalog lookup 49
2.3.9. Post-GWAS *in silico* functional characterization 49
2.3.10. Power Calculation 50

2.4. Results 52
2.4.1. Patients demographics 52
2.4.2. Population structure 52
2.4.3. Sample and SNP QC 53
2.4.4. Genotype imputation 54
2.4.5. Genome-wide association testing 55
2.4.6. Assessment of GWAS associated SNP 58
2.4.7. Assessment of SNP at GWAS suggestive-level association 65
2.4.8. GWAS catalog lookup 68
2.4.9. *In silico* functional analysis of significant SNP 70

2.5. Discussion 74
2.5.1. Non-HLA novel association with RA susceptibility in BSA 74
2.5.2. MHC/HLA region 77
2.5.3. GWAS lookup of previous associations 79
2.5.4. Strengths and Limitations 80

CHAPTER 3: PHARMACOGENETICS OF METHOTREXATE RESPONSE IN BLACK SOUTH AFRICAN RHEUMATOID ARTHRITIS PATIENTS ON METHOTREXATE MONOTHERAPY 82

3.1. Abstract 83

3.2. Introduction 84

3.3. Materials and Methods 86
3.3.1. Study participants 86
3.3.2. Patient data collection and extraction 86
3.3.3. SNP selection 87
3.3.4. Genotyping and imputation 87
3.3.6. Statistical analysis 88
3.3.7. *In silico* function prediction annotation 89

3.4. Results 90
3.4.1. Demographic and clinical characteristics of study participants 90
3.4.2. SNP extraction in H3Africa array data 90
3.4.3. Genotype and allele frequencies 91
3.4.4. Exact replication of genetic association of 36 index SNP association with MTX response 97
3.4.5. Assessment of regional association with MTX response 99
3.4.6. Functional annotation predictions of significant SNP 102

3.5. Discussion 104

CHAPTER 4: OVERALL DISCUSSION AND CONCLUSION 108

4.1. GWAS study 109

4.2. Pharmacogenetics 111

4.3. Future studies 113

REFERENCES 115

Reference list: 116

APPENDICES: 148

APPENDIX A: Approval of title and Ethics Certificate 148

APPENDIX B: The 2010 ACR/EULAR classification criteria for rheumatoid arthritis 150

APPENDIX C: RegulomeDB variant classification scheme. 151

APPENDIX D: GWAS SNPs that reached genome-wide statistical significance of association with increased risk of RA in the chromosome 6 region with pre-imputation and imputed data. Data stored here.

APPENDIX E: GWAS non-MHC SNPs that reached genome-wide statistical suggestive level of association with increased risk of RA with imputed data. Data stored here

APPENDIX F: The EULAR response criteria after 6 months on methotrexate monotherapy 152

APPENDIX G: The genotype and allele frequency of the SNP selected for replication in BSA RA patients on Methotrexate for at least 6 months.. 153

APPENDIX H: Plagiarism report 161

LIST OF FIGURES

Figure 1.1: An illustration of the pathophysiology of rheumatoid arthritis (Firestein et al. 2017). 8

Figure 1.2: Illustration of the positions of amino acid polymorphisms associated with increased risk of RA in the classical HLA-DRB1 locus in ACPA+ patients across studies populations (Okada et al. 2019.) 19

Figure 1.3: Non-HLA loci associated with susceptible to RA in Europeans, Asian and transethnic populations identified in GWAS Meta-analysis studies (Okada et al. 2019). 21

Figure 1.4: The number of individuals and population ancestry included in GWAS over the years (Martin et al. 2019). 22

Figure 1.5: Illustration of a potential MTX molecular mechanism of action that leads to an anti-inflammatory effect in RA patients (Pastore et al. 2015). 27

Figure 2.1: A flow diagram showing the selection process for the population control individuals in the study from Soweto and Dikgale (AWI-Gen study). 42

Figure 2.2: A Quanto based a priori estimate of statistical power to detect an association with common SNPS. 51

Figure 2.3: The principal components analysis (PCA) between cases (Soweto cases) and population controls (Soweto and Dikgale). 53

Figure 2.4: During the QC process, individuals and SNPS were removed after performing all quality control steps represented in a bar graph. 54

Figure 2.5: GWAS Quantile-quantile plots comparing the distribution of observed P-values to the expected -log10 P-values assuming a null hypothesis of no association. 56

Figure 2.6: GWAS Manhattan plots showing the strength of association for each SNPS represented as the -log10 of the P-value 58

Figure 2.7: Cluster plots examples of the four top GWAS SNPS on chromosome 6 HLA region 61

Figure 2.8: Regional plot of the chromosome 6 MHC/HLA region SNP strongly associated with risk for developing RA, using the pre-imputation data. 63

Figure 2.9: Cluster plots of the two GWAS SNP in CPT1A (rs3019593; rs3019596). SNP genotype clusters show clear separation from each other. 64

Figure 2.10: Regional/LocusZoom plot of the chromosome 11 region with SNP that reached the genome-wide significance of association with increased risk of RA (using the imputed data). 65

Figure 2.11: Regional plots of the SNP that reached the GWAS suggestive threshold of P ≤ 1.0 X 10-05 for increased risk for RA, using the imputed genotyping data. 67

Figure 2.12: Chromatin interaction and eQTL status of the RA susceptibility locus, CPT1A in FUMA. 73

Figure 3.1: The genotype and allele frequencies comparisons of a SNP in the ABCB1 gene, rs1045642, previously associated with Methotrexate response. 93

Figure 3.2: The genotype and allele frequencies comparisons of a SNP in the ADORA3 gene, rs3394, previously associated with Methotrexate response. 94

Figure 3.3: The genotype and allele frequencies comparisons of a SNP in the MTHFD1 gene, rs2236225, previously associated with Methotrexate response. 95

Figure 3.4: The genotype and allele frequencies comparisons of a SNP in the *SLC19A1* gene, rs1051266, previously associated with Methotrexate response. 96

Figure 3.5: The genotype and allele frequencies comparisons of *ATIC* SNP (rs7563206) gene previously associated with Methotrexate response.....97

Figure 3.6: Regional plot on chromosome 9 showing the significantly associated imputed SNP and their positions relative to the genes in BSA compared to in European ancestry. 101

LIST OF TABLES

TABLE 1.1: METHOTREXATE RESPONSE PHARMACOGENETICS META-ANALYSIS STUDIES CONDUCTED IN PATIENTS OF EURASIAN ANCESTRY…..……..……..29

TABLE 2.1: DEMOGRAPHIC AND CLINICAL CHARACTERISTIC OF RA CASES AND CONTROLS 52

TABLE 2.2: TOP 20 SNP ON CHROMOSOME 6 IN THE HLA REGION THAT REACHED GENOME-WIDE SIGNIFICANCE LEVEL ($P \leq 5 \times 10^{-08}$) ASSOCIATION (SORTED BY P-VALUE).....60

TABLE 2.3: NON-HLA SNP THAT REACHED GENOME-WIDE SIGNIFICANCE AT $P \leq 5 \times 10\text{-}08$. THE SNP MAPPED TO THE INTRONIC REGION OF CPT1A. THE SNP GENOTYPED ON THE H3AFRICA ARRAY ARE PRESENTED IN BOLD 64

TABLE 2.4: GWAS CATALOGUE LOOKUP AND REPLICATION ANALYSIS OF RA-ASSOCIATED RISK SNP 69

TABLE 2.5: THE TOP 35 HLA SNP WITH HIGH PROBABILITY PREDICTION TO INFLUENCE GENE FUNCTION 71

TABLE 2.6: FUNCTIONAL PREDICTION AND ANNOTATION OF GENOME-WIDE SIGNIFICANTLY ASSOCIATED SNP IN THE CPT1A GENE. SNP IN BOLD WERE GENOTYPED ON THE H3AFRICA ARRAY AND THE OTHER TWO SNP WERE IMPUTED 72

TABLE 3.1: DEMOGRAPHIC AND CLINICAL CHARACTERISTICS OF THE STUDY PARTICIPANTS.....90

TABLE 3. 2: SUMMARY SHOWING ALLELE AND GENOTYPE FREQUENCIES BETWEEN PAIR GROUPS OF GOOD RESPONDERS (GR), MODERATE RESPONDERS (MR), NONRESPONDERS (NR) AND RESPONDERS (R) 92

TABLE 3.3: EXACT REPLICATION OF 36 PREVIOUSLY ASSOCIATED SNP WITH METHOTREXATE RESPONSE 98

TABLE 3.4: SNP SIGNIFICANTLY ASSOCIATED WITH METHOTREXATE RESPONSE IN THE REGIONAL ANALYSIS. MULTIPLE TESTING WAS CORRECTED FOR USING 1000 PERMUTATIONS (P ADJ) ..100

TABLE 3.5: FUNCTIONAL ANNOTATION PREDICTION OF SNP SIGNIFICANTLY ASSOCIATED WITH METHOTREXATE RESPONSE (SORTED BY BASE POSITION).....103

ABBREVIATIONS

AA	:	African American
ACCP	:	Anti-Cyclic Citrullinated Peptide
ACPA	:	Anti-Citrullinated Protein Autoantibody
ACPA-	:	Anti-Citrullinated Protein Autoantibody negative
ACPA+	:	Anti-Citrullinated Protein Autoantibody positive
ACR	:	American College of Rheumatology
AE	:	Adverse Effect
AFR	:	African
APCs	:	Antigen-Presenting Cells
As	:	Arsenic
ASN	:	Asian
BSA	:	Black South African
CADD	:	Combined Annotation Dependent Depletion
Cd	:	Cadmium
CI	:	Confidence Interval
CLEAR	:	Consortium for the Longitudinal Evaluation of African Americans with Early Rheumatoid Arthritis
Co	:	Cobalt
Cr	:	Chromium
CRE	:	Cis-Regulatory Elements
CRP	:	C-Reactive Protein
Cu	:	Copper
DC	:	Dendritic cells
DMARD	:	Disease-Modifying Anti-Rheumatic Drugs
DNA	:	Deoxyribonucleic Acid
EBV	:	Epstein-Barr virus
ENCODE	:	Encyclopaedia of DNA elements
eQTL	:	Expression Quantitative Loci
ESR	:	Erythrocyte Sedimentation Rate
EULAR	:	European League Against Rheumatism
EUR	:	European
FLS	:	fibroblast-like synoviocytes
GEO	:	Gene Expression Omnibus
GTEx	:	Genotype-Tissue Expression
GWAS	:	Genome-wide association studies
H3K4me1_Enh	:	Monomethylation of histone h3 lysine 4 of active enhancer
Hg	:	mercury
HLA	:	Human Leucocyte antigen
IDB	:	Identity by Descent
IL	:	Interleukin
JIA	:	Juvenile idiopathic arthritis

KGP	:	1000 Genomes Project
LD	:	Linkage disequilibrium
MAF	:	Minor Allele Frequency
MCP	:	Metacarpophalangeal
METEOR	:	Measurement of efficacy of Treatment in the Era of Outcome in Rheumatology
MHC	:	Major Histocompatibility Complex
MLS	:	Macrophages-Like Synoviocytes
MMPs	:	Matrix Metalloproteinases
Mn	:	Manganese
MTP	:	Metatarsophalangeal
MTX	:	Methotrexate
MTX-GLU	:	MTX polyglutamation
Mφ	:	Macrophage
NSAID	:	Non-Steroidal Anti-Inflammatory Drugs
OR	:	Odd Ratio
Pb	:	Lead
PC	:	Principal Components
PCA	:	Principal Component Analysis
PIP	:	Proximal Interphalangeal
PolyPhen	:	Polymorphism Phenotyping
Q-Q	:	Quantile-Quantile
QC	:	Quality Control
RA	:	Rheumatoid arthritis
RF	:	Rheumatoid factor
ROS	:	Reactive Oxygen Species
SE	:	Shared Epitope
SIFT	:	Sorting Tolerant from Intolerant
SNP	:	Single Nucleotide polymorphism
TCR	:	T-cell receptor
TNF	:	Tumor necrosis factor
VEP	:	Variant Effect Predictor
WHO	:	World Health Organisation

CHAPTER 1: LITERATURE REVIEW AND INTRODUCTION

1.1 Rheumatoid arthritis

Rheumatoid arthritis (RA) is a chronic autoimmune disease characterised by systemic chronic inflammation and destruction of the synovial lined joints. It is the most common among inflammatory arthritis such as systemic Lupus Erythematosus, juvenile idiopathic arthritis, and primarily affects the small joints of the hands, with symmetrical involvement of the peripheral joints such as the hands and feet (Cojocaru et al. 2010). Left untreated or insufficiently treated, RA can lead to accumulated joint damage, structural bone damage, excruciating pain, and significant disability. This could result in the health of RA patients being mostly compromised, and quality of life significantly reduced as patients often require assistance with basic activities.

The systemic nature of the disease means that the inability to sufficiently treat RA may lead to extraarticular organ manifestations like the skin, eyes, lungs, heart, renal, nervous and gastrointestinal being also affected (Crostein 2007). Patients with these extra-articular manifestations, which lead to diseases such as cardiovascular, respiratory, musculoskeletal and digestive diseases, often have higher mortality compared to the general population (Cojocaru et al. 2010; Turesson et al. 2004; Turesson et al. 2002). Rheumatoid arthritis patients have a lower life expectancy compared to the general population (Tobon et al. 2010; Gonzalez et al. 2007; Bongartz et al. 2007; Gabriel et al. 2003). In a prospective study, RA patients were reported to have a mortality rate of 54% higher (up to 10 years) than the general population (van den Hoek et al. 2017).

Economically, patients with RA are less productive and depend on government grants for survival (Pollard et al. 2005). Available data show that disability remains high among individuals with RA. Most of the countries affected are low-income countries such as those in Africa, where RA patients continue to work with high levels of disease activity, thus increasing the levels of disability (Sokka et al. 2010). Rheumatoid arthritis was ranked as the 42nd highest contributor to global disability in a list of 291 conditions studied (Cross et al. 2014).

1.2 Classification and Diagnosis of RA

1.2.1 RA history

Although it was initially believed that RA is a modern disease, evidence from ancient medical writing and paleopathological findings on skeletal remains have suggested that RA may be an ancient disease. The clinical features chronicle indicative of RA-like condition was first described in the 1800s by Dr. Augustin Jacob Landre-Beauvais, a Saltpetriete asylum resident physician in France. Dr. Landre-Beauvais observed that unlike gout, which was predominant in males, this condition affected women more than men. He, therefore, postulated that disease was characteristically new and called it "*la goutte asthenique primitive*" or "Primary Asthenic Gout" (reviewed in Entezami et al 2011).

Dr. Garrod, an English physician, improved on the characterization of the disease in the late 1850s by chemically testing and reported excess uric acid exclusively in the blood of Gout patients and not in patients with other forms of arthritis. He further reported a distinct profile of what was then known as "primary Asthenic Gout", to other forms of arthritis, which lead to him naming the condition as Rheumatic Gout (Garrod, 1859). It was not until towards the end of the 19th century that Dr. Garrod, coined the name "Rheumatoid arthritis", after studying worldwide ancient skeletal sample findings and observed skeletal damage. Furthermore, he divided RA into three categories of acute, chronic and irregular (Zyrianova 2012), suggesting that the disease is heterogeneous in nature. Paleopathological studies also confirmed RA as a distinct disease with chronic behavior (Joshi 2012; Entezami et al. 2011). However, although studies of skeletal remains to establish the age and origin of RA were being performed, challenges still remained of distinguishing RA from other rheumatic diseases.

1.2.2 RA classification criteria

The complex and heterogeneous nature of RA in conjunction with a poorly understood aetiopathogenesis and symptom similarities with other inflammatory rheumatic diseases made the diagnosis of RA very difficult (Harris 1990). In an attempt to distinguish and define RA from other rheumatic disease for clinical studies, several adaptive classification criteria have been developed. These include those from the American Rheumatism Association (1956 to 1961), the Rome classification criteria adaptations (1966 – 1987), the American College of Rheumatology (ACR) 1987 (reviewed in MacGregor 1995), to the most recent ACR/European League Against Rheumatism (Aletaha et al. 2010) were developed.

1.2.2.1 The American College of Rheumatology

In 1987, the American College of Rheumatology (ACR) used hospitalised patients to develop a set of classification criteria for RA (Arnett et al. 1988). The classification criteria focused on distinguishing patients with established RA from other types of rheumatic diseases using a set of seven clinical and laboratory parameters. The criteria included that patients should have had the following symptoms persistent for six weeks, which include physical assessment of morning stiffness in and around the joints an hour before maximal improvement, simultaneous swelling of at least three joints, synovitis of at least one wrist area and symmetric involvement of the same joint area on both sides of the body. In addition, patients should present with rheumatoid nodules and raised rheumatoid factor (RF) antibodies, which recognise the Fc subunit of Immunoglobulin G. The presence of any four of the seven symptoms was accepted as a classification of RA. The classification criteria successfully distinguished RA from other inflammatory diseases with approximately 94% sensitivity and 89% specificity. However, it focused more on classifying RA based on late-stage features (Aletaha et al. 2010). It was observed that joint damage and nodules accumulate over time and are rarely present in the early stages of disease (Machold et al. 2002; Van der Heijde 1995). In addition, these criteria did not identify individuals with early arthritis who went on to develop RA (Morvan et al. 2009). This necessitated the classification criteria review to identify patients early and prevent individuals from reaching the persistent and erosive disease state (Finckh et al. 2006).

1.2.2.2 ACR/European League Against Rheumatism

To achieve this, the 2010 ACR and European League Against Rheumatism (EULAR) joint working group developed an updated classification criteria aimed at identifying early RA individuals for research and therapeutic intervention with disease-modifying anti-rheumatic drugs (DMARD) (Aletaha et al. 2010). The new criteria focused on early-stage features, which included that patients should have synovitis in at least 1 joint which cannot be explained by other conditions, instead of the previously proposed three joints (Appendix B). Furthermore, patients should achieve a weighted minimum total score of six from individuals scores derived from four domains, which include the number and sites of swollen or tender joints involved (score range 0 – 5). Joints that can be involved include large joints of the shoulders, elbows, hips, knees, ankles and small joints of the hands such as metacarpophalangeal (MCP) and proximal interphalangeal (PIP), thumb interphalangeal joints, as well as the wrist and foot

second to fifth metatarsophalangeal (MTP). Additionally, patients should present with elevated serologic autoantibody levels (score range 0 – 3). In addition to having elevated levels of RF autoantibodies previously used in the 1987 ACR, the 2010 ACR/EULAR required that patients present with elevated levels of anti-cyclic citrullinated protein autoantibody (ACPA), weighted based on antibody levels, which further classify RA as either seropositive if above or seronegative disease if below a certain threshold. Patients should also present with elevated inflammation markers (score range 0 – 1) such as the erythrocyte sedimentation rate or C-reactive protein (CRP). Unlike the 1987 ACR classification criteria, the 2010 ACR/EULAR classification criteria also factor in the duration of symptoms persistence >6 weeks (score range 0 – 1 in two levels). Because of its focus on early disease classification, the 2010 ACR/EULAR classification criteria do not incorporate the assessment of radiographic changes into the final score (Aletaha et al. 2010; Neogi et al. 2010). However, this gap was reviewed and radiographic changes such as bony erosions or periarticular decalcification were added (Kay and Upchurch 2012).

1.3 Epidemiology of RA

1.3.1 Worldwide

Rheumatoid arthritis occurs with a global prevalence estimation of 1% and an annual incidence of around 0.4/1000 reported for an adult Caucasian population with RA (Viatte et al. 2013; Silman and Pearson 2002). Furthermore, RA affects females twice as often as males, peaking in the 4th – 5th decade of life (Silman and Hotchberg 2001). Low global prevalence estimates of 0.24% have also been reported (Cross et al. 2014). However, prevalence based on geographical location and socioeconomic status has also been observed. Specifically, a higher prevalence of 6% has been reported in North American Native Americans and a low prevalence (0.2-0.3%) observed within Japanese and Chinese populations (Wang et al. 2016; Silman and Pearson 2002). Furthermore, a systematic meta-analysis of regional RA rates for six low to middle income countries revealed low prevalence estimates of 0.37% (95% CI:0.23-0.51) for Eastern Mediterranean regions, 0.40% (95% CI:0.23-0.57% for Southeast Asian regions, 0.42% (95%CI:0.30-0.53%) for Western Pacific regions, modest meta-estimates 0.62% (95% CI:0.47-0.77%) for European regions and 1.25% (95% CI:0.64-1.86%) for American regions (Rudan et al. 2015). The study also reported observing no significant difference in RA prevalence in a rural and urban setting (Rudan et al. 2015). Although the female to male ratio is similar, high-income countries

experience low RA prevalence of approximately 0.3% when adjusted for age (WHO 2008), which is also what is observed for low- and middle-income countries (Brennan-Olsen et al. 2017; Cross et al. 2014). Paradoxically, however, in high-income countries, RA patients experienced a more severe outcome of the disease (Hifinger et al. 2015). These geographical and population variation in prevalence may suggest a role for environmental and genetic factors. Data has also shown that disease prevalence may be declining in other Caucasian populations in Western countries, where data collected over a decade indicated that RA prevalence has dropped from 2.1 million in 1995 to 1.29 million affected adult Americans in 2005 (Helmick et al. 2008). Globally, however, approximately 18 million, consisting of 3.16 million male and 14.87 million females individuals in middle and income countries suffered from RA in 2010 (Rudan et al. 2015).

1.3.2 Africa

Due to many reasons, which include lack of robust epidemiological studies of African ancestry populations, several clinical RA studies have reported that the disease was generally rare, with an earlier age of onset and more seronegative (Elshafie et al. 2016; Malemba et al. 2012; Malemba and Mbuyi-Muamba 2008; Kalla and Tikly 2003; Mody 1995; Brighton et al. 1988; Meyers et al. 1977; Beighton et al. 1975; Solomon et al. 1975). From the first two cases of RA described in the 1950s, several millions of RA cases have since been reported four decades later (Dowman et al. 2012; Usenbo et al. 2015). Current studies indicate that the prevalence is increasing with urbanization, similar to that observed in Europeans (Mody 2017; Kalla and Tikly 2003; Solomons et al. 1975). For example, a prevalence of between 0.6 and 0.9% has been reported from 5000 predominantly black adults in the Democratic Republic of Congo, Kinshasa (Malemba et al. 2012). Determining the prevalence of RA in Africa has been challenging mainly due to the difference in diagnostic and classification methods used. However, notwithstanding the challenges alluded to disease classification and based on a systematic analysis of 10 population-based studies published between 1968 and 2009, the prevalence of RA was estimated to be approximately 0.36%, which translated to 2.3 million affected individuals across Africa (Dowman et al. 2012).

In addition, RA appears to have great geographical and population variations (Tishkoff and Williams 2002), and population-based studies suggest an urban-rural prevalence gradient. For example, population-based studies identified no cases of RA in 154 Botswana rural

participants (Truswell et al. 1968), in 543 South African rural participants >18 years (Brighton et al. 1988) and in 1994 Nigerian participants. Moreover, a low prevalence of 0.3% was observed in 1070 participants >15 years from rural Lesotho (Moolenburgh et al. 1984) and 0.3% in 5120 rural Egyptians (Abdel-Naser et al. 1997). A recent rural-urban comparison study of 192 Senegalese participants reported that 21% and 79% of RA patients were from a rural and ab urban setting (Lekpa et al. 2012). Furthermore, a systematic review and meta-analysis of 12 African population cross-sectional studies reported high estimates in urban settings, which ranged from between 0.1% in Algeria, 0.6% in DRC and 2.5% in South Africa (Usenbo et al. 2015). In contrast, low estimates were reported for South African rural (0.07%), Egyptian (0.3) and Lesotho rural (0.4) population (Usenbo et al. 2015). Similarly, a retrospective hospital-based study in urban Nigeria identified 128 RA cases (Ohagwu et al. 2017), suggesting an increasing trend in areas previously free from disease. The observed increasing trend and pattern of RA can be attributed to industrialization or urbanization of low-income countries and improved diagnostic or classification criteria which identifies diseases early (Mody 2017). However, a low prevalence (0.13%) has also been reported in 249 Urban participants of Algerian ancestry (Slimani et al. 2014), which may suggest that RA is generally rare in that population.

1.4 Pathophysiology

The mechanisms that lead to RA expression are very complex and not yet completely understood. However, it is generally accepted that RA is defined by a complex autoimmune process that undergoes four clinical phases, which include pre-clinical, disease onset and inflammation phases, as well as joint damage and extra-articular manifestations (Figure 1.1).

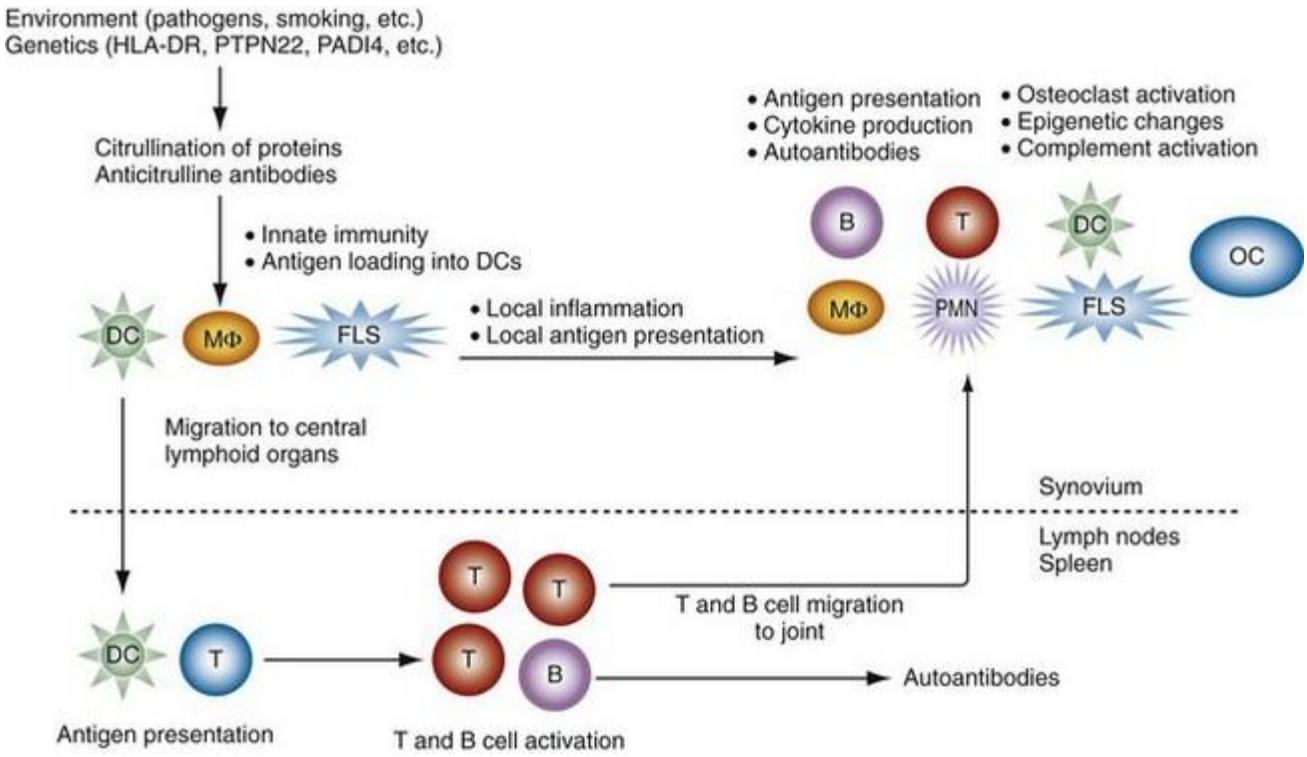


Figure 1.1: An illustration of the pathophysiology of rheumatoid arthritis showing the role and interaction of environmental factors, genetic factors, immune cells (DC-Dendritic cells, Mφ - Macrophage, FLS – fibroblast-like synoviocytes cells) and autoantibodies in the four phases of disease (Firestein et al. 2017).

1.4.1 Pre-clinical phase

The pre-clinical phase of RA is characterised by the production of autoantibodies RF and/or ACPA, and chronic inflammatory process is characterised by the production of cytokines. It is for this reason that Stanich and colleagues (2009) characterised RA as a syndrome. The disease generally progresses from joint synovitis, swelling and damage to systemically affecting extra-articular organs. It has been shown that RF and/or ACPA are present in an individual in the preclinical phase a decade before development of RA (Smolen et al. 2018; van der Woude et al. 2010; Isaacs 2010). Figure 1.1 shows an illustration of how genetic and environmental factors can interact to trigger RA onset. In a genetically predisposed individual (*HLA-DRB1* shared epitope (SE) allele), an interaction between genetic and environmental risk factors may generate neoantigens by citrullination of arginine to citrulline residues (Kidd et al. 2008). This is shown by a sharp increase in ACPA levels, which suggests a break in immune system tolerance to a citrullinated neoantigen (Balsa et al. 2010). Several processes, such as smoking or silica exposure, could lead to increased citrullination via the increased synthesis and activity of peptidyl arginine deiminase (PADI) enzymes in the lungs, (Bongartz

et al. 2007; Klareskog et al. 2006). The neoantigens bind to *HLA-DRB1,* more specifically to the antigen-presenting groove, and trigger an immune response by activation of antigen-presenting cells (APCs). These cells increase phagocytosis, and process and transports the arthritogenic peptides to the lymph nodes or thymus. Once in the thymus, the APCs display the peptides on their surface and activate CD4+ T helper cells to interact with B-cells which produce autoantibodies to RF and ACPA (Muller et al. 2015; Holers 2013). The post-translation modification alters the binding affinity of the neoantigen to the disease-predisposing HLA-complex, thus binding with higher affinity than the unmodified proteins. This primes the immune system, to which the altered arthritogenic peptides are not exposed to the T-cell receptor (TCR) during development, and thus not deleted in the thymus and therefore escape tolerance induction (Raposo et al. 2018).

1.4.2 Disease onset and inflammation

The activated CD4+ T helper cells stimulate B-cells to proliferate and differentiate into plasma cells and produce antibodies (Kim and Berek 2000). The CD4+ T helper cells and ACPA enter the blood circulation and migrate to the joint synovium. Once in the synovium, CD4+ T-cells secrete cytokines IFN-Y and IL17 and also interact with B-cells to produce antibodies. These CD4+ T-cells secrete cytokines, recruit macrophages-like synoviocytes (MLS), which also secrete TNF alpha, IL1 and IL6, and others. These pro-inflammatory cytokines, together with the T-cell cytokines stimulate the fibroblast-like synoviocytes (FLS) to proliferate and secrete matrix metalloproteinases (MMPs) and other small-molecule mediators such as prostaglandins (Bartok and Firestein 2010). The overproduction and overexpression of pro-inflammatory cytokines facilitates interaction between T- and activated B-cells, which increases the production of RF and ACPA among others and activates synovial cells (FLS and MLS) to proliferate.

1.4.3 Joint damage

The interaction results in overproduction of pro-inflammatory cytokines such as IL6, implicated in persistent inflammation and joint damage (Choy et al. 2002). Over the years, ACPA titers increase sharply (Sokolove et al. 2012). The increase in FLS and immune cell hyperplasia causes inflammation. The FLS, macrophages and mast cells along with pro-inflammatory cytokines such as TNF, IL1, and IL17 induce the secretion of MMPs, which break down the proteins and promote cartilage destruction (Konttinen et al. 1998). Similar

findings were observed in TNF-driven arthritis mouse studies (Hayer et al. 2016). However, studies have shown that the FLS is overrepresented and the most destructive cell type for cartilage and produces high levels of collagenases and stromelysins (Keyszer et al. 1998). In addition, increased FLS expression is observed at the direct site of cartilage invasion in the pannus, suggesting FLS migration potential (Keyszer et al. 1998). Indeed, one study reported observing aggressive invasion into cartilage explants in immunocompromised mouse models compared to controls (Muller-Ladner et al. 1996). Bone erosion is generally caused by immune cell infiltration and production of inflammatory cytokines which stimulate T-cells surface-bound receptor activator of nuclear factor kappa-B ligand (RANKL) protein, which binds to receptor activator of nuclear factor kappa-B (RANK) on the surface of osteoclasts (Sobacchi et a. 2019; Axmann et al. 2009). Osteoclasts can degrade the bone matrix by producing protease such as cathepsin K (Skoumal et a. 2005). The antibodies in the synovium bind to their targets and form immune complexes and activate the complement system, which is a family of 9 small proteins. These proteins function in an enzymatic cascade to cause joint inflammation and injury.

1.4.4 Extra-articular manifestation

The chronic inflammation, cytokines, and growth factors released by the infiltrated immune cells, together with hypoxia resulting from synovial hyperplasia cause angiogenesis, thus enabling infiltration of the joint by more inflammatory cells (Elshabrawy et al. 2015; Honorati et al. 2006). The inflammatory cytokines in the synovium and FLS can migrate to other joints and distant tissue to cause polyarticular disease. For example, IL1 or 6 canmigrate to the brain and act as pyrogens which induces fever, as well as result in the formation of rheumatoid nodules in visceral organs, vasculitis in blood vessels, atheromatous and fibrofatty plaques in vessels, anemia and rheumatoid lung (Lefevre et al. 2015; Cojocaru et al. 2010; Lefevre et al. 2009).

1.5 Risk Factors

The exact cause of RA remains elusive. However, it is generally accepted that the disease manifests after an interaction between genetic and an environmental factor in genetically predisposed individuals. Several risk factors have since been associated with increased risk of developing RA, which includes environmental factors (i.e. smoking, infections, dietary,

silica exposure, Vitamin D deficiency, microbiome), being of the female gender and genetic factors.

1.5.1 Environmental factors

Environmental factors previously associated with increased risk of developing RA are described below and include smoking, heavy metals, diet, age, gender, microbiome, and infection.

1.5.1.1. Smoking

Smoking is the most studied and validated risk factor with overwhelming evidence demonstrating susceptibility to the development of RA. The interaction with genetic factors was demonstrated in a series of 13 monozygotic twin pairs discordant for RA and smoking, where the smokers developed the disease in 12 of the 13 pairs (Silman et al. 1996). Smoking was shown to increase the risk of developing RA by 26% with an odds ratio of 1.9, which doubles with an increase in the 20 pack-year history of tobacco used compared to non-smokers (Hedstrom et al. 2018; Di Giuseppe et al. 2014). A meta-analysis of observational studies reported that smoking could increase the risk of developing RA by 40% in seropositive RA patients and heavy smokers (Sugiyama et al. 2010) even among smokers with low lifelong exposure (Di Giuseppe et al. 2014). A similar dose-dependent risk was also observed in African Americans (AA), where the risk of developing RA was limited to patients with more than 10 pack-years of exposure and was increased in individuals positive for at least one copy of the *HLA-DRB1* SE alleles (Mikuls et al. 2010).

Smoking is associated with the production of elevated levels of pro-inflammatory cytokines and ACPA antibodies through citrullination of protein antigens (Klareskog et al. 2005; Mattey et al. 2002) and increased disease activity (Sokolove et al. 2012). Furthermore, smoking can negatively impact on disease outcome in seropositive RA and severity in patients with at least one copy of the *HLA-DRB1* SE (Kallberg et al. 2011; Saag et al. 1997). Moreover, studies have implicated smoking as an independent risk factor for developing extra-articular manifestation such as pericarditis, neuropathy, ophthalmological manifestations and glomerulonephritis in RA patients (Turesson et al. 2002). Although inconclusive, several mechanisms of action to explain the effects of smoking in susceptibility of RA have been suggested. These include (1) increased production of oxidative stress which increases rheumatoid inflammation due to an altered antioxidant system caused by presence of free

radicals (Kalpakcioglu and Senel 2008), (2) decreased apoptosis due to an impaired fas (CD95)-Fas ligand (CD178) (Peng 2006), (3) autoantibodies through citrullination of proteins (Scott et al. 2010) and (4) epigenetic changes such as DNA methylation (Zeilinger et al. 2013). However, conflicting results have also been reported, where tobacco smoking was associated with a reduction of radiographic progression in early RA (Vesperini et al. 2013). In addition, Naranjo et al. (2010) reported observing no clinical status difference associated with smoking in a multinational database. Smoking has also been associated with influencing treatment efficacy. RA smokers have a reduced responsiveness to treatment, with studies showing that higher doses of DMARD are required to achieve efficiency compared to non-smokers or ex-smokers (Saevarsdottir et al. 2011; Westhof et al. 2008).

1.5.1.1 Microbiota and infections

The altered composition or imbalance of the normal flora, also termed dysbiosis, has also been associated with an increased risk of developing RA. An increase of *Porphyromonas gingivalis* DNA, also implicated in the development of periodontal disease, has been observed in RA patients (Konig et al. 2016; Martinez-Martinez et al. 2009). Furthermore, immune response with increased titers of *P. gingivalis* antibodies have been identified in individuals at high risk of developing RA (Mikuls et al. 2012) and in RA patients (Hitchon et al. 2010; Mikuls et al. 2009). Animal mice model studies have also shown that immunization with *P. gingivalis* induced or facilitated the severe form of RA (Maresz et al. 2013; Kinloch et al. 2011; Bartold et al. 2010). Several arthritogenic molecular mechanisms of action associated with microorganisms, that may increase the risk of RA, have been suggested. These include citrullination of proteins by PADI enzyme which triggers ACPA production and molecular mimicry (Wegner et al. 2010; Foulquier et al. 2007), induction of inflammation and direct joint damage (Bartold et al. 2010; Cantley et al. 2011).

Another bacterium, *Proteus mirabilis,* mostly associated with urinary tract infection, has also been associated with RA and increased antibodies against it have been reported in RA patients (Puntis et al. 2013; Rashid et al. 2007; Newkirk et al. 2005; Tiwana et al. 1999). Similarly, increased Mycoplasma DNA, glycoglycerophospholipids and antibodies were all detected in RA patients (da Rocha et al. 2011; Kawahito et al. 2008). Furthermore, it was reported that immunization with *Mycoplasma arthritidis* induced or worsened arthritis in mice (Cole and Griffiths 1993). Gut microbiota studies have also shown that diversity is decreased

in RA patients compared to the general population, with a marked increase in rare l bacterial specie such as Actinobacteria in disease patients (Chen et al. 2016). Increased levels of *Prevottela copri* has also been found in treatment naïve early disease when compared to patients with established RA and non-RA individuals (Alpizar-Rodriguez et al. 2019; Pianta et al. 2014). These microorganisms have been shown to potentially increase the risk of developing RA by either neo-antigen generation or citrullinating host peptides and induce autoimmune responses (Alpizar-Rodriguez et al. 2019; Larsen 2017; Wegner et al. 2010). In addition, several studies have reported the association of Epstein-Barr virus (EBV) infection with RA and other autoimmune diseases (Balandraud and Roudier 2018; Trier et al. 2018). Furthermore, EBV DNA, proteins, and antibodies have all been detected in RA patients (Lunemann et al. 2008; Klatt et al. 2005). Similar to bacteria, viruses also increase the risk of developing RA by molecular mimicry and inducing a strong immune response (Fujiwa and Takei 2015; Klatt et al. 2005).

1.5.1.2 Heavy metals

The association between exposure to heavy metals such as cadmium (Cd), mercury (Hg), arsenic (As), copper (Cu), lead (Pb), cobalt (Co), manganese (Mn) and chromium (Cr) and development of RA has been previously described in earlier studies (Irfan et al. 2017; Pedersen and Permin 1988). The authors reported observing a high prevalence of RA in painters who were exposed to heavy metals. Although not much was known about the pathogenesis of RA at the time, another study suggested that heavy metals may contribute to the development of RA through the generation of reactive oxygen species (ROS) (Quinonez-Flores et al. 2016; Mateen et al. 2016). These ROS such as superoxide ions, hydrogen peroxide and nitrogen oxide induce cellular oxidative stress, which alters cellular proteins to generate neoantigens and cause the immune system to overreact to its own proteins (Salomon-Escoto et al. 2011). High serum levels of heavy metals such as Cd, Co, CR, Cu, and Pb have been observed in 78 RA patients when compared to 75 controls (Hashmi and Shah 2012). Similar findings have been reported recently when 100 RA patients were compared to 100 controls (Irfan et al. 2017). Cadmium is also one of the ingredients in cigarettes, which may in part at least, explain the citrullination of protein by Cd (Cates and Hutchinson 2014; Mohamed et al. 2012). High levels of Cd were also observed in the blood, urine, and scalp of smoking and non-smoking RA patients (Afridi et al. 2013; Afridi et al. 2012), which suggests that it may play an independent role. More recently, Joo et al. (2019)

showed that there is a linear relationship between RA prevalence and increase in serum level of Cd in 53,829 females of Asian ancestry. Exposure to heavy metals has also been associated with RA disease progression and extra-articular manifestations. For example, patients living in farm soils contaminated with Cu had increased white blood cell, inflammatory markers and disease activity scores compared to patients from low-copper contamination setting (Yang et al. 2016). Furthermore, extremely high levels of urinary Cd were observed in RA patients that developed rheumatoid pulmonary nodules (Joo et al. 2019; Murphy et al. 2019; Murphy et al. 2017).

1.5.1.3 Diet

Studies into the contribution of diet to RA pathogenesis remain inconclusive. However, diet appears to play a role in both the risk of developing the disease as well as in protection against RA. For example, several studies have reported that sugary drinks which contain high-fructose corn syrup, fruit drinks, and apple juice increased the risk for developing RA by threefold in 1,209 US adults between the ages of 20 – 30 years (DeChristopher et al. 2016). Other beverages that showed an increased risk of seropositive RA were caffeinated coffee and tea consumption in a Finland population (Heliovaara et al. 2000). Furthermore, the Iowa Women's Health Study confirmed that coffee consumption was associated with the risk of developing RA when evaluating 31,336 women between 55 and 69 years of age (Mikuls et al. 2002). However, no association was detected between RA and consumption of coffee or tea in younger women (Karlson et al. 2003). More recently, a large prospective study of 76,853 older women (between 50 – 79 years) reported detecting no association between coffee consumption and RA onset but a modest association between daily caffeinated, non-herbal tea consumption (Lamichhane et al. 2019). Long-term moderate consumption of alcohol was associated with a reduced risk of developing RA in 121,701 female participants (Lu et al. 2014). Similar observations were previously reported in a hospital-based study comparing 135 high alcohol intake participants to 378 controls (Hazes et al. 1990). However, no association was observed when 349 RA incident cases were compared with 1,457 controls (Voigt et al. 1994). Both cooked vegetables, olive and fish oil were independently associated with protection against the development of RA (Linos et al. 1999; Rosell et al. 2009), particularly Greek women (Shapiro et al. 1996). Although the protective mechanism was not clear, the antioxidant and high omega-3 fatty acid content of the olive oil has been suggested to play a role (Woodman et al. 2019).

1.5.1.4 Age and Gender

Studies have shown that women have an increased cumulative lifetime risk estimation of developing RA (3.6) than men (1.7%) (Ngo et al. 2014; Crowson et al. 2011). This has been confirmed by a systemic meta-analysis which revealed that women in six low- and middle-income countries had a fourfold increased chance of developing RA compared to men (Rudan et al. 2015). Although RA can occur at any age, the disease is more prevalent in individuals from their 4th decade of life and peak incidences have been observed at the 6th decade where the disease affects both genders equally (Myasoedova et al. 2010). One in 28 women and 1 in 59 men will develop the disease in their adult life (Lajas et al. 2003). However, intra-variations have been observed in women. Incidence rates of RA appear to be low in women on oral contraceptive (0.3 in 1,000 women-years) compared to women who have never taken oral contraceptives (0.65 in 1,000 women-years) or those that were previously on oral contraceptives (0.55 in 1,000 women-years) (Orellana et al. 2017; Silman and Pearson 2002). In addition, women of childbearing age but who have not given birth are at an increased risk of RA, although disease remission is observed in pregnant women (Ostensen and Villiger 2007). In addition, menopause appear to be associated with an increased risk of systemic autoimmunity, probably due to reduced ovarian function and hormonal imbalances (Aplizar-Rodrigues et al. 2017). Furthermore, both subfertility and the post-partum period after first pregnancy and breastfeeding appear to increase the risk of developing RA (Silman and Pearson 2002). In contrast to early disease onset observed in women, men have late disease onset and are mostly seropositive for RF and ACPA (Drosos et al. 1997).

1.5.2 Genetic factors

The involvement of genetic factors in the development, progression, and severity of RA is now well established. The contribution of a genetic component to a disease can be estimated in several ways: for example, by comparing concordance between monozygotic and dizygotic twin pairs and calculating the sibling's risk ratio (risk to sibling divided by population risk) (Risch 1990). This risk ratio in siblings was found to be between 2 - 15 for RA (Vyse et al. 1996). This suggests that genetic factors are important in the development of RA and account for disease clustering within families (Vyse et al. 1996). Indeed, overwhelming evidence has shown that RA runs in families (Jones et al. 1996) and the chances of developing RA when one has a family history increases fivefold (Frisell et al. 2013; Grant et al. 2001). Moreover,

siblings of RA probands have an increased risk, with relative risk estimated to be between 1.6-15.8 when compared to controls (Smolik et al. 2013; Seldin et al. 1999). Furthermore, mothers transfer susceptibility to RA to their offspring's more often than fathers (Koumantaki et al. 1997). Twin studies have revealed that monozygotic twins have a concordance of about 12% - 15% and dizygotic (fraternal) twins have 2%-5% compared to the general population at 1% (Silman et al. 1993; MacGregor et al. 2000). Furthermore, twin studies estimated an overall genetic contribution to RA susceptibility as being around 50-60% (MacGregor et al. 2000), which increases to 68% in ACPA+ and 66% in ACPA- patients (van der Woude et al. 2009). However, family history is not conclusive of risk of developing RA and the trait may skip a generation (Frisell et al. 2013), which suggests that exposure to currently unknown environmental factors also contributes to the development of RA. Rheumatoid arthritis is a complex multifactorial trait and therefore gene sequences cannot fully explain heritability, suggesting a role for epigenetic mechanisms and environmental factors (Kaminsky et al. 2009).

1.6 Genetic association studies of RA

Overwhelming evidence indicates that genetic variation at both Human Leucocyte Antigen (HLA) and non-HLA loci contribute to an increased risk of developing RA. The genetic contributions to RA have been identified through linkage and genetic association studies, using both candidate and genome-wide approaches. Ethnic and geographical heterogeneity have been shown to influence susceptibility and severity of RA, further supporting the role of genetic variation in RA.

1.6.1 HLA association with RA

The HLA complex, also referred to as the major histocompatibility complex (MHC) in humans, is located on chromosome 6 (6p21.3) and has consistently shown the strongest genetic association with susceptibility to RA across all studied populations. The HLA complex contains approximately 200 genes of which 40 encode the highly polymorphic cell surface molecules that play a role in mediating the cell's interaction with immune cells by presenting epitopes to T-cells (Forbes and Trowsdale 1999; Klein and Horejsi 1998; Bjorkman et al. 1987). The HLA genes are crucial for antigen processing and presentation and are divided into class I genes (*HLA-A, HLA-B,* and *HLA-C*) and class II genes (*HLA-DP, HLA-DM, HLA-DO, HLA-DR,* and *HLA-DR*) (Klein and Sato 2000). The HLA class II complex region accounts

for over 30% - 50% of RA heritability in all populations studied, including Africans (Terao et al. 2016; Yamamoto et al. 2015; Van der Woude et al. 2009; Deighton et al. 1989).

1.6.1.1 HLA-DRB1 Shared Epitope

Earlier studies used serological techniques to identify the association between the HLA locus (particularly *HLA-DR4,* later renamed *HLA-DRB1*) and increased risk of developing RA in Caucasians (Stastny 1978) and later in African ancestry individuald (Martel et al. 1989; Mody and Hammond 1994). This association was further refined using molecular typing techniques and described that the association was driven by expression of a conserved region $_{70}$QKRAA$_{74}$, $_{70}$QRRAA$_{74}$, $_{70}$RRRAA$_{74}$ in the third variable of *HLA-DRB1*, later termed the "shared epitope" (*HLA-DRB1* SE) (Gregersen et al. 1987). Expression of the SE alleles has been shown to confer dose-dependent risk for RA, where individuals having one and two copies of the SE allele were four and eight times more likely to develop RA. Furthermore, the presence of the SE allele often led to more severe disease when compared to individuals without the SE (Macgregor 1995; Weyand et al. 1992). Like susceptibility to RA, *HLA-DRB1* SE association with severity appears to be dose dependent. For example, several studies have suggested that RA patients homozygous for *HLA-DRB1**04:04 are more likely to develop rheumatoid nodules, extra-articular manifestations and are more likely to require joint surgery compared to heterozygous patients (MacGregor 1995; Weyand et al. 1992). However, patients that carried the *HLA-DRB1*04:01* alleles had a modest disease progression (Gough et al, 1994; Weyand et al. 1992), which may suggest that *HLA-DRB1*04:04* may play a critical role in RA disease severity and joint destruction. Other HLA-DRB1 SE alleles, *HLA-DRB1*04:05* and *HLA-DRB1*01:01,* have also been shown to increase the risk of RA severity and were found to be associated with the presence of destructive bony erosions, extra-articular manifestations and reduced likelihood of remission (Lin et al. 2007; Wakitani et al. 1997). In contrast, while most susceptible *HLA-DRB1 SE* alleles are also associated with severity, the $_{70}$DERAA$_{74}$ encoding *HLA-DRB1* alleles are associated with reduced risk of RA (van der Helm-van Mil et al. 2005), highlighting the potential protective effect of the SE locus.

1.6.1.2 Ethnic variation

Ethnic variation in the frequency of the *HLA-DRB1* SE allele has also been described. For example, although most African ancestry RA patients are seropositive, the frequency of *HLA-*

*DRB1*04* allele was found to be lower (23% of 66 cases) than in Caucasians (Meyer et al. 2011; Anaya et al. 2001; Reveille et al. 1996; McDaniel et al. 1995; Wordsworth et al. 1992). This finding, in conjunction with recent reports that showed that RA susceptible HLA haplotypes odds ratios are highly correlated between Eurasian and UK individuals of African ancestry, suggests that the *HLA-DRB1*04* association with RA in the African ancestry individuals may be through admixture with Caucasians (Traylor et al. 2017; Hughes et al. 2008). This finding is supported by the observation in sub-Saharan Black South African (BSA) RA patients, where about 90% of those with RA were found to carry at least carried at least one copy of the SE allele (Meyer at al. 2011). This population is relatively admixed mixed with European populations genetic variants (May et al. 2013), which may explain the association.

Fine mapping and conditional analysis of the *HLA-DRB1* SE association with RA using amino acid studies revealed several important findings. The *HLA-DRB1* SE individual amino acids association with risk of RA varied based on ethnicity. For example, although amino acid 11 and/or 13 conferred susceptibility to RA in European, Asian and African ancestry, amino acid 57 was observed only in African and Asian ancestry, 74 in Asian and European ancestry and 71 only in European ancestry (Govind et al. 2019; Danila et al. 2017; Traylor et al. 2017; Saxena et al. 2017; Reynolds et al. 2014; Okada et al. 2014; Ling et al. 2016; Raychaudhuri et al. 2012). Figure 1.2 represents an illustration of the *HLA-DRB1* SE and the amino acid positions conferring susceptibility across three populations stratified into ACPA+ and ACPA- groups, suggesting shared risk loci (Okada et al. 2019).

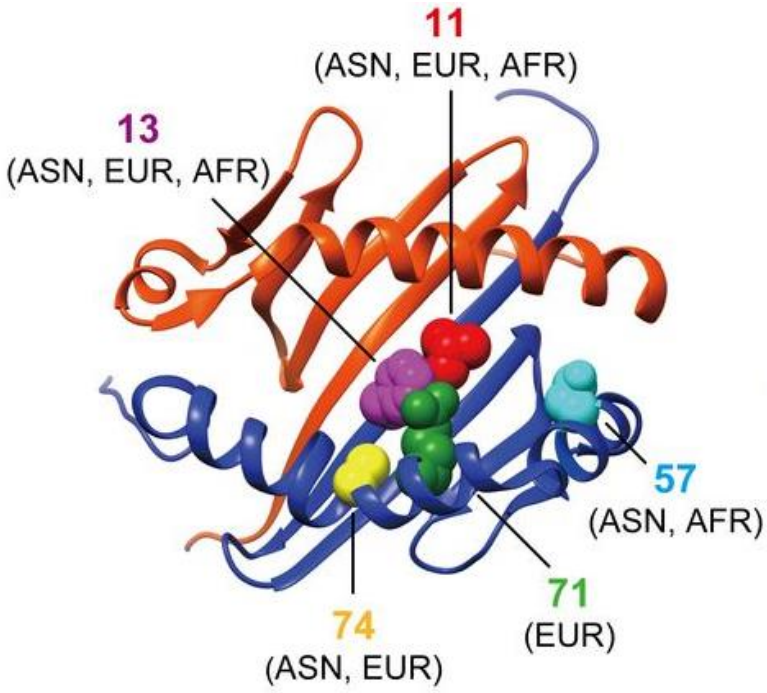


Figure 1.2: Illustration of the positions of amino acid polymorphisms associated with increased risk of RA in the classical HLA-DRB1 locus in ACPA+ patients across studies populations across studied populations. An overlap of risk loci is also demonstrated for the different amino acid positions, where there position 13 polymorphisms is prominent across the three studied populations (AFR – African ancestry, ASN – Asian ancestry, EUR – European) (Okada et al. 2019).

1.6.1.3 Non-HLA-DRB1 SE association with RA

Although the *HLA-DRB1* SE is clearly a very important region that confers the largest risk for RA susceptibility, it does not fully explain the observed genetic association. Furthermore, which independent HLA loci combinations that may better represent the estimated risk remains to be elucidated. This is compounded by the fact that the genes in the *HLA-DR* and *HLA-DQ* are in very strong linkage disequilibrium (LD) (Zanelli et al. 2000; Zaneli et al. 1995). Furthermore, current imputation reference panels contain limited HLA variants and genes and do not cover all the variation across different populations, further complicating the identification of independent signals. However, the presence of multiple independent associations within the HLA complex but outside the *HLA-DRB1 SE* have been previously suggested (Ding et al. 2009; Jawaheer et al. 2002). Independent associated to susceptibility to RA was observed with amino acid changes at position 9 in both *HLA-B* and *HLA-DPB1* (Raychaudhuri et al. 2012). A recent meta-analysis of eight studies with 592 cases and 935 controls confirmed the association of *HLA-DPB1*04:01* and *06:01 (Jiang et al. 2018). Other

non-*HLA-DRB1 SE* alleles associated with increased risk of developing RA in Eurasian populations include *HLA-DRB1*03:01* and *HLA-DRB1*07:01* (Vignal et al. 2009) and *HLA-DRB1*09:01* (Reynolds et al. 2014; Okada et al. 2014; Bang et al. 2013). Another study reported the association of asparagine amino acid polymorphism at position 77 in the *HLA-A* region with susceptibility to RA (Han et al. 2014). More recently, a coding variant in the *HLA-DOA* that alters the gene's expression levels in seropositive RA patients has been identified (Okada et al. 2016).

Cumulatively, the above findings confirm the overwhelming evidence that the HLA locus confers increased risk to the susceptibility of the development and severity of RA across studied populations. However, the findings also show that although it is a very important and a necessary region in the development of RA, the HLA locus on its own is not sufficient to cause disease (Newton et al. 2004). This Suggests a role for non-HLA genes in the pathogenies of RA.

1.6.2 Non-HLA genes association

The completion of the human genome sequence and 1000 Genomes Project (KGP) in conjunction with international collaborations and rapid advancements in array technology underpinning genome-wide association studies (GWAS), have facilitated the identification of relevant RA genetic risk loci (Okada et al. 2012; Stahl et al. 2010). The past two decades have seen the identification of over a 100 non-HLA SNPs associated with ACPA+ RA susceptibility mostly in European and Asian populations using candidate gene studies and GWAS (Okada et al. 2019). Figure 1.3 shows some of GWAS RA genetic risk loci validated in European and Asian ancestry (Okada et al. 2019). These findings have improved the general understanding of the contribution of the genetic architecture of RA pathogenesis. Furthermore, validation of many of these risk loci in transethnic populations suggests that some of the RA genetics risk loci are shared among populations of different ancestry. These genetic variants are captured in the curated GWAS catalog (https://www.ebi.ac.uk/gwas/), which is a publicly accessible summary of human genetic association research produced by the US National Human Genome Research Institute in partnership with European Bioinformatics Institute.

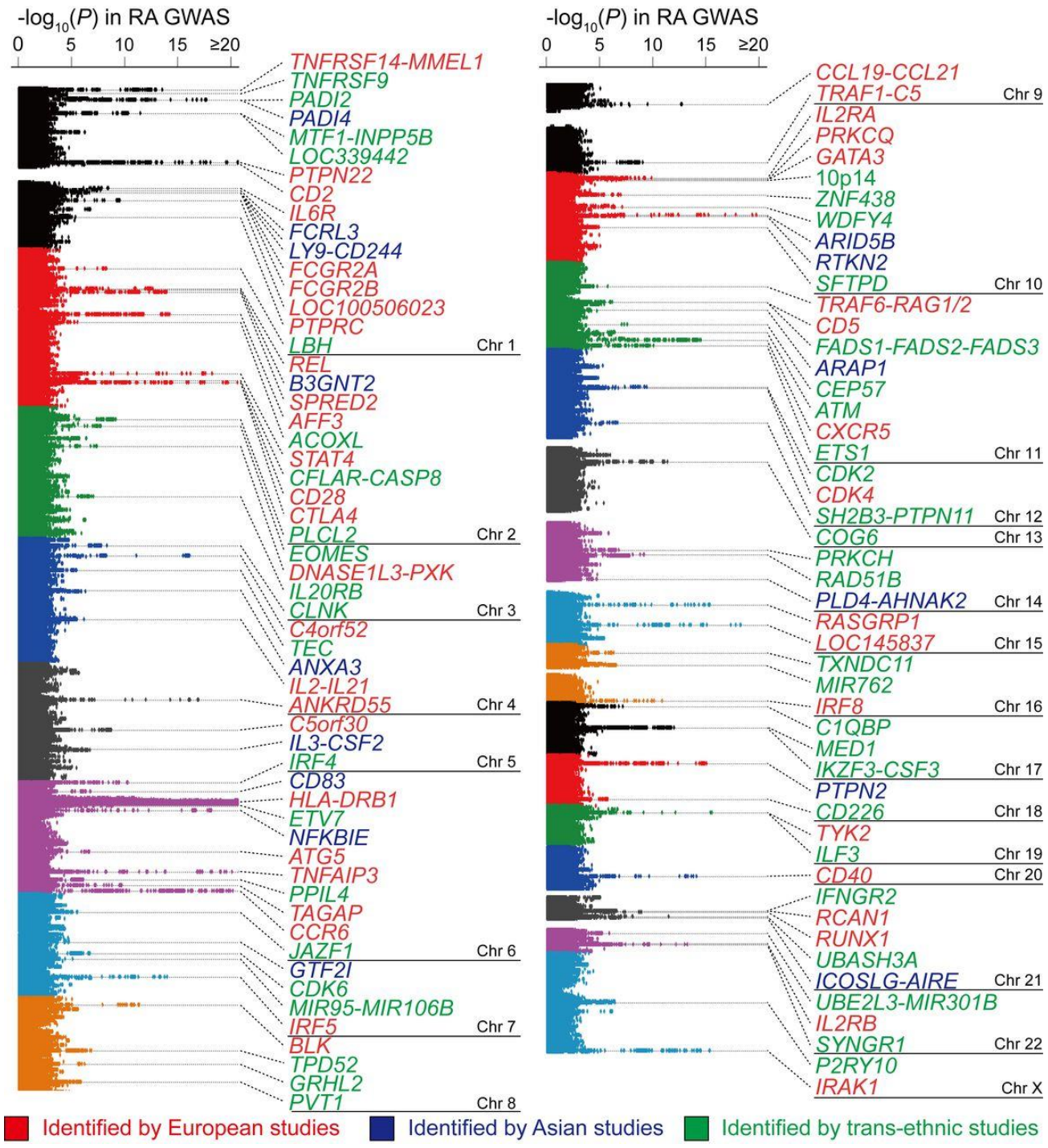


Figure 1.3: Non-HLA loci associated with susceptible to RA in Europeans, Asian and transethnic populations identified in GWAS Meta-analysis studies. Gene loci first identified in European, Asian and transethnic populations are coloured red, blue and green respectively. The statistical significance of the association in the studied population is also represented by the Manhattan plot (Okada et al. 2019).

Unfortunately, for a variety of reasons, the database is dominated by studies of European ancestry and to date, a very small proportion of participants have been of African ancestry (Figure 1.4) (Martin et al. 2019).

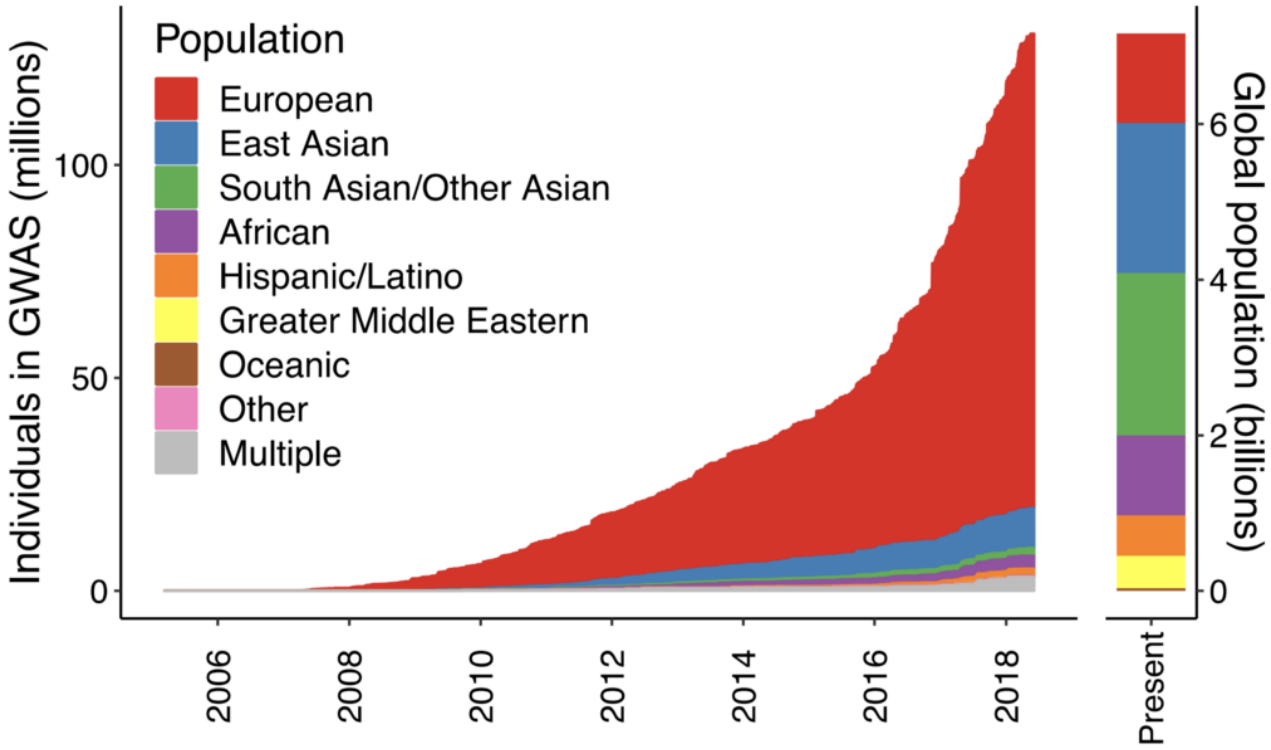


Figure 1.4: The number of individuals and population ancestry included in GWAS over the years. The population ancestries are represented by different colours. The majority of GWAS are still conducted in European (red), East Asian (blue) and South/other Asian ancestries (green). African ancestry individuals (purple) still make a very small proportion (Martin et al. 2019).

Specifically, of the 40 studies conducted, only three studies included African ancestry populations. Of the three, one included African Americans individuals (Laufer et al. 2019) and the other included individuals from North Africa (Saxena et al. 2017). Both populations have previously been shown to be genetically admixed with European ancestry and cannot be used as proxy for African Ancestry populations. This means that only one study, which included only 263 BSA RA patients (Govind et al. 2014), is captured in the GWAS catalog. Thus, attesting to the dearth of GWAS data in populations of African ancestry. Unlike the HLA locus, most non-HLA variants associated with risk of developing RA have not yet been replicated across all studied ethnic groups. GWAS conducted to date show that the non-HLA variants identified to date confer low to moderate risk and explain approximately 5% and 6%

of heritability in European and Asian ancestry (Okada et al. 2014). The validated RA risk loci have shown that there are ethnic-specific risk loci exist. For example, a SNP (rs2476601) in the coding region of protein tyrosine phosphatase non-receptor type 22 (*PTPN22)*, was first identified in a GWAS with 2,000 RA cases and 3,000 controls of European ancestry (WTCCC 2007). This loci, *PTPN22* (rs2476601-T), is the second strongest risk loci (OR = 1.9) after the HLA region in European ancestry. Since then this region has been replicated in many independent studies of European ancestry (Plenge et al. 2007; van Oene et al. 2005; Simkins et al. 2005; Begovich et al. 2004).

1.6.2.1 PTPN22

However, the *PTPN22* risk loci did not replicate in a GWAS of RA with 7,069 cases and 20,727 controls of Asian ancestry (Kochi et al. 2010; Lins et al. 2010), even after analysis of high throughput sequencing data (Lee et al. 2009). Similar results were observed in a sub-Saharan African population of 263 BSA RA cases and 374 controls using the Immunochip (Govind et al. 2014). This suggests that the risk loci may be European ancestry specific. Since then, several other SNPs in or near genes such as *TNFAIP3* (rs10499194)*, OLIG3* (rs6920220)*, CD40* (rs4810845)*, REL* (rs13031237)*, CTLA4* (rs231735) and *BLK* (rs2736340) with moderate effect sizes (~1.3) have been identified and validated in European ancestry and summarised in Figure 1.3 (Okada et al. 2019, Stahl et al. 2010; Gregersen et al. 2009; Thomson et al. 2007; Plenge et al. 2007). To date, studies in patients of European ancestry have confirmed over 40 non-HLA genetic variants relevant to RA pathogenesis.

1.6.2.2 PADI4

In contrast, several studies have shown that a SNP in intron 3 of the peptidyl arginase deiminase 4 (*PADI4*) locus (rs2240340-T) is the second strongest associated risk loci (OR = 1.9) in individuals of Asian ancestry (Kochi et al. 2009). The *PADI4* variant has been replicated in many GWAS in patients of Asian ancestry (Yang et al. 2015; Hou et al. 2013; Suzuki et at. 2013) but has not been confirmed in European ancestry, suggesting that it is Asian ancestry specific. The difference in RA risk loci may highlight the inherent ethnic-specific heterogeneity and the possible influence of environmental factors.

1.6.2.3 Other Non-HLA genes

In Asian ancestry populations, a regulatory risk variant in *CCR6* (rs3093024) was identified in 7,069 cases and 20,727 controls and replicated in two independent case-control cohorts which included 3,662 cases and 15,873 in cohort 1 and 1,106 cases and 1486 controls in cohort 2 (Kochi et al. 2010). Other variants with modest effects have since been identified and validated in individuals of Asian ancestry and include SNPs in or near *AIRE* (rs2075876 and rs760426), and *NFKBIE* (rs2233434) (Figure 1.3) (Myouzen et al. 2012; Terao et al. 2011). Studies in individuals of Asian ancestry have confirmed over 15 SNPs that are relevant to RA pathogenesis, some of which are highlighted in Figure 1.3.

Emerging well powered, transethnic and multistage GWAS of Asian ancestry and European ancestry have also replicated some of the previously identified SNPs. Furthermore, the risk loci are common between the two ethnic groups, highlighting that transferability of RA risk loci. The SNP include those in or near *STAT4, BLK*, *AFF3,* and *CCL21*, *CCR6*, *DPP4* and *CDK5RAP2* (Freudenberg et al. 2011; Kochi et al. 2010; Stahl et al. 2010; Lee et al. 2010). In addition, a multi-ancestry comparative GWAS meta-analysis which included 4,074 RA cases and 16,891 controls in the discovery sample of Asian ancestry and 5,277 RA cases and 21,684 controls of European ancestry showed and confirmed that *NFKBIE and TNFAIP3* have common effects in both ancestries (Okada et al. 2012). Taken together, these findings suggested that there was a genetic risk loci overlap between European and Asian individuals, more than can occur by chance. Transethnic studies have since confirmed about ~14 SNPs common to European and Asian ancestry.

1.6.2.4 Non-HLA association in African ancestry

Very few well-powered genetic association studies have investigated the genetic association of RA in African ancestry populations. Earlier studies investigated the replicability of 27 non-HLA European ancestry genetic risk loci in African Americans (Viatte et al. 2012; Hughes et al. 2010). Hughes et al. (2010) reported that *CCR6*, *TAGAP, TNFAIRP3* replicated in 556 seropositive African American RA cases and 791 African American controls. However, this finding was not replicated in 263 RA BSA cases and 374 BSA controls using the Immunochip array (Govind et al. 2014). These findings confirm the inherent genetic difference between African and European ancestry individuals as well as and also within individuals of African ancestry. However, a more recent multi-staged multi-ethnic GWAS meta-analysis study,

which included 966 RA cases and 1,392 controls of African American descent reported a risk locus *PADI2*, which is independent of *PADI4* and overlaps between Eurasian and AA in a trans-ethnic fine mapping (Laufer et al. 2019). In addition, the study reported two AA specific risk loci in *GPC5* and *RBFOX1*. However, these finding warrant replication in an independent cohort.

In comparison to other populations, recent data show that participants of African ancestry remain poorly represented, and contribute only a very small percentage to the total number of individuals included in GWASs (Figure 1.3) (Martin et al. 2019). This is worrisome as possible advantages that come with studying African populations are missed. For example, owing to its deep evolutionary roots, African populations have small LD blocks, which offer an opportunity for fine mapping and makes narrowing down of genetic associations to the causal variant possible. This LD decay in African populations also means there is greater potential for fine mapping to higher resolution to pinpoint causal variants that influence traits of interest compared to other global populations (Wu et al. 2013). The study of African population offers an opportunity to produce genetic associations that can be extrapolated across other populations more accurately as well as allowing for the identification of variants not present in other populations.

In addition, Africans live in a context of diverse environmental pressures and have a high level of genetic diversity, as mentioned above, which provide an opportunity for the discovery of novel genetic associations and consequent biological insights. Furthermore, studying African populations offers an opportunity to investigate the effects or impact of infectious diseases in the development of RA. A better understanding of the genetic architecture of RA in African populations has the potential to build generalised and improved prediction models. Studies have shown that African populations exhibit extreme allele frequency differences involved in several genes and pathways compared to 26 world populations (Sulovari et al. 2017). Combined with the low LD, this makes studying of African populations very attractive for the discovery of novel genetic associations and improved understanding of human disease-related genetics (Tekola-Ayele and Rotimi 2015). Taken together, this supports the inclusion of African populations in global disease-association studies and demonstrates a potential to improve genomic discoveries for medical purposes at an accelerated pace.

1.7 RA treatment

There is currently no preventive treatment or a cure for RA. Treatment options include disease-modifying antirheumatic drugs (DMARD) such as Methotrexate, Sulphasalazine Leflunomide, non-steroidal anti-inflammatory drugs (NSAID), glucocorticoids, and biologics such as anti-TNF alpha drugs are used to treat RA patients. Although the mechanism of action of treatment is poorly understood, it functions by reducing or clearing joint inflammation, thus reducing disease severity (Sizova 2008; Finckh 2006).

1.7.1 Methotrexate

Methotrexate (MTX) is the most widely used DMARD because of its affordability, long-term efficacy and acceptable safety profile. It has therefore been called the “anchor drug” for the treatment of RA (Pincus et al. 2003). The precise mechanism of action of MTX in the treatment of RA is unclear. However, it is thought to inhibit genes critical in DNA synthesis, repair and cellular replication in the folate pathway (Chan and Cronstein 2002). This inhibition consequently results in the accumulation of adenylate and adenosine, which leads to the systemic reduction of antigen dependent T-cell proliferation and anti-inflammation effects in the synovium (Kaltsonoudis et al. 2012; Wessels et al. 2008).

1.7.1.1 Methotrexate anti-inflammatory effect

Although there have been several suggested mechanisms suggested that lead to accumulation of adenosine, they all have in common the polyglutamation of MTX in common. Polyglutamation is the addition of glutamates by folylpolyglutamate synthetase (*FPGS*) to the MTX molecule once inside the red blood cell. The polyglutamation of MTX to MTX-GLU increases the cellular half-life compared to the MTX wild type. Two prevailing theories exist, with one suggesting that the MTX-GLU inhibits aminoimidazole-4-carboxamide ribonucleotide (*AICAR*) transformylse, which leads to the accumulation of AICAR and adenosine whose activated receptor exhibit anti-inflammatory effect (Cao et al. 2018; Cronstein and Sitkovsky 2017; Zhao et al. 2016; Kremer 2006; Cronstein 1996). The other suggests that accumulation of MTX-GLU inhibits dihydrofolate reductase (*DHFR*) and thymidylate synthase (*TS)* involved in purine biosynthesis (Friedman and Cronstein 2018; Budzik et al. 2000). Figure 1.5 illustrates a potential MTX mechanism of action in RA patients (Pastore et al. 2015). More recent studies have reported that MTX inhibits both T-cell and FLS activity, suggesting a role in the inhibition of NF-κB activity (Bergstrom et al. 2018).

Cumulatively, the above-mentioned studies provide evidence of the complex nature of MTX action and insight into the prioritization of potential target genes.

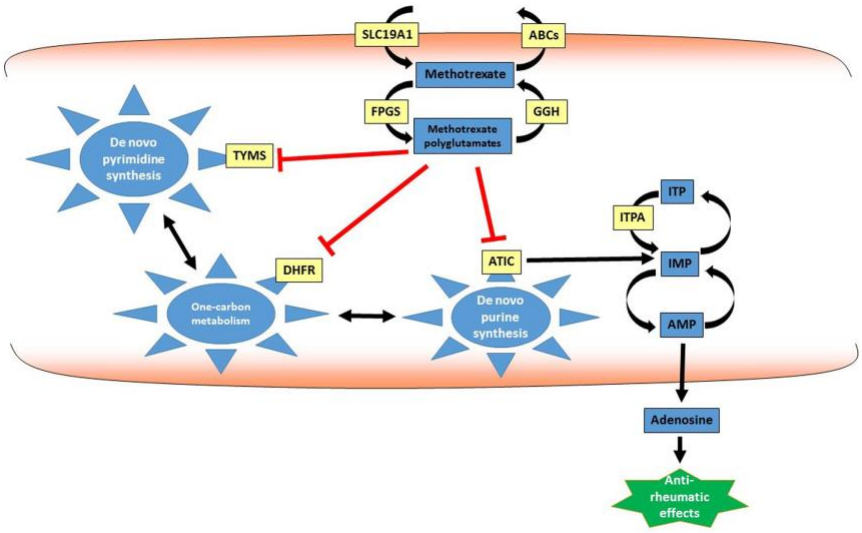


Figure 1.5: Illustration of a potential MTX molecular mechanism of action that leads to an anti-inflammatory effect in RA patients. The role of each gene from the point MTX enters the red blood cell via the human solute carrier (SLC family of genes), polyglutamation by *FPGS* and its inhibitory effects to either one or all three genes i.e. *TYMS, DHFR, ATIC* is demonstrated. This inhibitory effect is thought to result in the anti-rheumatic effects via the accumulation of adenosine (Pastore et al. 2015).

Several comparative studies revealed that low-dose MTX is effective and has a better safety profile than other DMARD (Braun 2011). In addition, early initiation of aggressive treatment of RA with MTX monotherapy or in combination with other DMARD (i.e. Sulphasalazine and leflunomide) increased remission and reduced radiographic progression similar to biologic agents in RA (Soukup et al. 2016; Santos-Moreno et al. 2015; Choy et al. 2008). Thus, a delay in diagnosis and initiation of therapy significantly results in significant functional impairment (Hodkinson et al. 2012). Although efficacious and very important in RA treatment, about 40% of RA patients do not benefit from MTX treatment (Salliot and van der Heijde 2009). Furthermore, adverse effects (AE) such as hyperhomocysteinemia, gastrointestinal disorders, hepatotoxicity, pneumonitis, infection and bone marrow suppression often lead to discontinuation of treatment in up to 35% of patients (Wang et al. 2018; Gilani et al. 2012). Although the AE are mostly observed with high doses of MTX, hepatotoxicity, pulmonary damage and myelosuppression are also observed with low doses in some RA patients (Kivity et al. 2014; Attar 2010). The inter-patient response variation also leads to treatment inefficacy, with the dropout rate approaching 51% in a South African DMARD naïve cohort

(Hodkinson et al. 2012). Thus, there is a need for biomarkers that will enable treatment monitoring in RA patients taking MTX.

1.7.2 Biologics

Biologics are treatment agents developed from proteins. Unlike DMARD whose mode of function is not completely understood and are not specific, biologic therapies are designed to block a specific key chemical or cellular pathway involved in RA inflammation. Biologic therapies used in RA treatment include TNF inhibitors, T-cell co-stimulation inhibitors, Janus Kinase (JAK) inhibitors, IL-6 receptor-blocking monoclonal antibodies and IL1-R-binding monoclonal antibodies (Boyce et al. 2018; Langdon and Haleagrahara 2018; Smolen et al. 2016). Recent studies have shown that efficiency is increased when biologics are used in combination with DMARD (Emery et al. 2018). Although very effective in many RA patients, they remain prohibitively very expensive for widespread use, particularly in low-income countries such as those in Africa.

1.7.3 Methotrexate pharmacogenetics

The majority of pharmacogenetic studies have been conducted using the candidate gene approach to detect associations. These studies were predominately conducted in patients of Eurasian ancestry with sample sizes ranges between 61 and 309 (https://www.pharmgkb.org/). Although using small sample sizes and often underpowered to detect an association of rare SNPs, these studies effectively demonstrated that genetic variation in genes encoding enzymes involved in the folate-related genes of MTX metabolism. These include MTX transporter enzymes (ABCs, *RFCs*), MTX-PG genes (*FPGS, GGH),* MTX-PG target genes (*DHFR, TYMS, ATIC, MTHFR*), influenced drug responses (Qiu et al. 2017; Restrepo et al. 2016; Ranganathan 2008; Dervieux et al. 2006; Ranganathan et al. 2004). Although inconclusive, these studies have also shown that drug response exhibited interpersonal and interethnic variations, thus affecting effective disease treatment and management. To improve sample sizes and consequently statistical power, meta-analysis and systemic reviews were conducted. Although very a very useful approach, validation of results remains a challenge. For example, a SNP in RFC1 (rs1051266) was found to be associated with MTX response in patients of European ancestry (Kung et al. (2014). However, the results did not replicate in another population (Li et al. (2016). A summary of meta-analyses performed to date is presented below (Table 1.1).

Table 1.1: Methotrexate response pharmacogenetics meta-analysis studies conducted in patients of Eurasian ancestry

Gene	Population	Number of studies	Cases sample size	Study outcome	Reference
RFC1 (rs1051266)	European	87	1,019 responders; 588 non-responders; 478 toxicity; 886 non-toxic	SNP is associated with MTX efficacy but not toxicity	Kung et al. (2014)
RFC1 (rs1051266)	Eurasian	17	1,187 responders; 588 non-responders; 978 toxicity; 1,649 non-cases	SNP was associated with MTX efficacy but after stratification, SNP was associated with efficacy in patients of Asian ancestry	Li et al. (2016)
MTHFR (rs1801133 and rs1801131)	Transethnic	8	1,514 RA patients	No association with *MTHFR* SNPs	Lee and Song (2010)
	Predominantly Caucasian patients	4	603 RA Patients	*MTHFR* was not associated with MTX response	Morgan et al. (2014)
ABCB1 C3435T	Transethnic	14	-	Variant was not associated with efficacy but toxicity	Lee et al. (2016)
ATIC (C347G)	Eurasian	9	1056 RA patients	Variant associated with non-response in European and not in Asian. GG+GC haplotype was associated with toxicity in European but not in patients of Asian ancestry.	Lee and Bae (2016)
28 SNPs	European	1	194 RA patients	FPGS and ABC1 genetic variants were associated with response	Moya et al. (2016)
88 SNPs in 28 genes	Eurasian	42	4822 RA patients	*RFC (*rs1051266) was associated with toxicity	Qiu et al. (2017)
25 SNPs	European	1	915 RA patients	MTRR (rs1801394), SNP was replicated	López-Rodríguez et al. 2018

In AA, the only African ancestry population studied to date, it was shown that *MTHFR* SNPs rs1801131 and rs1801133 occur at a low frequency of 0.11 and 0.13 compared to European ancestry population where a frequency of 0.30 and 0.34 were observed respectively (Hughes et al. 2006). However, Ranganathan et al. (2008) reported an association between *MTHFR* (rs1801133) with AE in African Americans, which conflicted with previous findings. Furthermore, studies have shown that the *MTHFR* SNPs are associated with AE in European ancestry but not in African American, which could be attributed to the allele frequency difference. However, the inverse was observed with another *MTHFR* SNP, rs4846051, which occurred at a higher frequency in African Americans (0.33) compared to Caucasians (0.08) and was associated with AE in African Americans only (Hughes et al. 2006), suggesting locus heterogeneity. However, the results remain inconclusive and lack of pharmacogenetic studies in patients of African ancestry further complicates matters. One of the limitations of candidate gene studies is that they are based on a priori hypothesis about the role of the selected genes and require prior knowledge of the gene's role in the biology of the phenotype (Alghamdi and Padmanabhan 2014).

To circumvent this gap in knowledge, GWAS have since been conducted to investigate genetic variants associated with MTX response that may have been missed with candidate gene studies. A GWAS, in conjunction with SVM analysis of MTX treatment response in 297 responders and 160 poor responders of Asian ancestry identified seven gene loci albeit at a slightly weaker than the genome-wide statistical significance (Senapati et al. 2018). In addition, the study replicated three gene loci in *DHFR, FPGS,* and *TYMS* previously associated with response (Senapati et al. 2018). In another GWAS study of 1424 early RA patients from two European ancestry consortia, the strongest evidence for association was with *NRG3* locus rs168201 (Taylor et al. 2018). However, the association did not reach genome-wide association significance, which could be due to the modest sample size. Taken together, these findings highlight the need for more pharmacogenetic studies in patients of African ancestry.

1.8 Functional annotation of GWAS of RA

The past decade has seen a rapid increase in the number of RA GWASs, which has culminated in the confirmation of over 100 associated genetic risk loci. Several of these GWASs confirmed that RA risk SNPs have been identified over the past five years due, at

least in part, to large international collaborations, and advances in technology and statistical methods. The discoveries have improved our understanding of RA pathogenesis considerably. However, over 90% of GWAS discoveries are enriched for associated markers in the non-coding regions of the genome (Kumar et al. 2013). These polymorphisms do not lead to a change in protein structure, thus making interpreting, translating and elucidating pathways of disease susceptibility, progression, and outcome of post-GWAS analyses a challenge. This is because GWASs identify associations with SNP markers and not necessarily the causal SNP, as a result of the presence of LD (Ardlie et al. 2002; Gabriel et al. 2002). Consequently, SNPs in strong LD within a haplotype in will have similar strengths of statistical significance, thus requiring empirical and experimental evidence to identify the functionally relevant SNP (Edwards et al. 2013; Schaub et al. 2012). There is currently a pressing need to understand the functional influence of non-coding region in pathogenesis.

While functional laboratory-based or animal-based studies to characterise GWAS loci are often laborious, genomic and epigenomic data generated by international collaborations such as the 1000 Genomes Project (Edwards et al. 2013), Encyclopaedia of DNA elements (ENCODE) project (ENCODE project consortium 2012), the NIH Roadmap Epigenomics Genome Project (Romanoski et al. 2015), Gene Expression Omnibus (GEO) (Edgar et al. 2002), the Genotype-Tissue Expression (GTEx) project (GTEx Consortium 2017) and others, provide extensive functional characterisation of tissue cell-type-specific putative cis-regulatory elements (CREs). These data facilitated the development of publicly available, web-based functional annotation tools such as Ensembl's Variant Effect Predictor (VEP) (McLaren et al. 2016), web ANNOVAR (wANNOVAR) (Chang and Wang 2012; Wang et al. 2010), RegulomeDB (Boyle et al. 2012) and HaploReg (Ward and Kellis 2012), that can be used to prioritise GWAS identified disease-associated SNPs.

More specifically, VEP is a publicly available computational toolset with access to an extensive range of genetic annotation of coding and non-coding variant transcripts (McLaren et al. 2016). Furthermore, this tool also offers protein deleterious prediction algorithms which generate scores for SNPs in coding regions, including Sorting Tolerant From Intolerant (SIFT) (Vaser et al. 2016) and Polymorphism Phenotyping-2 (PolyPhen2) (Adzhubei et al. 2013) as well as non-coding SNPs, including Combined Annotation Dependent Depletion (CADD) and Loss of Function (LofTool) (Fadista et al. 2017). Similar to VEP, wANNOVAR provides functional annotation of genetic variants which can also, among other things, report

conservation levels and predict functional consequences (Chang and Wang 2012). RegulomeDB is another publicly available database that characterises CREs in non-coding regions of the genome and generates a score (1 – 7) based on the experimental evidence for a particular SNP (Boyle et al. 2012). Depending on how many regulatory elements a variant overlaps, the tool generates a score reflecting the alteration, where a score of 1 would be interpreted as highly likely to influence binding and 7 as less likely. This can therefore be used to prioritise disease-associated SNPs (Appendix C).

Instead of analysing each SNP separately, RegulomeDB and HaploReg combine polymorphisms and analyses into haplotype blocks and estimate the LD between different SNPs, providing the utility for fine mapping of GWAS findings (Ward and Kellis 2012). Furthermore, HaploReg also provides SNP sequence conservation data by integrating evolutionary genome conservation data from Genome Evolutionary Rate Profiling (Cooper et al. 2005). GTEx is a publicly available database of gene expression quantitative loci (eQTL) of human non-disease tissues (GTEx consortium 2017).

The functional annotation of non-coding SNPs is still a challenge and therefore discrepancies in prediction from different tools for the same SNP are not uncommon (Shaik and Banaganapalli 2019; McCarthy et al. 2014). To bridge the gap and improve confidence in the predicted results, more than one tool is used to predict the functionality of SNPs identified in GWAS. For example, Arshad et al. (2018) used five *in silico* tools which included SIFT and PolyPhen-2 along with three other tools to show and prioritise the glycine to glutamic acid change at position 120, glycine to tryptophan change at position 141 and a valine to methionine change at position 151 as major mutations in the native *TAGAP* gene product. Furthermore, Houtman et al. (2018) integrated GTEx, RegulomeDB and HaploReg v4.1 along with experimental evidence to show that previously identified RA SNP in *PTPN2* are located in transcription binding sites and are highly correlated with DNA methylation levels at four CpG sites and expression levels of long non-coding RNA in *LINCO1882* in whole blood.

In a juvenile idiopathic arthritis (JIA) GWAS meta-analysis with 2,571 patients and 15,886 controls of European ancestry, RegulomeDB and HaploReg were integrated to show that the identified SNPs overlapped with regulatory regions, influenced gene expression and altered binding affinity in immune-related pathways (McIntosh et al. 2017). Walsh et al. (2016) used HaploReg to define SNPs in LD with the GWAS identified SNP and GTEx whole blood

dataset to map eQTL identified in RA diseased whole blood. However, studies that only used a single tool to assess the potential functional significance have also been reported. For example, Thalayasingam et al. (2018) used GTEx to confirm and prioritise cis-eQTLs common to CD4+ and B lymphocytes at RA risk loci and genes in specific immune cell types obtained from 344 incident RA patients. Cumulatively, the studies described above demonstrated the utility of *in silico* functional annotation in prioritising genes for experimental validation.

1.9 Study Rationale

Rheumatoid arthritis is a complex multifactorial inflammatory disease whose etiology remains poorly understood. The rapid advancement of GWAS technologies and statistical methods have accelerated the identification of RA genetic risk loci. More specifically, GWAS have identified and validated over 106 RA risk loci which have improved the current understanding of disease pathogenesis. Furthermore, pharmacogenetic studies have identified several SNPs associated with response to MTX. More recently, studies evaluating the genetics of RA have integrated GWAS findings with biological functional data, which has enhanced the interpretation of GWAS findings and thus provided insight into the pathogenesis of RA. Indeed, much progress has been made in understanding the genetics of RA. However, despite the progress made, there is currently a dearth of genetic data on susceptibility to RA and response to MTX treatment in populations of African ancestry, and more specifically in BSA. The majority of RA risk loci have been identified in populations of European and Asian ancestry. There is only one large, high-density genome association study of RA in BSA that was performed using the Immunochip SNP array, enriched for SNP markers previously associated with immune-related disorders. Although a very important study, it used a small sample size and the Immunochip SNP array, which was limited in its scope to identify novel associations. It has recently been shown in transethnic genetic studies that some, but not all, RA genetic risk loci may be shared between studied populations (e. g. HLA locus).

African populations tend to have small LD blocks, thus identifying an association in an African population in conjunction with biological functional data may assist in fine mapping and narrowing the associated region down in order to identify the causal loci. Furthermore, higher genetic diversity and unique genetic variants provide an opportunity for novel discoveries that may help elucidate the pathogenic mechanism of RA. Furthermore, assessing the pharmacogenetic variants that influence MTX metabolism in European and Asian populations could assist in understanding the heterogeneity of MTX response in Africans and could potentially identify biomarkers to predict drug response. It is envisaged that the information gathered in this study will expand the current knowledge of RA pathogenesis and the pharmacogenetic impact on MTX efficacy in BSA patients with RA. Such findings could potentially assist in accurate diagnosis and intervention through the genotyping of response biomarkers (i.e. SNPs) in BSA patients in a clinical setting.

1.10 Aim and objectives

The aim of this study was to use a GWAS approach to identify the genetic variants associated with susceptibility to seropositive RA and to evaluate pharmacogenetic variants associated with methotrexate response in Black South Africans (BSA).

Specific aims:

1. To identify novel and replicate previously reported RA associated genetic loci at the HLA locus and non-HLA loci in a Black South African population in a case-control genome-wide association study.
   - Genotype RA patients using the H3Africa SNP genotyping array.
   - Perform genotype imputation of the H3Africa array data using an African reference panel.
   - Perform a case-control genetic association study using data from ethnically and geographically aligned population controls (case:control ratio of ~1:3) genotyped on the H3Africa SNP genotyping array

2. To perform *in silico* functional analysis for significantly associated regions and genes.

3. To evaluate whether previously identified loci (either the exact variants or variants in the surrounding region) are associated with the response to MTX therapy in BSA RA patients treated with MTX
   - Select MTX responders and non-responders from the cohort using the METEOR database
   - Perform exact and regional association of SNPs previously associated with response to MTX in RA patients.

CHAPTER 2: GENOME-WIDE ASSOCIATION STUDY IN BLACK SOUTH AFRICANS WITH RHEUMATOID ARTHRITIS

2.1. Abstract

Rheumatoid arthritis (RA) is a complex chronic inflammatory disease with no cure. The heritability of RA is estimated to be 60% and no large genome-wide genetic association studies have been performed in African populations. This study used a case-control genome-wide association study approach to assess genetic susceptibility to RA in 577 Black South Africans (BSA) seropositive RA cases and 1612 population controls, by using a logistic regression analysis assuming an additive model. Association was adjusted for sex, age, smoking, and the first three principal components to adjust for possible population structure. This study confirmed the strong association between increased risk and genetic variants in the HLA region. Furthermore, this study identified four novel single nucleotide polymorphisms (SNPs) associations in the intron regions of the *CPT1A* gene, which has been shown to be differentially expressed in RA patients, compared to unaffected controls. GWAS results were followed by integration with functional annotation of variants obtained from tissues and cells relevant to RA pathogenesis, revealing that the associated SNPs are likely involved in the regulation of the expression of *CPT1A* and other nearby genes. Our study is the first and largest GWAS of RA in sub-Saharan Africa.

2.2. Introduction

Rheumatoid arthritis (RA) is a chronic systemic inflammatory disease afflicting approximately 1% of the adult population worldwide (Silman and Pearson 2002). The disease is characterised by symmetrical chronic inflammation and damage of the synovial lined joints, primarily of the hands and feet, that can lead to significant disability if untreated (Firestein 2003). In addition, the presence of rheumatoid factor (RF) and the more specific anti-citrullinated protein antibodies (ACPA), remain important serological determinants of an RA diagnosis according to the ACR-EULAR classification criteria (Aletaha et al. 2010). RA affects women twice as often as men, peaking in the 4th – 5th decade of life (Silman and Hotchberg 2001). RA patients experience higher rates of premature death compared to the general population (Sokka et al. 2008), with cardiovascular diseases reportedly accounting for half the cases in a meta-analysis study of 111,758 RA patients (Avina-Zubieta et al 2008).

RA in African ancestry populations was initially thought to be uncommon and milder than in Caucasians (Solomon et al. 1975). However, several recent studies in urban African ancestry populations have shown that the disease often has a profound impact on functional disability, economic burden and health-related quality of life (Mody et al. 1988; Kalla and Tikly 2003). Furthermore, significant radiographic changes of RA are frequently observed in African ancestry RA patients (Hodkinson et al. 2012; Mody and Meyers 1989). In Soweto, South Africa, women with RA in conjunction with low income and lack of healthcare services were found to have compounded social exclusion and loss of independence (Schneider et al. 2008). It is estimated that a third of BSA with RA require financial assistance from the state (Mody et al. 1988).

The exact etiology of RA remains elusive. However, it is now widely accepted that RA manifests from a complex multifactorial interaction between genetic, environmental, infectious, diet, microbiome, and hormonal factors in genetically susceptible individuals (Glossop et al. 2006; Silman and Pearson 2002; Vaughn et al. 1988), which results in immune dysregulation associated with the pathogenesis of RA. This immune dysregulation leads to loss of self-tolerance. Consequently, cellular events such as aberrant proliferation, migration, invasion of fibroblast-like synoviocytes and production of pro-inflammatory cytokines by activated synovial tissue macrophages, cause chronic inflammation and hyperplasia in the RA inflamed joint (Bottini and Firestein 2013; Firestein 2003).

Twin studies strongly support a genetic contribution to RA risk where a fourfold higher concordance is observed in monozygotic, compared to dizygotic twins. Heritability is estimated to be approximately 60% (MacGregor et al. 2000). Moreover, family studies estimate RA prevalence to be three times higher in first-degree relatives compared to the general population (Frisell et al. 2013). Genetic susceptibility in the presence of unfavorable environmental conditions, such as smoking, results in the formation of citrullinated peptides in the lungs and the production of ACPA (Hutchingson et al. 2001).

The HLA region located on chromosome 6 shows the strongest genetic association with RA and accounts for over 30% of the heritability in all populations, including Africans. More specifically, the variation at the *HLA-DRB1* locus accounts for a third of the genetic risk. Alleles of *HLA-DRB1* that encode a specific amino acid sequence with a common motif sequence i.e. –RAA at position 70-74, termed the 'shared epitope', was the first to be associated with RA (Gregersen et al. 1987). In BSA, earlier studies using low-resolution techniques showed that, like in Caucasian populations, the HLA class II antigens, DR4, DR1, and DR10, have an increased frequency among BSA people with RA (Mody et al. 1989; Martell et a*l*. 1990; Pile et al. 1992). Recent work by Meyer and colleagues (2011) using a medium to high-resolution identification of *HLA-DRB1* alleles showed a strong association with certain *HLA-DRB1* alleles especially *0401, *0404, *0405, *0410, with RA. In a study using the single SNP Immunochip array, the previously known association with *HLA-DRB1* was confirmed in BSA (Govind et al. 2014).

However, fine mapping of the amino acid beta-sheet of *HLA-DRB1* has revealed interesting associations and ethnic variation. In particular, amino acid on positions 11 and/or 13 conferred susceptibility to RA in European, Asian and African ancestry, amino acid 57 in African and Asian ancestry and amino acid 74 in Asian and European and amino acid 71 only in European (Danila et al. 2017; Traylor et al. 2017; Saxena et al. 2017; Reynolds et al. 2014; Okada et al. 2014; Ling et al. 2016; Raychaudhuri et al. 2012). Govind et al. (2019) confirmed the association of the amino acid polymorphisms at positions 11, 13 and 33 with increased risk of RA in BSA.

The completion of the 1000 Genomes Project (KGP) (The 1000 Genomes Project Consortium 2015), African Genome Variation Project (AGVP) (Gurdasani et al. 2015) and the Human Heredity and Health Africa (H3Africa) projects enabled the development of genotyping arrays

and reference panels with better representation of African populations. One of these is the recently developed H3A Consortium SNP genotyping array (The H3Africa Consortium 2014). This GWAS array contains approximately 2.4 million SNPs enriched for common African variants and mapped to the Genome Reference Consortium Human build 37 (GRCh37) (The H3Africa Consortium 2014).

Non-HLA genetic association overlap between Europeans and African Americans (AA) from the Consortium for the Longitudinal Evaluation of African Americans with Early Rheumatoid Arthritis (CLEAR) has been reported in candidate gene studies. More specifically, twenty of 23 genes confer risk in both Europeans and AA (Hughes et al. 2010). However, AA are genetically an admixed population and thus do not represent the full genetic diversity of all African populations. In BSA, the *PTPN22* gene variant rs2476601 that confers the strongest non-HLA risk in Europeans and AA does not confer risk in BSA, as it is not polymorphic in this population (Govind et al. 2014; Tikly et al. 2010). In a study using the Immunochip, Govind et al. (2014) identified several novel non-HLA associations in BSA. These observed novel associations included SNPs in intergenic regions between (*LOC389203*|*RBPJ, LOC100131131*|*IL1R1, KIAA1919*|*REV3L, LOC643749*|*TRAF3IP2),* SNP in the intron and *UTR of IRF1* and SNP the intronic region of *ICOS and KIAA1542* (Govind et al. 2014). The study also replicated a known association of a SNP in the *RBPJ* gene (rs874040) on chromosome 4 with RA in Europeans (Stahl et al. 2010). However, these findings warrant validation in a larger BSA cohort. There is currently a paucity of genomic data on African populations, which results in deficiencies in understanding the immunogenic basis of the pathogenesis of RA. Understanding the aetiology and pathogenesis of RA is essential to the development of better biomarkers and to predict treatment response.

The aim of the research in this section was to identify genetic variants associated with increased risk for RA in BSA, using a GWAS approach.

2.3. Materials and methods

2.3.1. RA Cases

A total of 647 unrelated seropositive RA patients (584 previously unstudied and 63 individuals were part of the Immunochip study) from the Soweto, Johannesburg metropolitan area, were included for this case-control genetic association study for RA. The patients were recruited from the outpatient Arthritis Clinic of the Chris Hani Baragwanath Academic Hospital (CHBAH), Soweto. To be included in the study, patients had to have been diagnosed with RA by fulfilling the 2010 ACR/EULAR classification criteria for RA (Aletaha et al. 2010). The patients received a medical history interview, physical examination and routine laboratory testing for every visit, as deemed fit by the consulting physician. Furthermore, the patients were consenting, unrelated BSA and were at least 18 years of age at disease onset. The patients were considered 'black' if they self-reported all four grandparents as being BSA. All participants gave written informed consent and the study was approved by the Human Research Ethics Committee (Medical), University of Witwatersrand (HREC protocol number M1706109, Appendix A).

2.3.2. Population controls

As controls for this study, we used data from an existing cohort at the Sydney Brenner Institute for Molecular Bioscience (SBIMB). Population control individuals were selected from the AWI-Gen study of the Human Heredity and Health in Africa (H3Africa) Consortium (Soweto and Dikgale, South Africa). The AWI-Gen (Africa Wits-INDEPTH Partnership for Genomic Studies) study is a population cross-sectional study of adults aged 40 to 60 years (Ramsay et al. 2016). This study recruited participants from South Africa, Kenya, Burkina Faso, and Ghana, but we only selected individuals from South Africa to be used as controls to ensure that they are ethnically matched and to reduce potential population sub-structure as a confounder in this study (see selection criteria below Figure 2.1). In addition to the genomic data (also described below), limited clinical data were used in the analyses. Informed written consent was obtained from each study participant. The collection of these samples under the AWI-Gen study was approved by the Human Research Ethics Committee (Medical), University of the Witwatersrand (HREC protocol number M121029).

2.3.2.1. Population controls selection

The controls were selected from only two of the AWI-Gen South African study sites, from Soweto (Gauteng Province) and Dikgale (Limpopo Province), as these sites had appropriate ethnically matched controls. Following stringent quality control, a total of 3,052 controls were included for assessment. The controls were then further selected for inclusion in the study as depicted in the flowchart (Figure 2.1) and detailed below. A Principal Component Analysis (PCA) was conducted using EIGENSTRAT SMARTPCA (Patterson et al. 2006) and visualised in Genesis (http://www.bioinf.wits.ac.za/software/genesis/) to assess the genetic structure and to remove extreme population outliers. Samples with missing data were removed and, then, to achieve a 1:3 case-control ratio, samples were further removed at random. In the final analyses, 1612 controls were used.

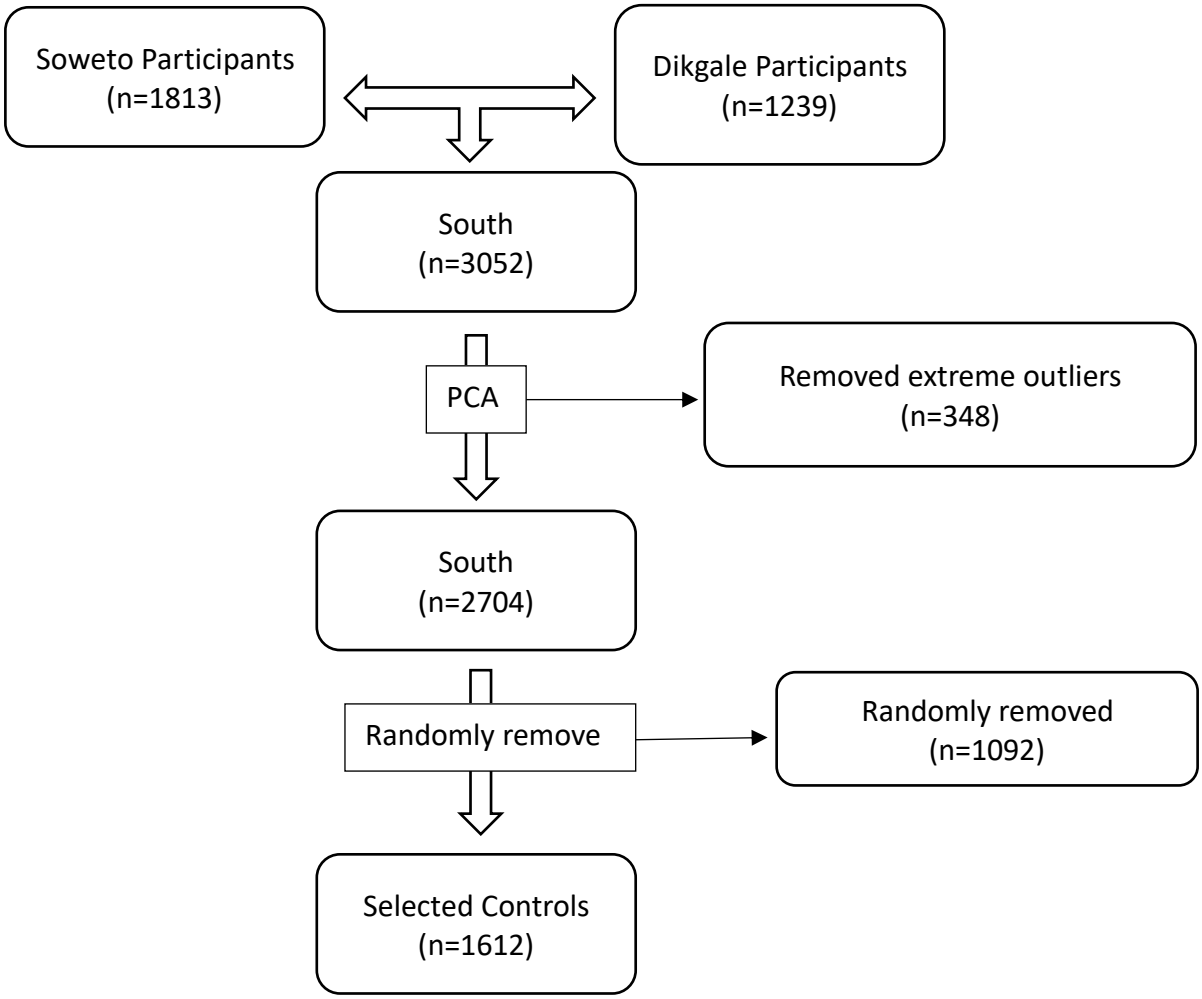


Figure 2.1: A flow diagram showing the selection process for the population control individuals in the study from Soweto and Dikgale (AWI-Gen study).

2.3.3. Patient data collection and quality controls

Patient demographic information, including age, sex, smoking history and clinical and serological (ACPA and RF) data were collected and stored on the password protected online international Measurement of Efficacy of Treatment in the Era of Outcome in Rheumatology (METEOR) database (http://www.meteorfoundation.com/). The South African patient data in this database is managed from the Chris Hani Baragwanath Academic Hospital Arthritis Clinic and is used to collect and store patient clinical and demographic data at each clinic visit. The data that was not captured or incompletely captured in METEOR was then obtained manually from patient files with the help of a research administrator. Patients with inconsistent information captured in METEOR were excluded from the study if the inconsistency could not be resolved.

2.3.4. Laboratory methods

2.3.4.1. DNA extraction and quality control

Blood samples (in EDTA as anti-coagulant) from 584 participants with RA were collected under a separate protocol and ethics approval (certificate number **M10707**) from a previous study titled “The genetics of rheumatoid arthritis in Black South Africans” (Govind et al. 2014). DNA was already extracted and stored for 384 samples. I extracted DNA for the remaining 200 samples at the SBIMB using the salting-out method (Miller et al. 1988), with minor modifications. DNA extraction using this method takes place in two parts and spans two days. These DNA samples were available and stored at -80°C at the SBIMB Biobank, Parktown, Johannesburg, South Africa.

The first part of the DNA extraction involved completely thawing samples to room temperature. Each sample was transferred to an appropriately labeled 50 ml NUNC™ tube and the tube was filled with cold sucrose-Triton lysing buffer, which lyses the red blood cells. The tubes were inverted several times until completely mixed and placed in a centrifuge (Beckman Coulter, USA). Using this procedure, the blood was centrifuged at 2400 rpm at 4°C for 10 min. After that, the supernatant was poured off gently into a bleach-containing container, ensuring that the pellet did not dislodge. The pellet was washed with 20 – 25 ml of cold sucrose-Triton X lysing buffer. After that, the tubes were placed in a -20°C freezer for 5 minutes and then spun down for 5 minutes. The supernatant was poured off as before. This pellet washing step was repeated at least twice to ensure that most of the red cell debris was

completely removed. About 3 ml T20E5, 200 µl of 10% SDS and 500 µl of the proteinase-K mixture were added to the pellet containing tube and mixed by inversion. The tubes were then incubated overnight at 42°C. This step lyses the white blood cells to release the DNA into the solution and to degrade the proteins.

The following morning about 1 ml of saturated NaCl was added to the lysate containing tube and mixed by vigorous agitation for about 20 seconds. The tubes were placed at -20°C for 5 to 10 minutes. The tubes were then spun down for 30 minutes until a white pellet appeared at the bottom of the tube. After that, the supernatant was transferred to an appropriately labeled 50 ml NUNC™ tube. The tube with the pellet was then discarded. To the supernatant, about two supernatant volumes of absolute ethanol were added and mixed by gentle agitation. This allowed the DNA to aggregate in the solution. The DNA was then "fished" out from the solution and placed in an appropriately labeled 1.5 ml cryovial. In order to remove excess salt, the DNA was then washed with ice-cold 70% ethanol. The DNA was air-dried by placing the tubes on the bench and then resuspended in 400 µl TE buffer.

2.3.4.2. DNA quality control and normalization

In order to reduce or prevent downstream process failures, the quality and concentration of the extracted DNA was assessed using a NanoDrop™ Spectrophotometer (ThermoFisher Scientific™). The quality and concentration were assessed by measuring the ratio of absorbance of 1 µl DNA at 260/280nm wavelength and results recorded. DNA samples with a ratio of absorbance of approximately 1.8 to 2.0 were accepted as good quality and an absence of protein contamination. The DNA samples were normalised to at least 30 ng/µl and 35 µl was transferred to Infinium Genotyping Assay bar-coded Midi Plates (Illumina®) in preparation for genotyping. A DNA sample manifest was prepared to provide information on each sample, and the samples were shipped on dry ice to the service provider by the SBIMB Biobank personnel.

2.3.4.3. SNP genotyping with the H3Africa SNP array

DNA samples from of 647 RA cases were genotyped at Illumina® FastTrack™ Microarray services (USA), using the H3Africa Consortium genotyping array, which contains approximately 2.3 million SNP, specifically designed to be enriched with common African

SNP (The H3Africa Consortium 2014). The controls were genotyped on the same array, as part of another study and the genotyping data were used in the present study.

2.3.5. SNP genotype data processing

Following pre-processing steps, quality controls steps were performed on 618 RA cases (29 samples failed genotyping), 1,612 controls and 2,267,346 SNPs on the array. The SNPs were mapped to the Genome Reference Consortium human Build 37 (GRCh37) using the H3ABioNet/H3Agwas pipeline (https://github.com/h3abionet/h3agwas) prior to GWAS analysis. Briefly, this pipeline performs individual sample and SNP QC as detailed below:

Sample QC – The pipeline first assesses and filters individuals with more than 2% genotype missingness. Next, individuals with discordant sex are removed by comparing the stated sex in the clinical database with the biologically determined sex from the genotyping array results, based on the homozygosity rate for markers on the non-recombining portion of the X chromosome (>0.8 for males and <0.2 for females). X chromosome homozygosity rates greater than 0.2 and less than 0.8 are flagged as potential problems. Samples flagged by this step were verified by checking the sex reported and captured in METEOR with the patient file and, where appropriate, the sex was corrected, or the individual was removed from the data. The sex discordancy could arise as a result of the collection tube being mislabelled or by incorrect capturing of sex data during the recruitment process. Those individuals for whom sex could not be verified were then removed from the analysis. The pipeline further does an individual related assessment as measured by “identity by descent (IBD)” metrics of independent SNP between pairs of individuals. The metrics generate IBD scores and when an IBD is 1, this indicates that the samples are duplicates or come from monozygotic twins, an IBD of 0.5 indicates first degree relatives, an IBD of 0.25 indicates second-degree relatives and an IBD of 0.125 indicates third-degree relatives. The pipeline cut-off is set to remove individuals with IBD > 0.2 to avoid introducing bias into the analysis and inflating the association. Next, the pipeline checks for extreme heterozygosity in the individual genotype data, which may indicate DNA sample contamination if too high and inbreeding if too low. Individuals with heterozygosity above 0.343 and below 0.15 were removed from the analysis.

SNP QC – The SNPs were removed from further analysis for any of the following reasons: (i) genotype missingness rate above 2%; (ii) Hardy-Weinberg equilibrium P-value <1X10^{-06};

(iii) MAF <0.05. The MAF cut-off of 0.05 was used in this study because of the small sample size and lack of statistical power to detect an association with low MAF.

Batch effect – Although the cases and controls were genotyped in different batches, they were, genotyped by the same service provider using the same genotyping platform (same chemistry). However, the potential for a batch effect to confound the genetic association needs to be addressed as this could result in spurious associations. The raw datasets of the cases and controls were compared by calculating and comparing the minor allele frequencies, genotyping quality and genotype call rate across all SNPs. No batch effect was identified between the cases and controls. PCA analysis also did not reveal significant batch effects.

In addition, this study assessed the presence of population stratification between the cases and controls by comparing the distribution of all individuals using PCA and based on the independent SNP in the genotype data using EIGENSTRAT SMARTPCA (Patterson et al. 2006).

2.3.6. Genotype imputation and post imputation quality control

To improve the statistical power to detect an association signal, imputation of the merged cases and controls dataset after QC processes was conducted. The imputation was performed with the help of Dr Dhriti Sengupta (SBIMB Post-doctoral fellow) using the Sanger Imputation Server and the African Genomes Resource as reference panel. Imputation was performed using the EAGLE2+PBWT pipeline which phases chromosomes (haplotypes) against the African Genome Resources reference panel using EAGLE2 (v2.0.5). The following pre- and post- imputation steps, and QC steps detailed below were performed on the merged data:

Pre imputation - SNPs with a missing call rate of above 0.05, MAF less than 0.01 and HWE P-value less than 0.0001 were removed. Individuals with missingness above 0.05 were also removed. Next, the SNPs reference alleles were matched to that of the GRCh37 reference alleles. SNPs that did not match were removed.

Imputation - The cleaned dataset was then imputed using the Sanger Imputation Server following the EAGLE2+PBWT pipeline which, phased chromosomes (haplotypes) against the

African Genome Resources reference panel using EAGLE2 (v2.0.5). The phased data was then imputed using the default algorithm PBWT and the same reference panel.

Post-imputation - VCFTOOLS and BCFTOOLS were used to remove SNPs with MAF less than 0.01, missingness above 0.05, HWE P-value less than 0.00001, poorly imputed SNPs with IMPUTE2 info score of less than 0.06.

Following the processes above, compressed comma separated value (CSV) files were received for each chromosome. The CSV files were converted to plink bed/bim/fam format and merged using a script kindly provided by Dr Jean-Tristan Brandenburg (SBIMB). Thereafter, the files were filtered for SNPs with at least a MAF 0.05, to identify only the common SNPs. The sample size was not powered to detect associations with a MAF less than 0.05.

2.3.7. Statistical analysis

Analysis of the data was carried out mainly in PLINK v1.9 (Purcell et al. 2007; Chang et al. 2015) and R statistical software (v3.5.1) (R core Team 2018) using R package “qqman” (Turner 2017). The association testing was conducted using the imputed data. However, an extra step was conducted to assess the HLA region with pre-imputed data to assess the accuracy of the imputation in this region.

2.3.7.1. Association testing

Genome-wide association testing was conducted using logistic regression assuming an additive model. This model compares the MAF between cases and controls to estimate the effect sizes for the association of the minor allele with the trait. The genetic association was adjusted for sex, age, smoking, and the first three principal components (PCs) as covariates. The results were then visualised using R-studio.

To ensure that results observed are true genetic association signals and not influenced by other factors, other than the disease itself, confounders were assessed. Confounders are variables in a study that could be different between cases and controls (other than the disease itself) and may contribute to some degree of inflation of variant association with disease. To minimise the potential bias, the first three PCs were added as covariates to the GWAS analyses. Genetic inflation of association was also assessed using quantile-quantile

(Q-Q) plots, which is a statistical test to assess the expected distribution (X-axis) of variant association between a genotyped SNP, against the observed values (Y-axis) of the GWAS results under the null hypothesis of no association, was plotted. Early deviation from the diagonal line, in conjunction with genomic inflation factor (lambda) above 1, indicates the influence of association by the genetic architecture of the SNP association. The possibility of genomic inflation was assessed by calculating the genomic inflation factor or lambda gc (λgc) in R, which augments the Q-Q plot. The lambda gc was calculated using the P-values of GWAS results, as the median of the resulting chi-square test divided by the median of chi-square distribution, all divided by 0.45499364 for 1 degree of freedom. A lambda gc of approximately 1 suggests that there is no indication of population stratification which may inflate the association. When genomic inflation is observed, PCAs are used to adjust for population structure. The PCs were calculated using independent SNPs in the dataset, which were obtained by pruning SNP in LD and extracting the independent SNPs.

2.3.7.2. HLA imputation evaluation

The HLA region has been shown to be hyperpolymorphic with genes in high LD, which introduces complexities in imputation. Moreover, the H3Africa SNP array itself has high enrichment of SNPs from this region. This extra step of association testing of this region with post QC imputation was done to assess the accuracy of imputation and to compare the results with the observed association with genotype data.

2.3.7.3. GWAS data visualization

Manhattan plot

The GWAS summary statistics were visualised using a Manhattan plot generated using the R package qqman (Turner 2017), which represents each SNPs position on the chromosomes on the X-axis and the negative log base 10 P-values of the association on the y-axis. Based on the number of SNPs genotyped and that passed QC, the genome-wide significance level was set at $P \leq 5 \times 10^{-08}$ and genome-wide significance suggestive level of an association at $P \leq 1 \times 10^{-05}$.

LocusZoom plot

Regional analyses for regions surrounding the lead SNP that reached the genome-wide significannce ($P \leq 5 \times 10^{-08}$) and suggestive levels ($P \leq 1 \times 10^{-07}$) were performed using

LocusZoom to generate single plots using the African population data set for PLINK summary statistics options (http://locuszoom.org/) (Pruim et al. 2010). LocusZoom uses the LD and gene annotation information from the publicly available KGP database to generate regional plots. The plots display the strength of SNP association, gene positions, and population-specific SNP LD calculations. In this study, the KGP African ancestry population LD structure of hg19 genome build was used.

2.3.8. GWAS catalog lookup

To assess replication of previously identified RA risk loci, SNPs published in the NHGRI-EBI GWAS catalog (Accessed 22 January 2019) (https://www.ebi.ac.uk/gwas/home) were compared with SNPs in this study. The GWAS catalog captures common genetic variants associated with common disease reaching a GWAS threshold of $P \leq 1 \times 10^{-06}$. Most of the RA risk loci captured in the GWAS were identified in participants with predominantly European and Asian ancestries. The RA associated SNPs were searched using the words "Rheumatoid arthritis". The non-redundant RA-associated SNPs from the GWAS catalog were then compared to GWAS summary statistics in our study. A nominal P-value of $P \leq 0.05$, which with Bonferroni correction becomes $P \leq 0.0004$ in the present study, was considered significant replication of the GWAS catalog associated SNPs.

2.3.9. Post-GWAS *in silico* functional characterization

The likely biological function of the associated SNPs from the genome-wide association study ($P \leq 5 \times 10^{-08}$) were assessed. The GWAS SNPs were first mapped to genes or proximity to genes using publicly available web-based databases such as the Ensembl Variant Effect Predictor (VEP) (http://grch37.ensembl.org/Tools/VEP) (McLaren et al. 2016). VEP was used to predict the effect of the variants in both coding and non-coding regions. VEP achieves this by utilising variant pathogenicity prediction tools for coding variants and non-coding variants, among others SIFT, PolyPhen, CADD, and LofTool with output scores based on the strength of functional evidence. VEP also provides output results for the different mRNA transcripts.

Concordance from two or more of the four prediction tools was accepted as good evidence for high-quality functional prediction. Since VEP does not report proximal genes for intergenic variants, wANNOVAR (http://wannovar.wglab.org/) (Wang et al. 2010) was used for this purpose. In addition, the non-coding variants were further assessed for regulatory activity

based on regulatory functional evidence on the expression of local genes mapped to the region. To achieve this, RegulomeDB (http://www.regulomedb.org/) (Boyle et al. 2012), an integrated regulatory database which incorporates data from the ENCODE project (CHIP-seq peaks, DNase I hypersensitivity peaks, DNase I footprints, transcription factors binding sites, biochemically validated promoter regions, chromatin state), Roadmap Epigenomics project, GEO data, eQTL data and published literature was used. It assigned categories for non-coding variant consequence based on the weight of experimental evidence for a regulatory role in those data sources.

RegulomeDB classifies variants into six different categories and subcategories based on evidence that the variant is likely to affect binding and is linked to the expression of the target gene, likely to affect transcription binding, less likely to affect transcription bind and variants with minimal transcription binding evidence. HaploReg v4 (http://compbio.mit.edu/HaploReg) (Ward and Kellis 2016), which used the same data sources and evolutionarily conserved genome sequences, was used to further annotate SNPs overlapping regulatory histone marks (promoter, enhancer, DNase I hypersensitivity).

Functional mapping and annotation (FUMA) (https://fuma.ctglab.nl/) (Watanabe et al. 2017), a web-based functional annotation tool was used to assess the eQTL function of a SNP and to map chromatin interactions of the GWAS significant SNP using the SNP2GENE option. Among many other functions, such as annotating eQTL using GTEx data (The GTEx consortium 2017), FUMA uses high-throughput chromatin conformational capture (HI-C) data to map chromatin by overlapping significant SNPs, and SNPs in LD with them, with the significantly interacting regions in the tissue cell types of interest. The SNPs are mapped to genes whose promoter regions (250 bp upstream and 500 bp downstream of transcriptional start site) overlap with another significant interaction (Watanabe et al. 2017). In cases where there was SNP functional prediction conflict between tools, the prediction made by two or more tools was accepted. All the tools mentioned above are compatible with the PLINK v1.9 data output format.

2.3.10. Power Calculation

Power was assessed using Quanto (http://biostats.usc.edu/Quanto.html). This case-control study is approximately 80% powered to detect associations of variants with an odds ratio

(OR) of 1.5 or above when the minor allele frequency is ≥ 0.05, in a study with 577 cases and 1612 controls after QC processes (Figure 2.2).

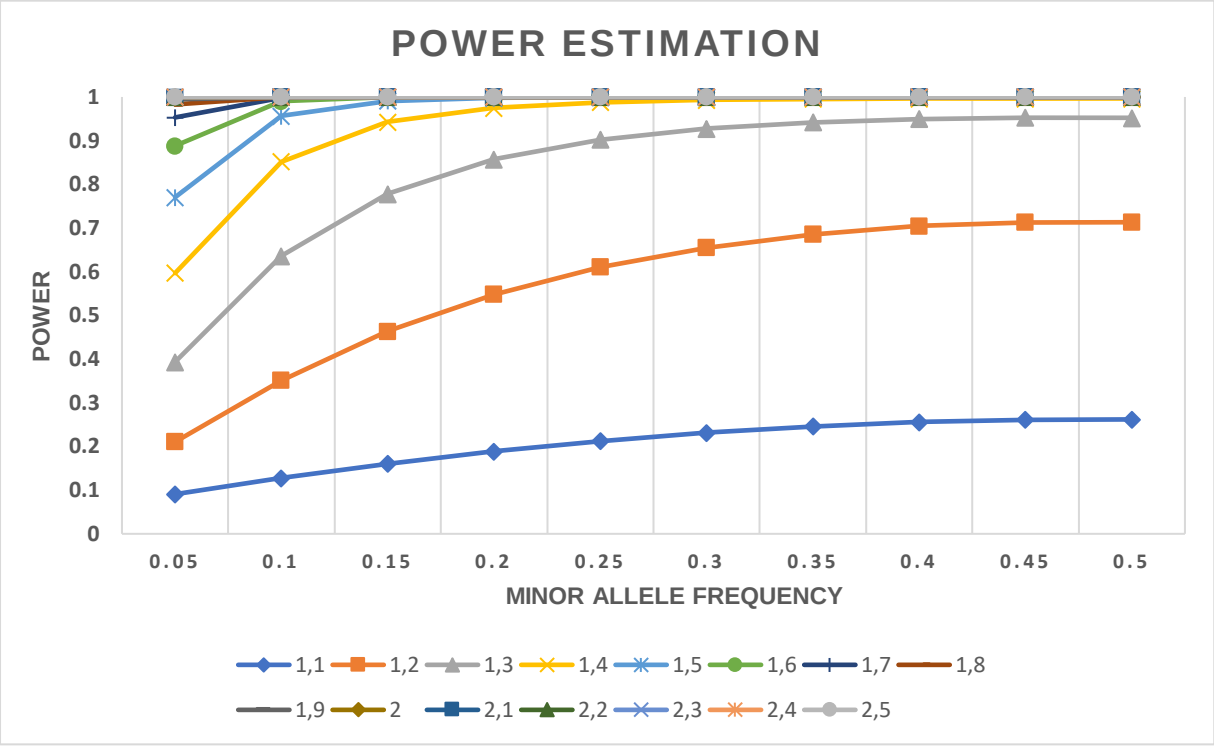


Figure 2.2: A Quanto based *a priori* estimate of statistical power to detect an association with common SNPs. The graph plots the power over minor allele frequency for odds ratios between 1.1 and 2.5 with 577 RA cases and 1612 controls. Each coloured line corresponds to a particular odds ratio. This study is approximately 80% powered to detect an association with an odds ratio from 1.5 and minor allele frequency ≥ 0.05.

2.4. Results

2.4.1. Patients demographics

A total of 2189 individuals (577 unrelated RA cases following attrition due to the QC process explained in section 2.3.3 below and 1612 population controls) were analysed in this study to evaluate the genetic associations and to identify variants that increase susceptibility to RA. The demographic and clinical characteristics of the 577 cases and 1612 control participants are shown in Table 2.1. There was a predominance of middle-aged females (86%) in comparison with males (14%) among the RA participants with an age range of between 27 and 86 years (mean 56 years). Furthermore, there were more RA affected non-smokers (79%) than smokers (19%).

Table 2.1: Demographic and clinical characteristic of RA cases and controls

Variable	Cases (n = 577)	Controls (n = 1612)
Sex, Female (%)	86	58
Age, Mean±SD (Years)	56±12.5	52±4.5
Smoking, Positive (%)	19	35.9
Disease duration, Mean±SD (Years)	4.0±2.4	-
ESR (mm/h), Mean±SD	35.6±25.4	-
CRP (mg/L), Mean±SD	26.1±36.7	-

n – number of individuals, ESR – erythrocyte sedimentation rate, CRP – C-reactive protein
SD – standard deviation

2.4.2. Population structure

The population structure was assessed using PCAs and comparing the distribution of the cases and controls. The results show that the cases and controls cluster together and are roughly evenly distributed indicating that there is relatively little population structure to bias the association analyses (Figure 2.3). However, to adjust for potential population structure indicated by the distribution clines of the individuals, the first three PCs were used to correct for potential confounding due to population sub-structure.

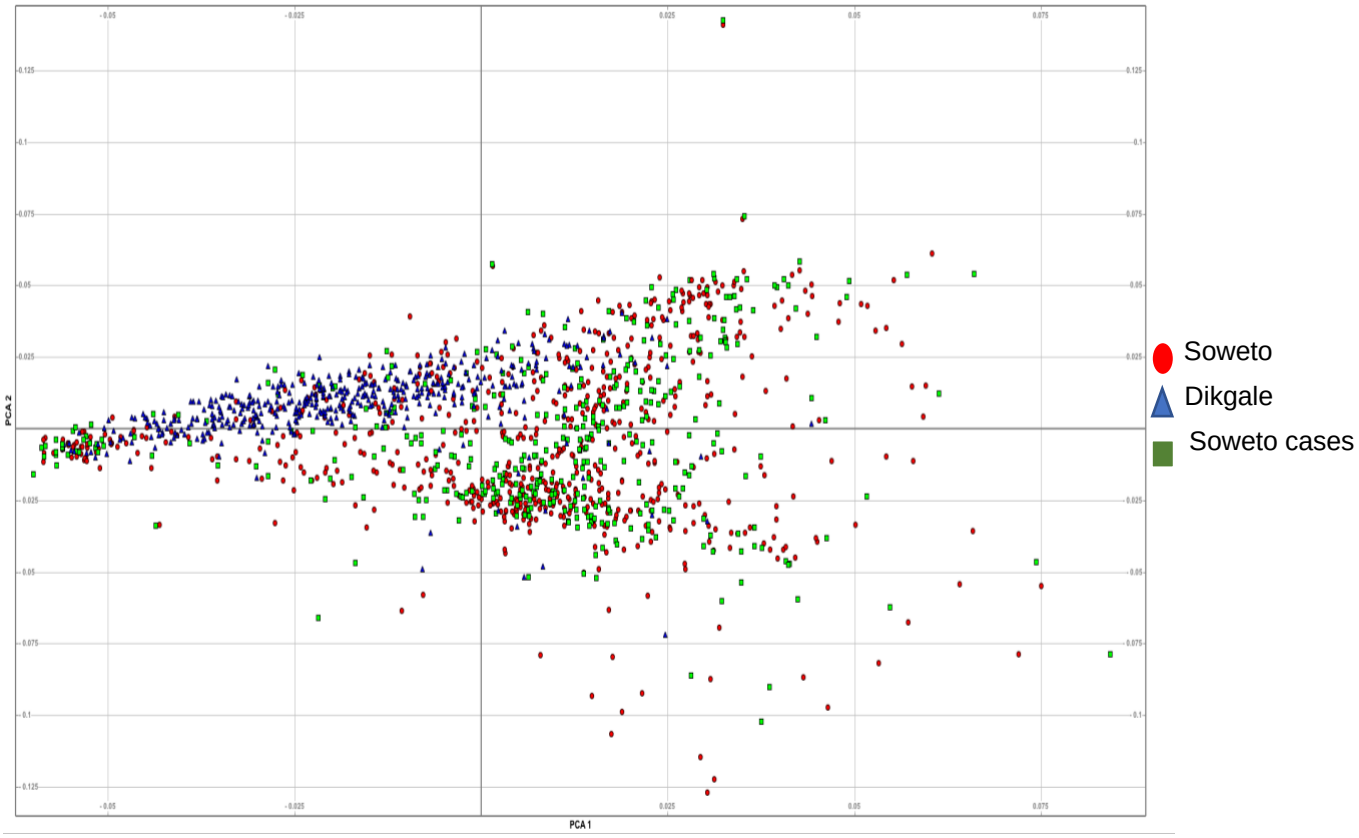


Figure 2.3: The principal components analysis (PCA) between cases (Soweto cases) and population controls (Soweto and Dikgale). Soweto cases are represented as green squares and boxes, red circles represent Soweto population controls and the blue triangles represent Dikgale population controls from the AWI-Gen study. PC 1 and PC 2 represent the fairly homogeneous distribution of cases and controls. PC1 captured 14.7% whilst PC2 captured 4.7% of the variation.

2.4.3. Sample and SNP QC

After stringent QC procedures on 2259 individuals and genotypes from 2,267,346 SNPs, a total of 577 RA cases, 1612 population controls (2189 individuals) and 1,405,229 autosomal SNPs passed QC and were available for analysis. A total of 29 samples failed the genotyping assay and were not available for analysis. The QC results revealed that 34 pairs of individuals (5%) had an IBD score of 1, which indicated that they were duplicates and one individual from each pair was removed based on the IBD cut-off of 0.2. In addition, 7 samples failed a sex check filter and sex could not be resolved and these samples were removed from the analysis. SNP genotyping QC filtered out 39,867 non-autosomal and 862,117 autosomal SNPs. Furthermore, 666,712 (77%) SNPs were rare variants and were removed with the MAF cut-off of less than 0.05, and of these 213,321 (32%) were monomorphic in the dataset.

A relatively small number of SNPs, 1945 (1%), were removed with HWE cut-off of $P \leq 10^{-06}$ (Figure 2.4).

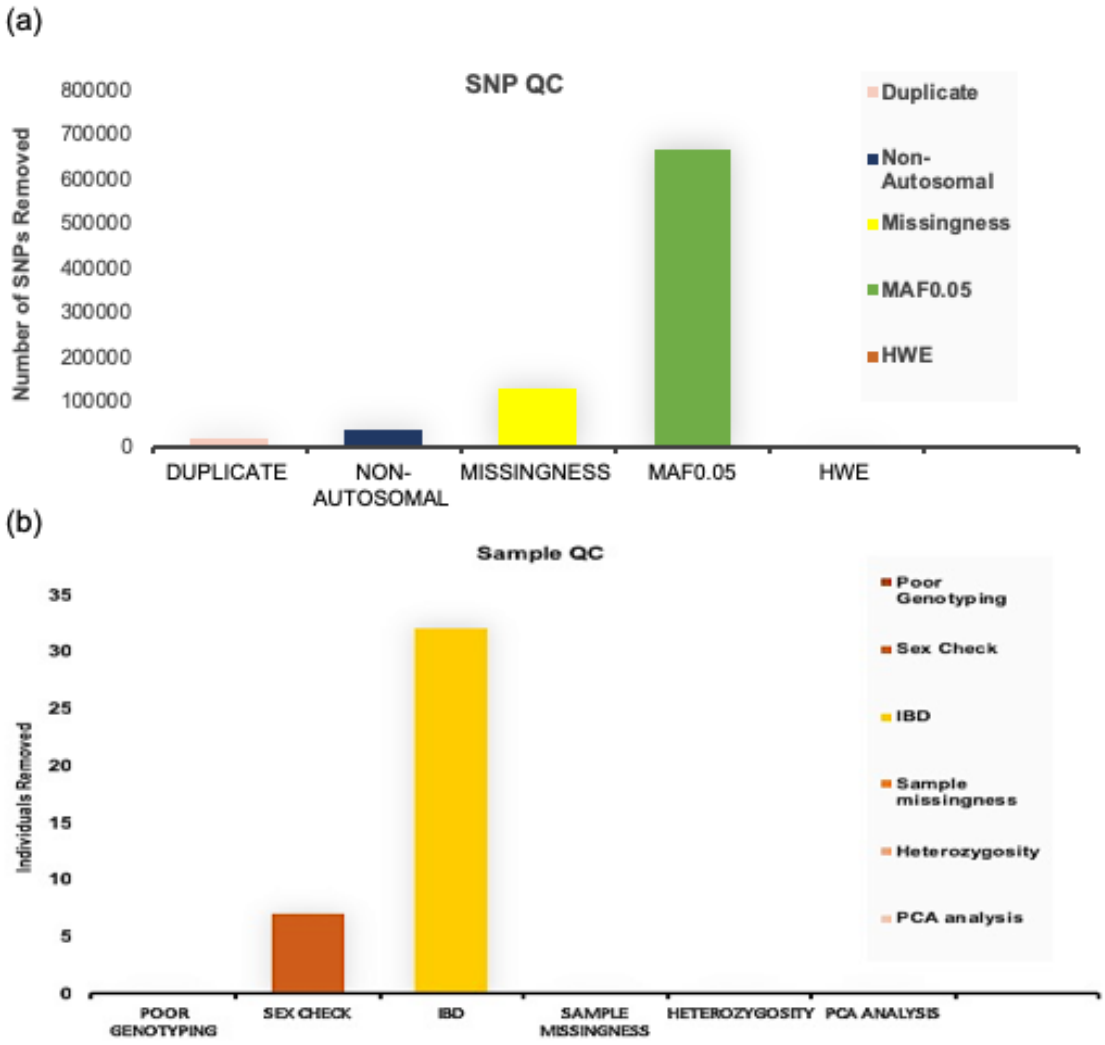


Figure 2.4: During the QC process, individuals SNPs were removed after performing all quality control steps represented in a bar graph. (a) Most SNPs were filtered out with the MAF cut off of 0.05 step (top bar graph). (b) The majority of the individuals were removed following IBD assessment which revealed that there were duplicates that had been assigned different unique identifiers because they were collected twice at different time points.

2.4.4. Genotype imputation

Imputation was performed and following post-imputation QC, 14,975,226 SNPs with an imputation quality score (r_2) of 0.8 at MAF cut-off of 0.01 remained. After filtering for a MAF cutoff of 0.05, a total of 7,986,485 SNPs were available for GWAS analyses.

2.4.5. Genome-wide association testing

To identify genetic variants associated with susceptibility to RA, a logistic regression model assuming an additive model was performed with the imputed data (7,989,485 SNPs). The analyses were adjusted for sex, age, smoking, and the first three PCs.

2.4.5.1. Quantile-Quantile Plots

Q-Q plots were generated using GWAS summary statistics to assess the possibility of genomic inflation due to population structure and other potentially confounding factors, indicated by a marked early rightward deviation from the diagonal line on the plot, assuming a null hypothesis of no association. The results show a marked deviation from the diagonal line (Figure 2.5a), however, after masking the highly associated extended HLA region (74,598 SNPs within range chr6:28,477,797 – 33,448,354) from the analysis, the deviation was no longer concerning (Figure 2.5b). This generated and estimated genomic inflation factor, lambda, of 1.01, which suggests that there was not confounding due to population structure.

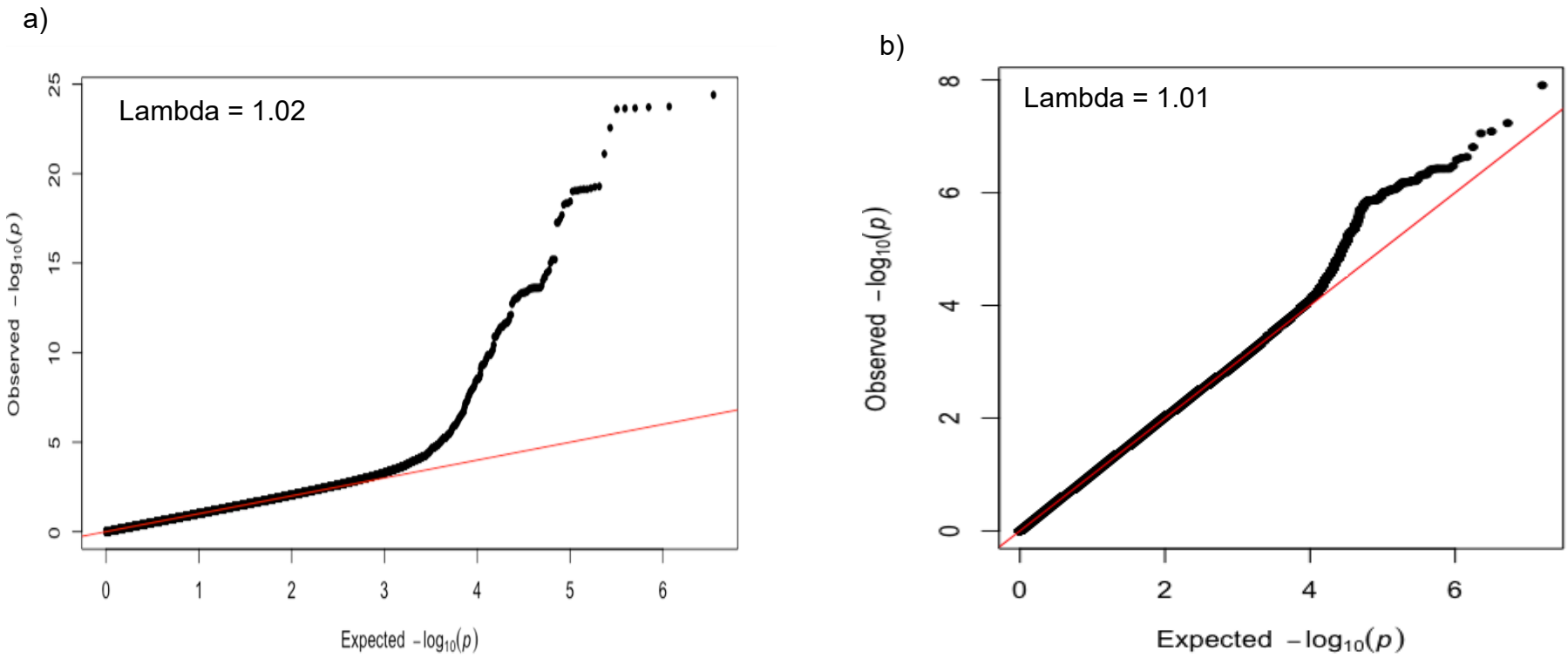


Figure 2.5: Quantile-quantile plots comparing the distribution of observed P-values to the expected -log10 P-values assuming a null hypothesis of no association. Every black dot represents a SNP. (a) represents the whole dataset and (b) represents the data after masking out data from the extended HLA region. Although the plot appears to be inflated in the graph (a), removing the highly associated and strongly linked extended HLA region shows that there is no significant genomic inflation.

2.4.5.2. Manhattan plot

To visualise the association test P-value of each of the SNP at the chromosome level, a Manhattan plot was generated using GWAS summary statistics following association analysis using the imputed data. Two peaks, one on chromosome 6 and the other on chromosome 11 reached the genome-wide significance threshold of $P \leq 5 \times 10^{-08}$ for association with increased risk of developing RA (Figure 2.6). In addition, although multiple SNPs showed trends towards association by reaching the suggestive threshold of significance of $P \leq 1 \times 10^{-05}$, only SNPs that showed a "tower" on chromosomes 2, 3, 4, 8 and 10 (Figure 2.6b) were selected for further analysis.

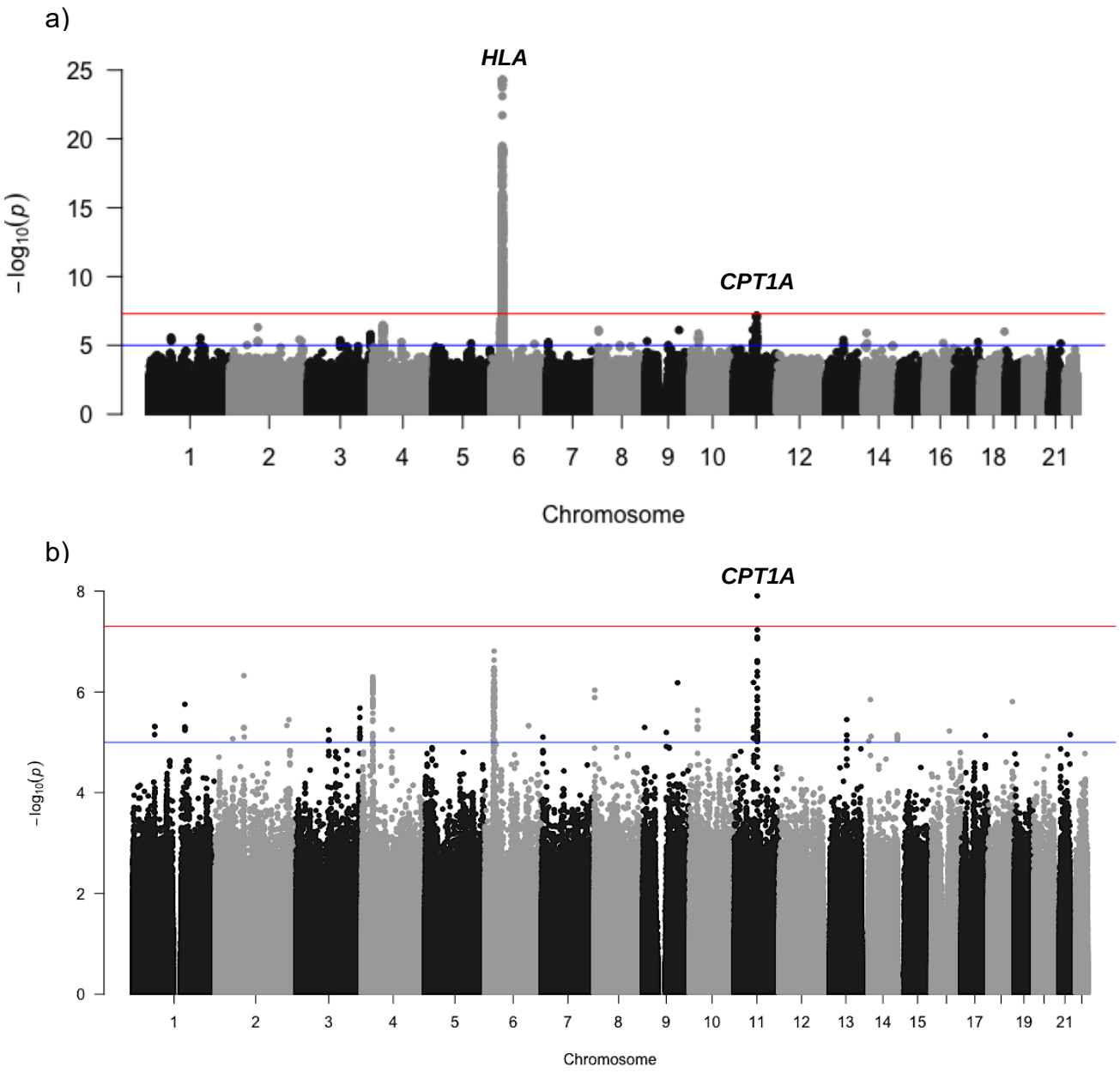


Figure 2.6: Manhattan plots showing the strength of association for each SNP represented as the -log10 of the P-value on the Y-axis and the chromosome position on the X-axis when RA cases were compared with controls using the imputed dataset. (a) represents the dataset with the extended HLA region on chromosome 6 and (b) the dataset without the extended HLA region data on chromosome 6. The blue line denotes the genome-wide suggestive association level ($P \leq 1 \times 10^{-05}$) and the red line denotes genome-wide association level ($P \leq 5 \times 10^{-08}$). The genes to which the SNP of the top association signals map are annotated.

2.4.6. Assessment of GWAS associated SNP

Two regions revealed SNP associations at genome-wide significance ($P \leq 5 \times 10^{-08}$) using the imputed data. They are the HLA region and the region around the *CTP1A* gene (a non-HLA

locus). The post-GWAS analysis for the significantly associated non-HLA locus was conducted using results from the imputed genotype data. The HLA region, however, was analysed twice by comparing the results for association using two different set of SNPs, one with just the genotyped SNP and the other with the imputed SNP. The reason is that there is a high number of polymorphic loci and strong LD between variants in the complex HLA region which can affect the imputation efficiency.

2.4.6.1. HLA region

A total of 240 SNPs from the chromosome 6 HLA region were associated with increased susceptibility to RA when using the pre-imputation genotyping data. The top 20 SNPs with the most significant P values are shown in Table 2.2. The SNP to gene annotation revealed that the strongest five potential novel SNP associations in the HLA region located to the intergenic region between *HLA-DRB1* and *HLA-DQA1* (rs617578, OR = 3.31, P = 3.99 X 10^{-25}, rs34855541, OR = 3.35, P = 1.79 X 10^{-24}, rs34261045, OR = 3.35, P = 1.94 X 10^{-24}, rs35265698, OR = 3.34, P = 2.47 X 10^{-24}, rs34350244, OR = 3.34, P = 2.21 X 10^{-24}) (Table 2.2). Analysis with the imputed genotyping data increased the number of GWAS significant SNP associations with a P-value less than the threshold of $P \leq 5 \times 10^{-08}$ from 240 to 966. However, it was observed that the top five most strongly associated SNPs from the pre-imputation association analysis along with another 36 associated SNPs (Appendix D) were lost during the imputation process and subsequent QC steps. These findings suggest that the reference panel could be sub-optimal for imputing SNPs in the HLA region in this population. This is being investigated in a parallel study and could be the result of poor imputation of low-frequency variants and complex LD structures. Therefore, for analysing SNP associations in the HLA region we reverted to using the pre-imputation dataset and will only report these results.

Table 2.2: Top 20 SNPs on chromosome 6 in the HLA region that reached genome-wide significance level ($P \leq 5 \times 10^{-08}$) association (sorted by P-value)

SNP	CHR	BP	A1	A2	MAF*		Genomic Region	Overlapping Gene	OR	P
					Case	Control				
rs617578	6	32574603	A	G	0.26	0.07	Intergenic	*HLA-DRB1* | *HLA-DQA1*	3.31	3.99E-25
rs34855541	6	32559825	G	A	0.25	0.07	Intergenic	*HLA-DRB1* | *HLA-DQA1*	3.35	1.79E-24
rs34291045	6	32560385	T	A	0.25	0.07	Intergenic	*HLA-DRB1* | *HLA-DQA1*	3.35	1.94E-24
rs35265698	6	32561334	G	C	0.25	0.07	Intergenic	*HLA-DRB1* | *HLA-DQA1*	3.34	2.47E-24
rs34350244	6	32561465	T	C	0.24	0.07	Intergenic	*HLA-DRB1* | *HLA-DQA1*	3.34	2.21E-24
rs34039593	6	32570311	G	T	0.25	0.07	Intergenic	*HLA-DRB1* | *HLA-DQA1*	3.23	2.29E-24
rs3998158	6	32681992	C	T	0.36	0.17	Intergenic	*HLA-DQB1* | *HLA-DQA2*	2.54	2.76E-23
rs602457	6	32573562	C	T	0.25	0.08	Intergenic	*HLA-DRB1* | *HLA-DQA1*	2.93	7.84E-22
rs1964995	6	32449411	C	T	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.16	5.15E-20
rs9391786	6	32448561	G	A	0.46	0.25	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.16	5.56E-20
rs9378264	6	32443451	A	G	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.16	6.59E-20
rs12195582	6	32444544	T	C	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.15	7.49E-20
rs9394099	6	32449160	T	G	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.15	7.49E-20
rs12194148	6	32444198	T	G	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.15	7.81E-20
rs28895244	6	32443820	A	G	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.15	8.81E-20
rs9378212	6	32445691	T	C	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.15	8.91E-20
rs13211921	6	32442836	G	T	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.15	9.79E-20
rs3104413	6	32582650	G	C	0.29	0.11	Intergenic	*HLA-DRB1* | *HLA-DQA1*	2.49	3.55E-19
rs1391371	6	32603798	T	A	0.29	0.11	Intergenic	*HLA-DRB1* | *HLA-DQA1*	2.49	4.51E-19
rs6931277	6	32583357	T	A	0.29	0.11	Intergenic	*HLA-DRB1* | *HLA-DQA1*	2.48	4.52E-19
rs9268839	6	32428772	G	A	0.46	0.26	Intergenic	*HLA-DRB9* | *HLA-DRB5*	2.12	5.38E-19

SNP- single nucleotide polymorphism, CHR – chromosome, BP – base position, OR – odds ratio, A1 – minor allele, A2 – major allele, MAF – minor allele frequency, P – P-value measure of statistical significance

*MAF calculates the frequency of the minor allele (A1) in both cases and controls.

HLA SNP Genotype cluster plots

Genotype cluster plots of the top four HLA region SNPs showing the strongest association indicated that the SNPs were well genotyped in the H3Africa array (Figure 2.7).

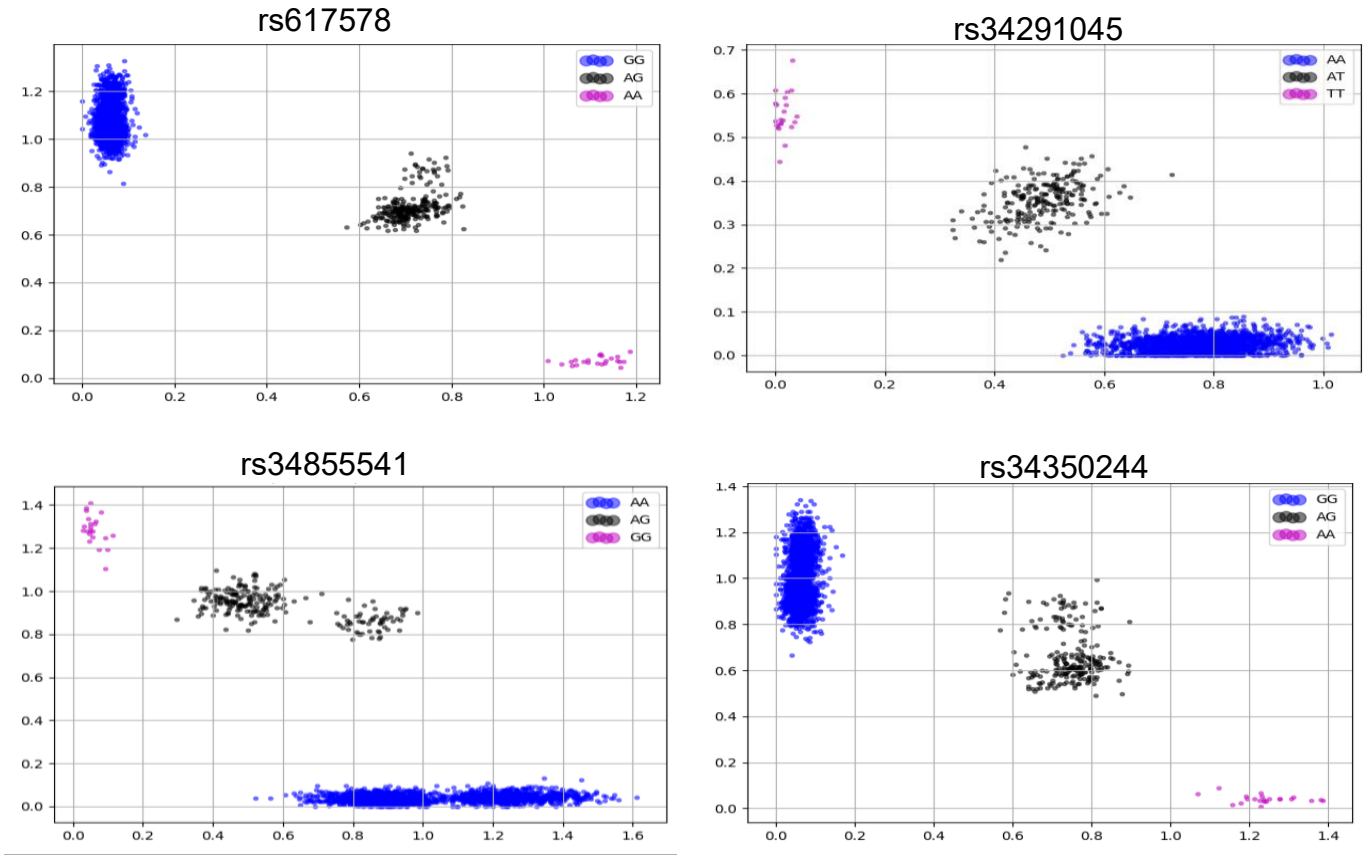


Figure 2.7: Cluster plots examples of the four top GWAS SNPs on chromosome 6 HLA region (rs617578, rs3429104, rs34350244, rs3485554). SNP genotypes cluster shows clear separation from each other.

HLA SNP Conditional Analysis

Conditional analysis was performed to identify independent associations in the HLA region. This was done using a genome-wide complex trait analysis (GCTA) software with an option for conditional and joint analysis (GCTA-COJO) (Yang et al. 2011). Briefly, GCTA-COJO uses GWAS summary statistics to perform stepwise conditional regression within a 10MB window and adjusts for LD between SNPs at a locus level (Yang et al. 2011). Using this process, two independently associated SNPs mapping to the intergenic region between *HLA-DRB1* and *HLA-DQA1* (rs617578) and an intergenic region between *HLA-DRB9* and *HLA-DRB5* (rs28377109), were identified to be potentially responsible for the signals in this chromosome 6 region. However, it needs to be noted that conditional analysis of the HLA region has its own limitations and therefore the results need to be investigated further.

HLA region LocusZoom

A LocusZoom plot of the associated HLA region was generated using the 1000 Genomes Project African population data for the LD (r_2) visualisation (Figure 2.8). The plot showed that the top index SNP (rs617578) is not in strong LD ($r_2 > 0.8$) with any other of the associated SNP within the HLA region. The lack of strong LD for this lead SNP lends support to the conditional analysis results which reported that the lead SNP represented an independent signal.

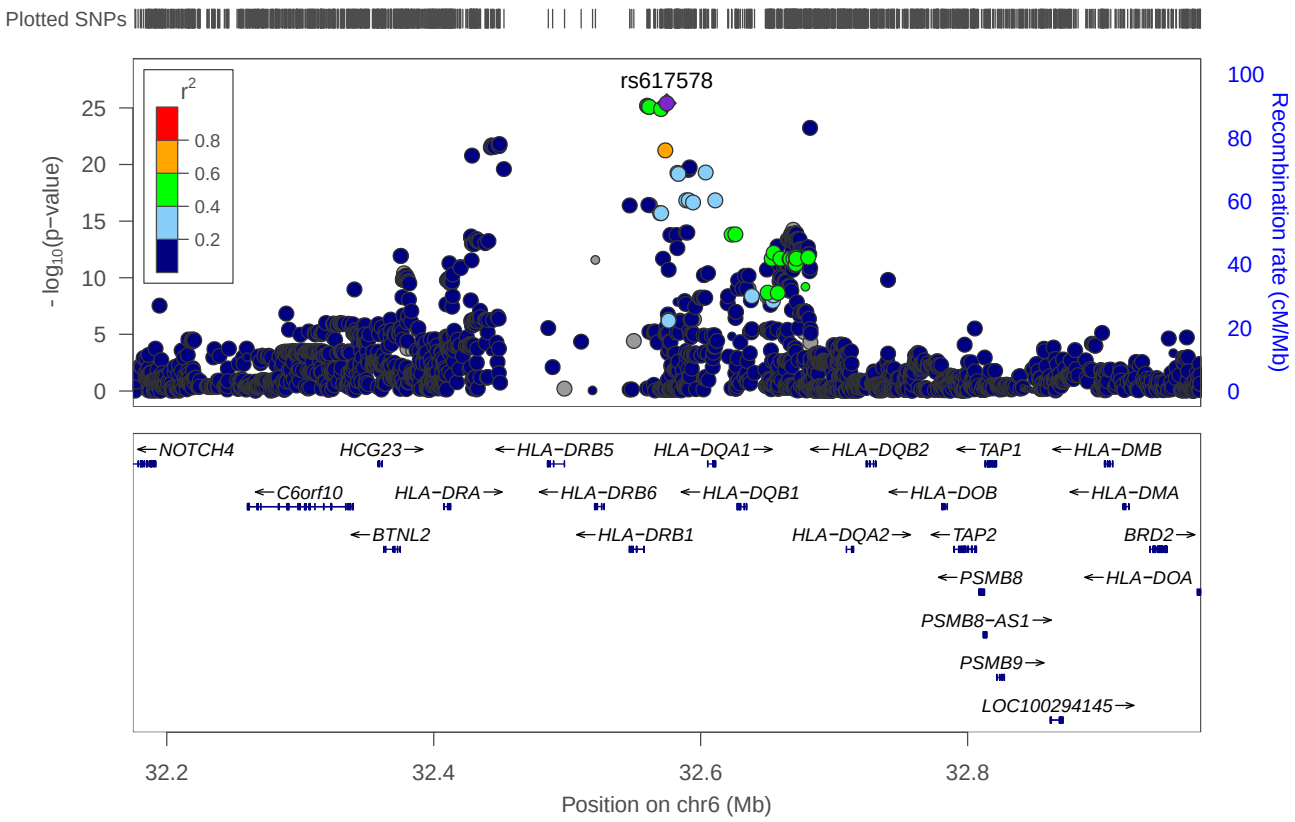


Figure 2.8: Regional plot of the chromosome six HLA region SNPs strongly associated with risk for developing RA, using the pre-imputation data. The plot shows the P values for the significantly associated SNPs and intergenic genes to which the lead SNP (rs617578) maps. The plot also shows that the lead SNP is not in strong LD with any other SNP in the African ancestry population in the KGP dataset, shown by the different colours, with red indicating complete LD (r_2>0.8) and dark blue indicating a very low degree of LD (r_2<0.2) in relation to the lead SNP.

2.4.6.2. Non-HLA locus, CPT1A, with genome-wide significant association

A set of four SNPs on chromosome 11 reached the GWAS statistical significance level of $P \leq 5 \times 10^{-08}$. The SNP to gene annotation revealed that the four associated SNPs map to intron 12 (rs57638434, OR = 1.70, $P = 7.11 \times 10^{-08}$) and intron 14 (rs3019593, OR = 1.68, $P = 7.26 \times 10^{-08}$, rs2123869, OR=1.61, $P = 7.79 \times 10^{-08}$, rs3019596, OR = 1.66, $P = 1.05 \times 10^{-08}$) of the *CPT1A* gene on the chromosome 11 (Table 2.3).

Table 2.3: Non-HLA SNPs that reached genome-wide significance at P ≤ 5 X 10^{-08}. The SNPs mapped to the intronic region of *CPT1A*. The SNPs genotyped on the H3Africa array are presented in bold

SNP	CHR	BP	A1	A2	MAF*		Genomic Location	Gene	OR	P
					Case	Control				
rs57638434	11	68546953	G	A	0.23	0.18	Intronic	*CPT1A*	1.70	7.1E-08
rs3019593	**11**	**68538939**	**C**	**T**	**0.24**	**0.18**	**Intronic**	***CPT1A***	**1.68**	**7.2E-08**
rs2123869	11	68538660	G	A	0.30	0.24	Intronic	*CPT1A*	1.61	7.7E-08
rs3019596	**11**	**68532388**	**G**	**A**	**0.25**	**0.19**	**Intronic**	***CPT1A***	**1.66**	**1.1E-08**

SNP- single nucleotide polymorphisms, CHR – chromosome, BP – base position, OR – odds ratio, A1 – minor allele, A2 – major allele, MAF – minor allele frequency, P – P-value measure of statistical significance.
*MAF calculates the frequency of the minor allele (A1) in both cases and controls

CPT1A Genotype cluster plots

The genotype clusters of two genome-wide significantly associated non-HLA SNPs (rs3019593, rs3019596) in the *CPT1A* gene showed that the genotypes were clearly separated (Figure 2.9).

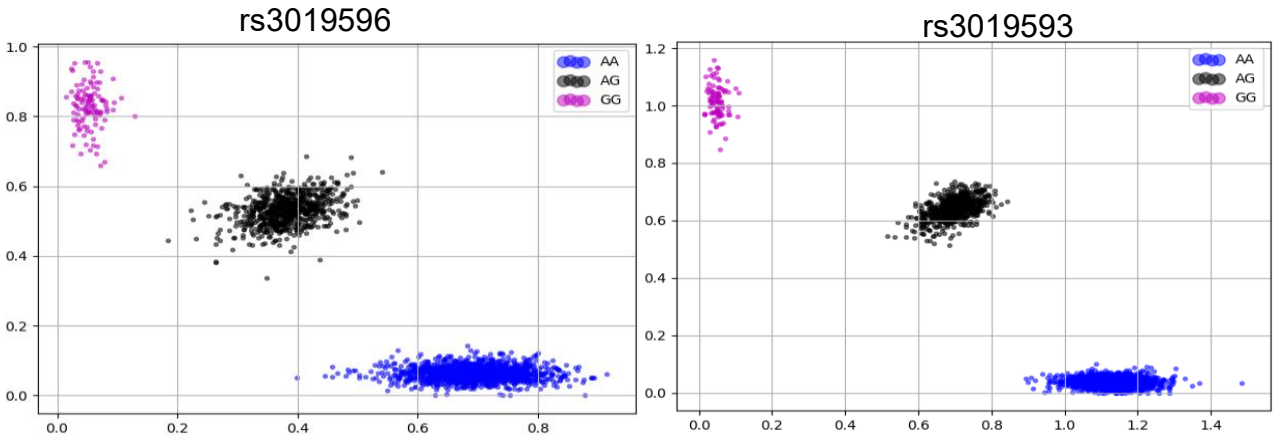


Figure 2.9: Cluster plots of the two GWAS SNP in CPT1A (rs3019593; rs3019596). SNP genotype clusters show clear separation from each other.

CPT1A associated SNP LocusZoom

The regional association was also assessed and a LocusZoom plot generated using the *CPT1A* gene index SNP (rs57638434) at the center of the plot (Figure 2.10). The plot shows that the lead SNP is in strong LD with several other SNP in the *CPT1A* locus. There are several other genes in close proximity of *CPT1A* including *GAL, MTL5*, *MRPL21,* and *IGHMPP2.*

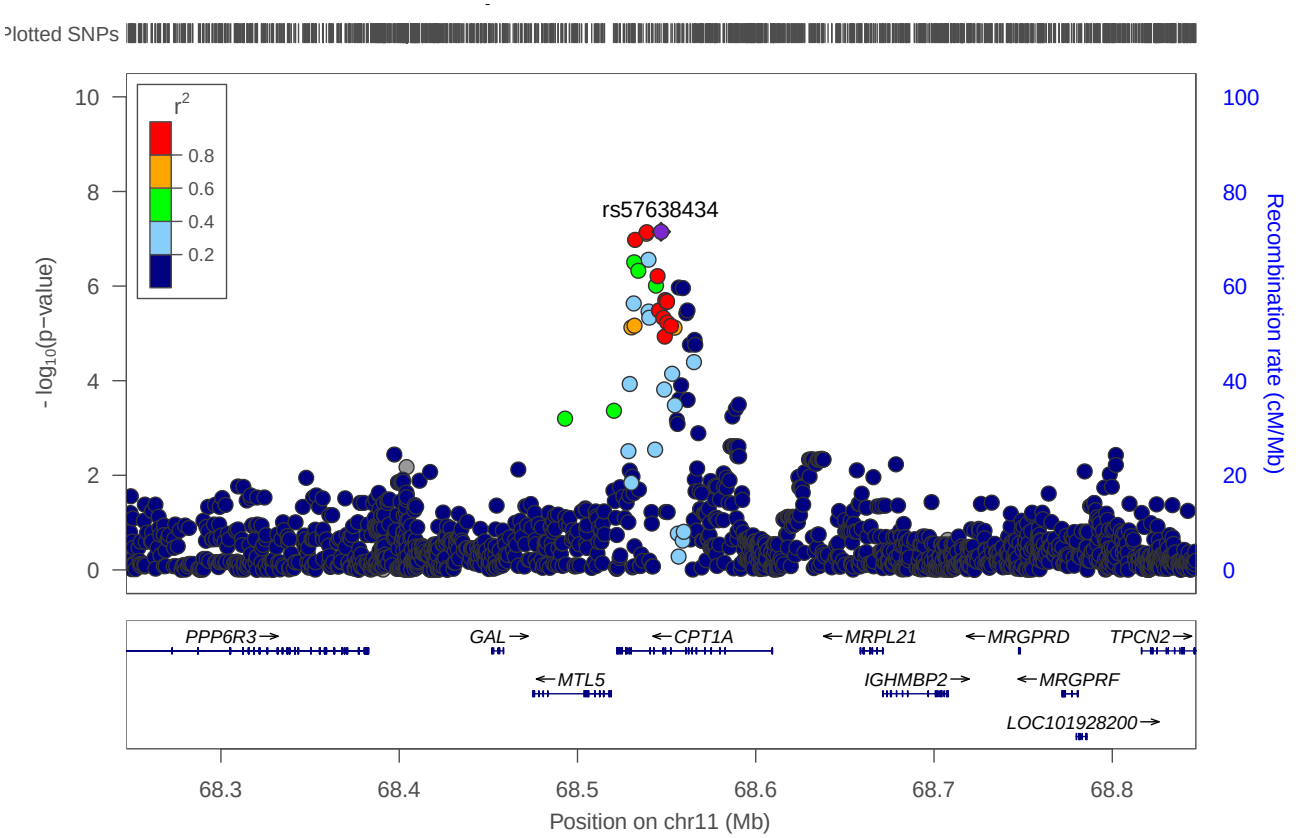


Figure 2.10: Regional/LocusZoom plot of the chromosome 11 region with SNP that reached the genome-wide significance of association with increased risk of RA (using the imputed data). The plot shows the P-values of the significantly associated SNP in *CPT1A* to which the lead SNP, rs57638434, maps. The plot also shows the level of LD between the lead/index SNP and other SNP, in the different colours, with red indicating near-complete LD (r_2>0.8) and dark blue indicating a very low degree of LD (r_2<0.2) in relation to the lead SNP.

2.4.7. Assessment of SNP at GWAS suggestive-level association

2.4.7.1. Non-HLA genome-wide suggestive associations

Five regions with a total of 737 SNPs showed suggestive association at a threshold of significance of P ≤ 1.0 X 10^{-05} (Figure 2.5b). Several SNPs that showed suggestive association mapped to the intergenic region between *LOC102723828* and *LOC101928622* on chromosome 4 (lead SNP rs115398984) **(Appendix E**). Other regions showing multiple SNPs associations include an intergenic region between *DCBLD2* and *MIR548G* on chromosome 3 (lead SNP rs111298380), intergenic region between *ST3GAL5-AS1* and *LOC90784* on chromosome 2 (lead SNP rs572443370) and an intergenic region between *LOC286083* and *DLGAP* on chromosome 8 (lead SNP rs1013012). Similarly, multiple SNPs (lead rs72826987) overlapping the intron region of *ABLIM1* on chromosome 10 and SNP

(lead rs7321945) in the intergenic region between *RPL12P34* and *KLHL1* showed suggestive associations (Appendix E). Future larger and better-powered studies from sub-Saharan Africa should examine the regions identified in this study further.

Non-HLA suggestive SNP LocusZoom

The LocusZoom plot of the chromosome 3 region showed that the SNPs that reached the suggestive threshold of association (Figure 2.11a) are in strong LD with the lead SNP (rs111298380) in the intergenic region between *DCBLD2* and *MIR548G.* In addition, there were numerous SNPs in strong and moderate LD with the lead SNP (rs115398984) in the dense chromosome 4 region (Figure 2.11b). Similar results are observed with the suggestive SNP on chromosome 10 (Figure 2.11d). Only one SNP is in strong LD with the lead SNP (rs111298380) on chromosome 3 (Figure 2.11c). Future larger and more powered studies should assess these regions for association with susceptibility to RA in sub-Saharan Africans.

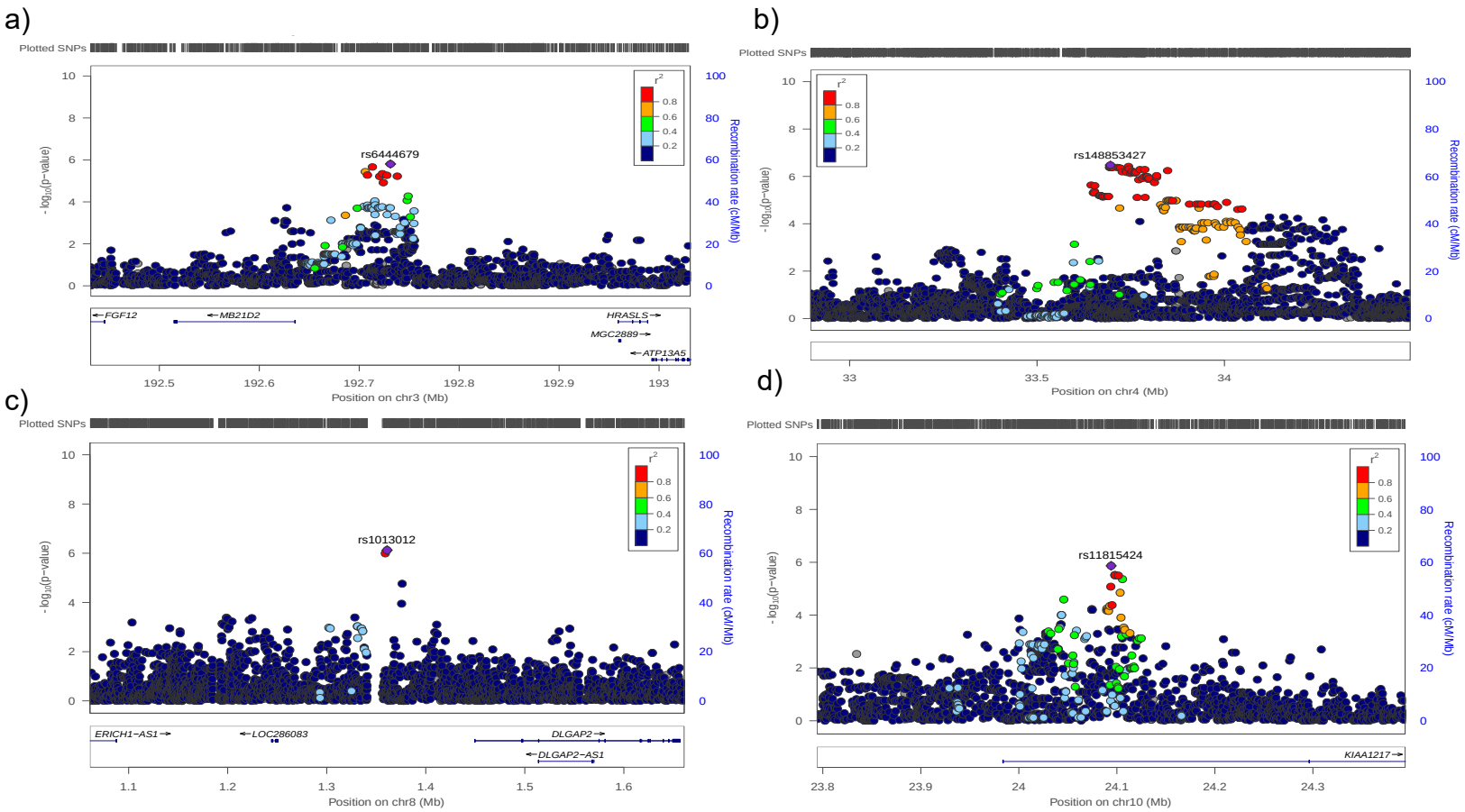


Figure 2.11: Regional plots of the SNPs that reached the GWAS suggestive threshold of P ≤ 1.0 X 10^{-05} for increased risk for RA, using the imputed genotyping data. The plot shows the strength of the suggestive associations and how the SNPs are positioned relative to genes to which the lead SNP map. (a) Several SNPs are in strong linkage with lead SNP in the suggestive regions on chromosome 3. (b) numerous SNPs are seen to be in strong with the lead SNP (rs115398984) mapping intergenic to LOC102723828 and LOC101928622 on chromosome 4. (c) Only one SNP was in strong LD with the lead SNP on chromosome 8. (d) Several SNPs are in strong LD with the lead SNP on chromosome 10. The plot was generated using the African ancestry population in the KGP for LD.

2.4.8. GWAS catalog lookup

The GWAS catalog (Accessed January 2019) contains a total of 685 non-redundant RA risk SNP. Checking for the presence of these SNPs in our imputed dataset revealed that only 227 SNPs were present at a MAF cutoff of 0.05. To look for regional replication, SNPs within a 50kb window were considered a single signal and the most significant SNP was selected and used. After this process, 109 RA SNPs from the GWAS catalog were considered independent signals and were available for replication confirmation.

To correct for multiple testing in the replication analysis, associated SNPs with a P-value of <0.0004 (Bonferroni correction - 0.05/109) were considered replicated and a P ≤ 0.05 was considered suggestive of replication. Following this process, 10 SNPs (discovered on the following groups: 4 in Eurasian, 3 in European, 2 in Asian, 1 in sub-Saharan Africa) were replicated in this study with varying strengths of association (Table 2.4). A SNP (rs12194148) intergenic between *HLA-DRB9* and *HLA-DRB5*, previously identified in a study of European ancestry participants at $P = 5 \times 10^{-58}$, was replicated at a significance level of $P = 1 \times 10^{-19}$. The remaining nine SNPs were replicated at a modest P ≤ 0.05, which is suggestive of replication. More specifically, two SNPs mapping to *MMELI* (rs2843401) and another mapping to the intergenic region between *GAPDHP64* and *NEFHP1,* previously identified in a European ancestry study with $P = 6 \times 10^{-06}$ and $P= 4 \times 10^{-08}$, were replicated at 0.001 and 0.005 respectively.

Two SNPs previously identified in an Asian ancestry study mapping to *ARHGEF3* (rs2062583, $P = 2 \times 10^{-06}$) and *BLK* (rs1600249, $P = 5 \times 10^{-06}$) were replicated at 0.012 and 0.018 in this study. Of those previously identified in Eurasian populations and mapping to *ANKRD55* (rs7731626, $P = 8 \times 10^{-23}$), intergenic between *B3GNT2*|*LOC101928724* (rs13385025)*,* intergenic between *AC104698.1*|*LBH* (rs10175798) and intergenic between *CDK12*|*NEUROD2* (rs1877030), were replicated at P = 0.003, 0.016, 0.017 and 0.05 (Table 2.4). Generally, higher SNP MAFs were observed in the GWAS catalog, which largely represents Eurasian populations. Taking into consideration the Bonferroni corrected replication, these results may suggest an absence of transferability of RA risk loci between Eurasian and sub-Saharan populations.

Table 2.4: GWAS catalogue lookup and replication analysis of RA-associated risk SNP

AWI-Gen								GWAS catalog				
CHR	**SNP**	**Gene**	**BP**	**A1**	**MAF**	**OR**	**P**	**Pop**	**MAF**	**OR**	**P**	**Author**
6	**rs12194148**	*HLA-DRB9|HLA-DRB5*	32444198	T	0.46	2.15	1.2E-19	E	N/A	N/A	5.00E-58	Padyukov, 2011
1	**rs798000**	*GAPDHP64|NEFHP1*	117280696	C	0.10	1.55	0.001	E	0.23	1.11	6.00E-06	Eyre, 2012
1	**rs2843401**	*MMEL1*	2528133	C	0.50	1.24	0.005	E	0.67	1.10	4.00E-08	Eyre, 2012
5	**rs7731626**	*ANKRD55*	55444683	A	0.18	0.74	0.003	EA	0.63	1.21	8.00E-23	Okada, 20143
3	**rs2062583**	*ARHGEF3*	56966246	G	0.39	1.23	0.012	A	N/A	1.59	2.00E-06	Freudenberg, 2011
2	**rs13385025**	*B3GNT2|LOC101928724*	62461120	A	0.11	1.36	0.016	EA	0.10	1.11	3.00E-06	Okada, 2014; Laufer, 2018
2	**rs10175798**	*AC104698.1|LBH*	30449594	A	0.26	1.25	0.017	EA	0.54	1.08	1.00E-09	Okada, 2014
8	**rs1600249**	*BLK*	11359638	T	0.11	1.38	0.018	A	N/A	1.30	5.00E-06	Freudenberg, 2011
11	**rs12421158**	PHRF1	609888	T	0.41	1.19	0.036	SS	0.34	1.72	4.00E-06	Govind, 2014
17	**rs1877030**	*CDK12|NEUROD2*	37740161	T	0.18	0.82	0.050	EA	0.10	1.09	1.00E-08	Laufer, 2018

SNP-single nucleotide polymorphism, CHR – chromosome, C1 – cases, C2 – controls, MAF-minor allele frequency for minor allele, A1 – minor allele, OR – odds ratios, P – P-value, RA – risk allele, Pop – population, EA – Eurasian, E – European, A- Asian, SS – Sub-Saharan African.

2.4.9. *In silico* functional analysis of significant SNP

2.4.9.1. HLA region associations

Of the 240 associated SNPs in the HLA region, 48 SNPs were predicted to have varying probability to potentially alter gene transcription based on the RegulomeDB score rating. More specifically, 35 of the 48 SNPs were scored at category 1, which is the highest probability of altering transcription factor binding and expression of a gene target (Table 2.5). Within category 1, 28 of 35 SNPs scored in the 1f category, which is a category for eQTL with either transcription factor binding or a DNase hypersensitivity peak. Four SNP in category 1 scored in the second highest 1b, and three SNPs scored 1d. Of the SNP that scored at category 1, most mapped to *HLA-DRA* gene (10 SNP), *HLA-DQB1*|*HLA-DQA2* intergenic region (8 SNP) and one SNP to *HLA-DQA1* (Table 2.5). Five SNPs scored in category 2 and eight scored in category 3.

Overall, of the 48 SNPs, more than half of the SNPs that mapped to a regulatory region were also an eQTL as predicted by HaploReg and GTEx (transformed fibroblasts and whole blood cells). More specifically, 24 SNPs mapped to a promoter region and of those 13 were also eQTL. Similar trends were observed for 41 SNPs that mapped to enhancers (28 eQTL) and 37 SNP DNase I hypersensitivity (23 eQTL). A total of 19 SNPs overlapped all three regulatory regions in whole blood tissue data and of those 10 also mapped to known eQTL. Furthermore, a missense SNP (rs7192) mapping to *HLA-DRA* gene which overlapped an enhancer and is an eQTL.

Table 2.5: The top 35 HLA SNP with high probability prediction to influence gene function

Gene	SNP	RegulomeDB	Regulatory Histone Marks			eQTL
			Promoter	Enhancer	DNase	
HLA-DQA1	rs9271589	1b	Yes	Yes	Yes	Yes
HLA-DRB9|*HLA-DRB5*	rs9378212	1b	Yes	No	Yes	Yes
HLA-DRB1|*HLA-DQA1*	rs9271588	1b	Yes	Yes	Yes	Yes
HLA-DQB1|*HLA-DQA2*	rs2647046	1b	Yes	Yes	Yes	Yes
HLA-DRB1| *HLA-DQA2*	rs9275224	1d	No	Yes	Yes	Yes
HLA-DRA	rs2213585	1d	No	Yes	Yes	Yes
HLA-DRB1	rs2516049	1d	No	Yes	Yes	No
HLA-DRB1|*HLA-DQA1*	rs9272218	1f	Yes	Yes	Yes	No
HLA-DRB1|*HLA-DQA1*	rs3104369	1f	Yes	Yes	Yes	No
HLA-DRB1|*HLA-DQA1*	rs9271586	1f	Yes	Yes	Yes	Yes
HLA-DRB1|*HLA-DQA1*	rs660895	1f	Yes	Yes	Yes	Yes
HLA-DRB9|*HLA-DRB5*	rs9391786	1f	No	Yes	Yes	No
HLA-DRB9	rs9268835	1f	Yes	Yes	No	Yes
HLA-DRB9	rs9268833	1f	Yes	Yes	No	Yes
HLA-DQB1	rs6905775	1f	No	Yes	Yes	Yes
HLA-DQB1	rs3828796	1f	Yes	Yes	Yes	No
HLA-DQB1|*HLA-DQA2*	rs7774434	1f	No	Yes	Yes	Yes
HLA-DQB1|*HLA-DQA2*	rs2647003	1f	No	No	No	Yes
HLA-DQB1|*HLA-DQA2*	rs4713581	1f	No	Yes	Yes	Yes
HLA-DQB1|*HLA-DQA2*	rs4713582	1f	No	Yes	Yes	Yes
HLA-DQB1|*HLA-DQA2*	rs4248168	1f	No	Yes	Yes	Yes
HLA-DQB1|*HLA-DQA2*	rs6457620	1f	No	No	No	Yes
HLA-DQB1|*HLA-DQA2*	rs2856692	1f	Yes	Yes	Yes	No
HLA-DRB1|*HLA-DQA2*	rs9275312	1f	No	No	No	Yes
HLA-DRA	rs9268659	1f	Yes	Yes	Yes	Yes
HLA-DRA	rs7192	1f	No	Yes	Yes	Yes
HLA-DRA	rs3763327	1f	No	No	No	Yes
HLA-DRA	rs7194	1f	No	Yes	Yes	Yes
HLA-DRA	rs2239803	1f	No	Yes	No	Yes
HLA-DRA	rs4935354	1f	No	Yes	Yes	Yes
HLA-DRA	rs7195	1f	No	Yes	No	Yes
HLA-DRA	rs2213586	1f	No	Yes	Yes	No
HLA-DRB9	rs9268832	1f	Yes	Yes	Yes	Yes
HLA-DRA	rs3129882	1f	Yes	Yes	No	Yes
HLA-DRB1	rs28366298	1f	No	Yes	No	No

2.4.9.2. CPT1A region associations

The functional annotation of the four genome-wide significant associated SNPs showed that the non-HLA SNPs were within an intron of *CPT1A*. The lead SNP (rs3019596) in intron 14 of *CPT1A* was in a consensus sequence that has strong evidence of binding to regulatory elements and overlapping an enhancer region. None of the four SNPs was predicted to be within a promoter region using HaploReg whole blood tissue-specific data. One SNP (rs3019593) overlapped a DNase I hypersensitivity region. The GTEx predicted that another SNP (rs2123869) in intron 14 of the *CPT1A* gene was an eQTL in whole blood tissue-specific data. All four SNP showed evidence functional impact as shown by a low LofTool score, however, CADD scores of less than 12.5 suggested that the SNP effects were not deleterious (Table 2.6).

Table 2.6: Functional prediction and annotation of genome-wide significantly associated SNP in the *CPT1A* gene. SNPs in bold were genotyped on the H3Africa array and the other two SNPs were imputed

Gene	SNP	RegulomeDB	Regulatory Histone Marks				LofTool	CADD
			Promoter	Enhancer	DNase	eQTL		
CPT1A	**rs3019596**	**1f**	**No**	**Yes**	**No**	**No**	**0.03**	**2.1**
CPT1A	rs57638434	5	No	No	No	No	0.03	4.4
CPT1A	**rs3019593**	**ND**	**No**	**No**	**Yes**	**Yes**	**0.03**	**0.8**
CPT1A	rs2123869	ND	No	No	no	Yes	0.03	2.9

SNP – single nucleotide polymorphism, DNase – DNase I hypersensitivity site, eQTL – expression quantitative loci, LofTool – Loss of Function, CADD – Combined Annotation Dependent Depletion

FUMA was used to confirm eQTL and map chromatin interactions of the GWAS significantly associated SNPs on *CPT1A.* Chromatin interaction was evaluated using Hi-C data of 14 tissue types, including whole blood, which contains long-range genome-wide chromatin interaction data. FUMA results showed and prioritised 12 genes the interact with the *CPT1A* region associated with susceptibility to RA either as eQTL or chromatin interaction, or both (Figure 2.12). Of the 12 genes, the risk locus (3019593) is a cis eQTL for five genes which include *CPT1A, IGHMBP2, MRPL21, TPCN2,* and *AP000807.1*. In addition, four genes (*MRGPRF, RP11-554A11-6, CCNDI,* and *SUV420H1*) showed evidence of chromatin interaction with the risk locus. Moreover, the risk locus interacts with three other genes, as both a cis-eQTL and chromatin interaction, to *MTL5, GAL,* and *PPP6R3.*

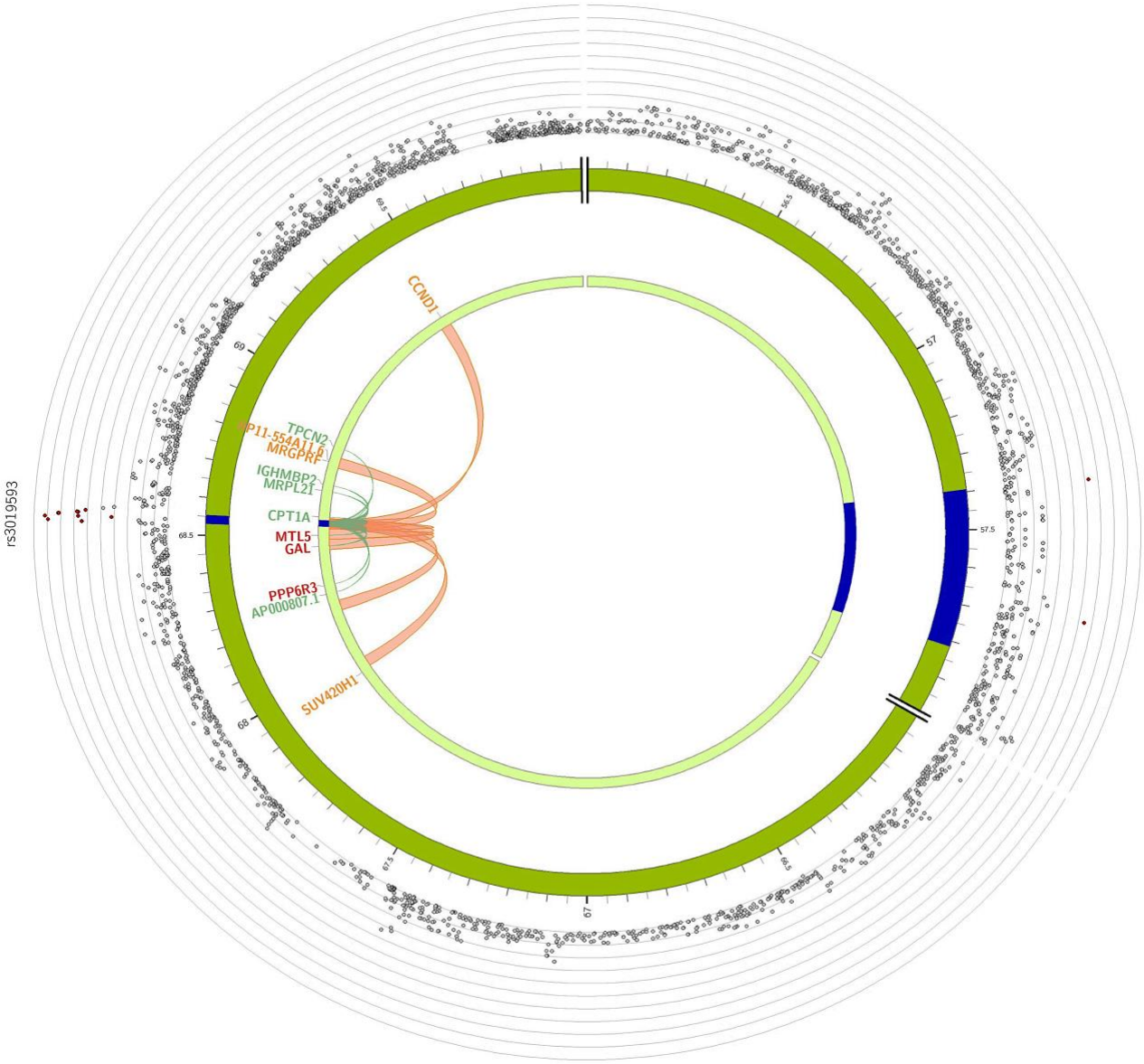


Figure 2.12: Chromatin interaction and eQTL status of the RA susceptibility locus, *CPT1A*. The outer layer represents the Manhattan plot displaying SNP from P ≤ 1 X 10^{-05}. Genome-wide significant SNPs are coloured based on the LD (r_2) to the top independent SNP (rs3019593) (red, r_2>0.8, and orange, r_2>0.6). The other SNPs are coloured in grey. Chromosome coordinates and genomic risk loci are highlighted in blue. Genes mapped by either Hi-C or eQTL are shown in the circle and coloured orange and green respectively. Genes mapped by both Hi-C and eQTL are coloured red. Chromatin interactions, eQTL or both are coloured orange, green and red respectively.

2.5. Discussion

Rheumatoid arthritis is a complex multifactorial disease whose pathogenesis remains poorly understood. Although GWAS has identified more than 100 SNPs associated with susceptibility to RA in European and Asian ancestry populations, there is a dearth of data on Africans. This GWAS study is the first to use an African-common-variant-enriched GWAS array to identify genetic variants associated with susceptibility to seropositive RA in black South Africans. Firstly, we identified a potentially novel association with genetic variation in the *CPT1A* gene. Secondly, we reaffirmed the well-established association of the HLA region with risk of seropositive RA. Thirdly, four intergenic risk loci between *LOC102723828* and *LOC101928622*, *DCBLD2* and *MIR548G*, *ST3GAL5-AS1* and *LOC90784* and *LOC286083* and *DLGAP* were found to have suggestive associations with susceptibility to RA. Finally, this GWAS did not show replication of any of the non-HLA SNPs previously identified in Eurasian ancestry populations (GWAS catalog). However, we observed some possibility for replication of nine of the previously identified signals, which might replicate even after consideration of multiple hypothesis testing corrections in a larger study. A more in-depth analysis that considers local replication might provide a better estimate of transferability of RA signals to Africans. Most of the SNPs identified were enriched in non-coding, intergenic and intronic regions, which may suggest pathways involving gene expression regulation as a possible molecular mechanism by which they contribute to disease pathogenesis in BSA. The functional prediction annotation of associations originating both in the HLA region and other regions revealed evidence of potentially influencing transcription factors binding to regulatory regions such as promoters, enhancers and DNase I hypersensitivity sites using relevant tissue-specific experimental data. Furthermore, some of the associated SNPs were predicted to be eQTL using data from whole blood and transformed fibroblasts, thus having the potential of influencing the tissue-specific gene expression.

2.5.1. Non-HLA novel association with RA susceptibility in BSA

This GWAS identified four SNPs in the carnitine palmitoyltransferase 1A (*CPT1A*) gene that were statistically significantly associated with increased risk of developing RA. These SNPs are located within introns 12 and 14 of *CPT1A*. *CPT1A* is a rate-limiting metabolic enzyme critical in the metabolization of fats into cell energy by attaching carnitine to long-chain fatty acids, making them permeable to the mitochondrial membrane where fatty acid oxidation (FAO) takes place (Roe and Ding 2001). The associated SNP (rs3019596) in intron 14, had

a variant allele that was functionally predicted to have a higher probability to modulate a B-cell enhancer region possessing the monomethylation of histone H3 lysine 4 of an active enhancer (H3K4me1_Enh) in chromatin marks. This supports a previous finding which has shown that H3K4me1_Enh was specific to identifying RA GWAS risk loci in genes from disease-relevant B- and T-cells (Walsh et al. 2016), suggesting a transcription factor binding modulation to target gene/s role for this SNP. This SNP is in almost complete LD with another SNP (rs3019593) in intron 14 of *CPT1A* and a SNP (rs5763844) in intron 12. Due to the lack of relevant functional prediction of rs5763844, it is assumed that the association of this SNP is due to being in LD with rs3019596 and/or rs3019593. Interestingly, rs3019593 overlaps a DNase I hypersensitive site and is predicted to be a *cis*-eQTL of *CPT1A* in whole blood (Westra et al. 2013), which suggests a gene expression regulatory role. Another *CPT1A* SNP in intron 14 is also predicted to be a cis-eQTL to *GAL*, which is a gene located about 50kb downstream to *CPT1A* and transcribed in the opposite direction from the other DNA strand. This finding suggests that intron 14 in *CPT1A* may have a gene regulatory role for its own expression and that of nearby genes. GWAS have shown that SNPs within 25kb of the *CPT1A* locus identified in this study were previously associated with RA related phenotypes. For example, SNP in intron 14 (rs186065533) and the 3' untranslated region (rs2278907) of *CPT1A* and a SNP intergenic to *CPT1A* and *TESMIN* were all associated with bone density (Madina-Gomez et al. 2018; Kichaev et al. 2019). In addition, a SNP (rs2924679) in intron 9 and a missense SNP (rs2229738) of *CPT1A* were previously associated with obesity and lipid metabolism (Comuzzie et al. 2012; Kettunen et al. 2012).

Using a similar population to the current study, an Immunochip study (~200,000 SNP enriched for regions previously associated to immune traits) reported seven non-HLA significant SNPs associations with RA in regions intergenic between *LOC389203*|*RBPJ, LOC100131131*|*IL1R1*, *KIAA1919*|*REV3L, LOC643749*|*TRAF3IP2* and the intron regions of *ICOS* and *KIAA1542* (Govind et al. 2014). These SNPs were not genome-wide significant in this current study. The difference may be due to methodological improvements used in the current study. Apart from adjusting for sex and age as in the Immunochip study, the current study also adjusted for smoking and the first three PCs which could also confound the association. Furthermore, all the four SNPs identified in this study are not typed on the Immunochip array, which demonstrates the utility of genome wide and imputation platforms for discovering associations.

The *CPT1A* association with increased risk of susceptibility to RA observed in this study appears to be African-ancestry specific as it is not reported in the GWAS catalog which reports predominantly European and Asian studies. Furthermore, two previous GWAS meta-analyses of RA with over 100,000 participants of Eurasian ancestries (Okada et al. 2014) and another with over 40,000 participants (Stahl et al. 2010) reported all four SNPs as not being associated with increased risk of developing RA (P-values around 0.3). In addition, two GWAS reported rs2123869 as not being associated with increased risk of RA using 1522 RA cases and 1850 controls North Americans and Swedish populations (P=0.048) (Plenge et al. 2007), and 400 cases and 400 controls Spanish population (Julia et al. 2008). The general trend in the previously mentioned GWAS show that all four SNPs have low effect sizes (OR =~1.1) for the four *CPT1A* SNP, which may suggest that the SNP occur in very low frequencies in those populations (MAF not reported). However, KGP reports the SNP MAF to be approximately 0.12 for Europeans and 0.09 for Asians. The higher MAF of 25 and effect size (OR = 1.7) may explain the association with increased risk of RA in our African ancestry cohort. This finding may highlight the inherent ethnic-specific heterogeneity.

Biological evidence from transcriptomics and qPCR studies have previously shown that *CPT1A* is differentially expressed in circulating CD4+ T-cells from 44 patients with systemic autoimmune rheumatic diseases (Kumar et al. 2018; Hudson et al. 2017; Spreafico et al. 2016), antigen-specific B-cells from RA and SLE patients (Abreu et al. 2019) and in a collagen-induced arthritis mouse model (Barbarroja et al. 2016). Furthermore, differentially expressed *CPT1A* has also been associated with increased autophagy of CD4+ T-cells, a critical process for the generation and maintenance of T-cells (Kumar et al. 2018; Puleston et al. 2014; Xu et al. 2014). The deregulation of autophagy in CD4+ T cells of RA patients was shown to result in the development of citrullinated peptides, development of autoantibodies, inflammation from T-cell hyperactivation and apoptosis resistance in a rodent model by suppressing the *FoxO* transcription factors necessary for cell cycle arrest (Zhu et al. 2017; van Loosdregt et al. 2016). More recently, a link between RA pathogenesis and deregulation of autophagy has been made (Vomero et al. 2018; Chen et al. 2018).

This study adds to this growing body of knowledge and offers insights into the genetics of RA pathogenesis. Taking into consideration the evidence above, we hypothesise that the identified risk locus in *CPT1A*, particularly the intron 14 variant, may play a regulatory role to target gene/s which may alter apoptosis as one of the possible molecular mechanisms of RA

pathogenesis in BSA. This hypothesis is supported by functional analysis of the independent SNP (rs3019593) association in intron 14 of *CPT1A*, and a DNA segment where variant alleles in predicted regulatory regions, as evidenced by chromatin interactions, are likely involved in the regulation of gene expression of three other nearby genes, *MTL5, GAL,* and *PPP6R3,* located within 200kb of *CPT1A*. *MTL5* and *GAL* have previously been shown to be upregulated in patients with osteoporosis. Furthermore, *MTL5* and *CPT1A* were two of three genes that were reported to be dysregulated in inflammatory diseases and in osteoporosis, another phenotype related to arthritis (Chen et al. 2018; Liu et al. 2017). Future studies should investigate replicating this association in larger RA cohorts of sub-Saharan ancestry.

2.5.2. MHC/HLA region

In this study, we demonstrated that SNPs that were genotyped on the H3Africa genotyping array were excluded from the imputed dataset, following post-imputation QC, leading to a loss of potentially important data. Several previous studies have also observed challenges with genotype imputation of the HLA region using established imputation methods (Pappas et al. 2018; Zheng et al. 2014; Dilthey et al. 2013). A common thread in those studies was that the low accuracy achieved with imputing the HLA region, specifically the *HLA-DRB1* locus, was attributed to a "SNP desert" and hyper-variation extending from *HLA-DRB3* to *HLA-DQB3* (Traherne et al. 2006, de Bakker et al. 2006). The high variation between haplotypes in these regions may have resulted in increased missing SNPs rates and in the systemic exclusion of *HLA-DRB1* SNPs from custom made arrays and reference panels (Pappas et al. 2018). We observed in our study that some of the SNPs, including the top five HLA associations using the H3Africa SNP array-based data, were lost doing the imputation process. These SNPs were either not on the reference panel or imputed at very low accuracy and quality and would, therefore, have been removed during the QC process. As the H3Africa SNP array has been specially enriched with SNPs for the HLA region (Mulder N personal communication), this would suggest that the Sanger African reference panel used for imputation in this study may be sub-optimal for this genomic region. We therefore only report on associations with HLA SNPs that were genotyped on the array. As mentioned before, an in-depth investigation of this phenomenon is being conducted by others.

Several previous studies have reported extensively on the HLA locus association with RA (Stastny 1978; Martell et al. 1989; Mody et al. 1989; Pile et al. 1992). Despite the technical

limitations outlined above, the findings from the current study confirmed the association of the HLA class II locus and its association with an increased risk of developing RA. Although the largest contribution to increased risk of developing seropositive RA has been largely attributed to *HLA-DRB1* alleles in most studied populations, including those with African ancestry participants (Viatte et al. 2015; Govind et al. 2019, Govind et al. 2014; Reynolds et al. 2014; Okada et al. 2014; Han et al. 2014; Raychaudhuri et al. 2012), divergent associations have also been reported in other populations. For example, a recent study has shown that *HLA-DQA1* locus conferred the strongest and major independent risk in seropositive RA Han-Chinese population after conditional analysis (Guo et al. 2018), thus highlighting the complexity of the region.

This study identified the strongest association with increased risk of RA with markers in the intergenic region between *HLA-DRB1* and *HLA-DQA1*. This locus (intergenic between *HLA-DRB1* and *HLA-DQA1*) has previously been reported by Govind et al (2014), in the first large genetic association study with the Immunochip array in RA using a subset of the participants in our study. However, the association dissipated when conditioned for the *HLA-DRB1* SNPs, thus suggesting that the association may be dependent on the *HLA-DRB1* locus. Our extended study shows that the region may be potentially independent of *HLA-DRB1* loci. This important discordance in outcome may be explained by an improvement in the conditional analysis methodology since the first study was published. Whereas the Immunochip study conditioned for independent signals by removing *HLA-DRB1* SNPs from analysis, which does not factor in the LD, this study used the GCTA-COJO algorithm, which conditions using the top SNP and takes into consideration the LD structure of the region in the study sample. Thus, the difference could indicate the possibility of multiple independent HLA region signals in BSA, as observed in other studies (Diaz-Gallo et al, 2018; Lee et al. 2008). It should, however, be noted that since the *HLA-DR* and *HLA-DQ* loci are in strong LD, it is difficult to completely rule out the possible influence of *HLA-DRB1* locus from the observed association (Raychaudhuri 2012). Thus, further larger studies should be carried out to confirm the independence of the associations in this study finding. In contrast to observations made in Caucasian and Asian populations, the *HLA-DP* and *HLA-B* loci were not statistically associated with RA in BSA (Furukawa et al. 2018; Jiang et al. 2018; Raychaudhuri et al. 2012; Okada et al. 2014).

Conditional analysis in the present study revealed two novel independent SNP associations, rs617578 in the intergenic region between *HLA-DRB1* and *HLA-DQA1*, and rs28377109 in the intergenic region between *HLA-DRB9* and *HLA-DRB5*. In GWAS, about 90% of associations are identified with markers in non-coding regions (Farh et al. 2015), and such variants are usually not the causal variant, but show association through LD with the causal variants. However, should the causal variant be in an intergenic region, it may have a role in the regulation of gene expression as a possible mechanism of action for RA pathogenesis. RA risk loci have previously been reported to be enriched in cell specific regulatory regions (Okada et al. 2014; Trynka et al. 2013). Indeed, the function prediction of rs617578 revealed that the SNP overlaps and probably modulates an enhancer region possessing histone H3 mono-methylation of lysine 4 (H3K4me1_Enh) signal, which was previously shown to be specific for identifying RA GWAS risk loci (Walsh et al. 2016). In addition, rs617578 overlaps a DNase I hypersensitivity region in whole blood data, which lends support to the hypothesis that RA associated SNPs are enriched in DNA regions that are open and active in CD4 T-cells (Okada et al. 2014; Trynka et al. 2013). Moreover, GTEx Consortium reports that rs617578, which is 17kb away from *HLA-DRB1*, may influence the expression target genes such as *HLA-DRB1*, *HLA-DQB1*, and *HLA-DQA2* in whole blood and *HLA-DQA2* in lung and thyroid cells, all of which have been shown to be relevant to RA pathogenesis (Lim and Kim 2019; The GTEx Consortium 2017). Therefore, this current study finding may offer further insight into the potential role of the intergenic region between *HLA-DRB1* and *HLA-DQA1* in the HLA risk, and may play a role in the pathogenesis of RA. However, the molecular basis of association and potential mechanisms of action remain poorly understood.

2.5.3. GWAS lookup of previous associations

This GWAS also replicated 10 non-HLA SNPs associations previously identified largely in studies with European and Asian ancestry populations. SNPs identified in Eurasian populations and replicated in this study generally had low MAF in the African study and statistical significance of association at $P \leq 0.05$. The most striking SNP association replication was observed with rs12194148 intergenic to *HLA-DRB9*|*HLA-DRB5,* previously reported in European ancestry at $P = 5 \times 10^{-58}$ and replicating in our study at $P = 1 \times 10^{-19}$. The nine other SNPs replicated at a modest $P \leq 0.05$, suggesting that SNPs has potential to replicate in larger cohorts. These initial findings demonstrate very limited transferability of the

Eurasian based RA risk loci in Africans and warrants a more in-depth assessment of transferability.

2.5.4. Strengths and Limitations

The strengths and limitations of our study are noted. The main strength of the study is that it is the first GWAS in the sub-Saharan region to investigate the genetic association of RA, thus offering ethnic-specific SNP-associations with RA susceptibility. Although this study is based on a modest number of BSA RA patients, this is currently the biggest GWAS of RA in the sub-Saharan region. Furthermore, coupled with the use of an African-common-variant-enriched GWAS SNP genotyping array and a specific African reference panel has much higher potential to detect African ancestry specific associations. This study complemented GWAS results using *in silico* functional annotation, which has the potential to assist in selecting and prioritising identified risk loci.

The primary limitation is the inability to detect rare and modest effect size RA-associated SNPs within our study. However, despite the modest sample size, this study identified an important novel locus association with ample supporting biological functional data for a role in the pathogenesis of RA. In addition, the cases and some of the controls, particularly from Dikgale, which is in a different province in South Africa, might have different environmental factors influencing disease association that may have influenced the analysis outcomes. The controls were population-based and were not screened for the absence of RA prior to commencement of the study, thus some of them may have RA, which could affect the results. The sex distribution was not optimal; however, this was adjusted for in the association testing. Finally, unavailability of suitable data from other African cohorts did not allow for replication of the novel signal. Meta-analysis using summary statistics from published African-American/non-African based studies might provide initial validation of the signal from this region.

In conclusion, this study identified a novel association with four markers in two intronic regions of the *CPT1A* gene, with increased risk of RA. Taking into consideration the previously reported gene expression findings highlighting a potential role for *CPT1A* in autoimmune diseases, this study offers insights into the genetic variants in *CPT1A,* particularly in intron 14, that may contribute to RA pathogenesis in BSA. Furthermore, this

study confirmed the association of the HLA locus with increased risk of RA and identified a potentially novel independent association in the intergenic region of *HLA-DRB1* and *HLA-DQA1* in people of sub-Saharan African ancestry. Future work is expected to validate these findings (probably in a larger cohort) and also contribute in narrowing down the causal variants, thereby providing greater insights into the biology of RA in African ancestry.

CHAPTER 3: PHARMACOGENETICS OF METHOTREXATE RESPONSE IN BLACK SOUTH AFRICAN RHEUMATOID ARTHRITIS PATIENTS ON METHOTREXATE MONOTHERAPY

3.1. Abstract

Methotrexate (MTX), an antifolate disease-modifying antirheumatic drug, is often used to treat RA. Previous studies have shown that variations in MTX polyglutamation genes play a role in MTX response in RA patients. However, very little is known about the disease aetiopathogenesis of RA and pharmacogenetics of MTX use in people of African ancestry. This study investigated 247 Black South African seropositive RA patients on MTX monotherapy to evaluate the association of 36 previously identified index single nucleotide polymorphisms (SNPs) spanning 19 genes and 7812 surrounding SNPs with MTX response. Patients were treatment naïve at baseline and were categorised after 6 months of MTX treatment into two groups of either EULAR good/moderate responders (n=186) or poor responders (n=61). A logistic regression analysis with an additive model adjusting for sex, age, smoking and disease duration was used for association testing. The regional replication of previously reported MTX response loci identified a region on chromosome 9 with 14 significantly associated SNPs mapping to intronic region of *CDK9* gene, of which four (rs1002095; rs2297214; rs3217740, rs3217751) are expression quantitative loci (eQTL) that potentially influence the expression of the *FPGS* gene involved in MTX polyglutamation and intracellular persistence of MTX polyglutamation. The regulation of the expression of the *FPGS* gene by eQTL variants mapping to the *CDK9* gene may influence response to MTX therapy in black South Africans with RA. Although our sample size had limited power to test many of the previously identified RA genetic associations with confidence, our results indicate that the pharmacokinetics of MTX might have a different genetic underpinning in African populations and underlines the necessity of larger studies in populations from Africa.

3.2. Introduction

RA is a complex autoimmune disease affecting approximately 1% of the adult population worldwide (Silman and Pearson 2002). Rheumatoid arthritis is characterised by chronic synovial inflammation of joints, primarily the small joints of the hands and feet, which leads to degradation of cartilage and bone. Left untreated, RA can lead to loss of function of the affected joint and disability (Firestein 2003). Currently, there is no preventive treatment or a cure for RA. Methotrexate (MTX) and DMARDs are currently used to treat the inflammation observed in RA (Morgan et al. 1997). Low dose MTX remain the anchor drug and is commonly used as a first-line drug for the treatment of RA (Cronstein 2005). Although successful, interpatient variability from MTX treatment, which leads to treatment dropout rates approaching 51% has been observed in a South African cohort (Hodkinson et al. 2012). Furthermore, AE such as hepatotoxicity and bone marrow suppression do occur and often lead to discontinuation of treatment in up to 35% of patients (Gilani et al. 2012).

The exact MTX mechanism of action remains elusive. It is, however, widely accepted that it functions by suppressing the immune response. Possible mechanisms of action including (1) adenosine signaling, (2) generation of reactive oxygen species, (3) decrease in adhesion molecules, (4) alteration of cytokine profiles, (5) polyamine inhibition and (6) folate antagonism. The folate antagonism is the most studied pathway and involves inhibition of dihydrofolate reductase (*DHFR*) by folylpolyglutamate synthase *(FPGS),* which is required for tetrahydrofolate production (Cronstein 1997). This inhibition of tetrahydrofolate leads to the reduction of purine and pyrimidine production by thymidylate synthetase (*TYMS)* and consequently DNA synthesis, which arrests the cell cycle at the S phase and causes apoptosis in primary T-cells (Visentine et al. 2012; Budzik et al 2000). This inhibition consequently results in the reduction of ATP and GTP, while the accumulation of adenylate and adenosine occurs leads to the systemic reduction of synovitis and anti-inflammatory effects (Kaltsonoudis et al. 2012; Wessels et al. 2006; Edno et al. 1996). The polyglutamation process is reversed by gamma-glutamyl hydrolase (*GGH*) enzyme, which removes glutamates from MTX. Intracellular retention of polyglutamated MTX (MTX-PG), which is the active compound, appears to be crucial for MTX anti-inflammatory activity (Yamamoto et al. 2016). A systemic review and meta-analysis study reported the role of genetic variants in genes that code for the folate pathway enzymes as determinants of treatment outcome in RA patients (Qiu et al. 2017). More recently, *FPGS* variants have been shown to influence

intracellular MTX-PGs retention and consequent response (Yamamoto et a*l*. 2016; Moya et al. 2016) and should be added in response prediction models (Muller et al. 2018).

The effective pharmacological activity of MTX depends largely on suppressing the replication rate of immune cells. Thus, it is plausible that the expression of genes involved in the cell cycle could be crucial in MTX treatment, eliciting its anti-inflammatory effect. The rate of apoptosis in inflamed RA joints is slower than immune cell proliferation and has been shown to reduce MTX efficacy (Raza et al. 2006; Cutolo et al. 2001). The cyclin-dependent kinases (CDKs) are critical to the cell cycle as they regulate progression through the four main cell replication phases (G1, S, G2, M) (Bose et al. 2013). More specifically, *CDK9* expressed on the surface of immune cells regulates the lifespan and apoptosis of immune cells (Wang et al. 2012; Sedlacek 2001). Recently, inhibition of *CDK*7 and *CDK*9 has been explored in the treatment of inflammatory arthritis (Hong et al. 2018; Hellvard *et al*. 2016). Therefore, it is likely that genetic variants that influence the dysregulation of genes involved in apoptosis may consequently influence anti-inflammatory treatment response.

Although the genetic association of SNPs with MTX efficacy has been described largely in European and Asian populations, there is currently a paucity of data for African ancestry populations. This study aims to evaluate previously identified MTX metabolism-associated genetic variants with MTX response in BSA.

3.3. Materials and Methods

3.3.1. Study participants

A total of 247 patients were recruited from the Arthritis Clinic at Chris Hani Baragwanath Academic Hospital (CHBAH), Soweto, South Africa. All patients fulfilled the 2010 American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) classification criteria for RA (Aletaha et al. 2010), were DMARD naïve, ACPA positive and had high disease activity at baseline. Patients were consenting, unrelated BSA and were at least 18 years of age at disease onset. They were considered 'black' if they self-reported all four grandparents as being BSA. Furthermore, all patients were treated with MTX monotherapy as the first-line treatment drug for at least six months and were treated with a similar treat to target protocol. Patients were started on a minimum MTX dose of 10-15 mg/week and this was increased to a maximum of 25 mg/week according to the disease activity, efficiency, and toxicity, at the discretion of the treating physician during follow up visits. Patients were categorised using the EULAR response criteria (Appendix F) into 2 groups, responders (good responders (GR), moderate responders (MR)) and non-responders (NR). The EULAR classification criteria uses the Disease Activity score (DA28) (Fransen and van Riel 2005), which is a score determined by the numerical sum of five outcome parameters 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 at six months. Patients had baseline data collected before treatment commencement and were followed up for six months, at which time their response was assessed. The patients received a medical history interview, physical examination, and routine laboratory testing at every visit as deemed fit by the consulting physician. Exclusion criteria included patients who had already been started on other DMARD therapy at the time of enrolment (e.g. Leflunomide). All participants in this study gave informed written consent and the study was approved by the Human Research Ethics Committee (Medical), University of Witwatersrand (HREC protocol number M1706109).

3.3.2. Patient data collection and extraction

Demographic, clinical patient and treatment data were retrieved from the international observational rheumatology registry (METEOR) (http://www.meteorfoundation.com/) managed at CHBAH. The collected data included sex, age, smoking history, clinical measurement such as swollen 28-joint count (SJC28), tender 28-joint count (TJC28), C-

reactive protein (CRP in mg/L) and erythrocyte sedimentation rate (ESR in mm). These measurements were used to calculate the disease activity scores (DAS28), a composite score which considers the SJC28, TJC28, ESR or CRP, and patient global health assessment. These measurements were collected at baseline just before the patient started MTX monotherapy and again just after 6 months follow up. The DAS28 is used to evaluate disease progression and response to therapy.

3.3.3. SNP selection

SNPs were chosen and evaluated for association with MTX response by comparing the minor allele frequencies between responders (GR combined with MR to form R) and non-responders (NR). The SNP associations with MTX response were evaluated in two parts. Firstly, the exact SNP from previous association studies was assessed and secondly, SNPs within 25kb in both directions of the index SNP were tested for association. A total of 52 SNPs previously associated with MTX response (in European and Asian ancestry studies) were retrieved from the pharmGKB (https://www.pharmgkb.org/) database. This database collects, curates and disseminates information about genetic variants that influence drug response (Whirl-Carrillo et al. 2012). The SNPs were extracted using query terms “RA” and “Methotrexate”.

3.3.4. Genotyping and imputation

Genomic DNA from study participants was extracted and genotyped at the Illumina® FastTrack™ Microarray services (Illumina, San Diego, USA), using the Human Heredity and Health Africa (H3Africa) Consortium genotyping array, which contains approximately 2.3 million SNP enriched for common African variants and mapped to the Genome Reference Consortium Human build 37 (GRCh37) (The H3Africa Consortium 2014). The H3A array data were enhanced by imputation using the Sanger African reference Imputation panel with and imputation score of above 0.8. The H3Africa SNP array data is hosted and managed by the University of Witwatersrand on behalf of the AWI-Gen group (Ramsay et al. 2016) (further information provided in Chapter 2). Quality control procedures were performed on the data using the H3ABioNet/H3Agwas pipeline (https://github.com/h3abionet/h3agwas). Briefly, this pipeline performs individual and SNP QC process. With regard to individual QC, they were removed if they had a genotyping failure rate with more than 2% missingness, discordant sex by comparing the X chromosome homozygosity rate (>0.8 for males and <0.2 for females)

and were closely related as measured by an “identity by descent” (IBD) cut-off of >0.2. SNPs were removed if they had a missingness rate above 2%, Hardy-Weinberg equilibrium P-value $<10^{-06}$; and minor allele frequency of less than 0.05.

3.3.5. SNP extraction in H3A SNP array data

The SNPs extracted from pharmGKB were checked if they were present in the H3A SNP array data. For local/regional replication, SNPs within 25kb in both directions of previously associated each index SNP were extracted from the H3A imputed data.

3.3.6. Statistical analysis

3.3.6.1. Genotype and allele frequency assessment

The differences in the genotype and allele frequency of the 36 SNPs overlapping 19 genes were compared between four groups containing GR, MR, NR, and R. Comparison was conducted between four pairwise group: (1) GR (n = 106) and MR (n = 80), (2) GR and NR (n = 61), (3) MR and NR and (4) R (n = 146) and NR. The difference in the frequencies between groups were evaluated using chi-square test ($\chi 2$) or Fisher’ exact test for cell counts less than 5, with 95% confidence interval in Plink v1.9 (Chang et al. 2015; Purcell et al. 2007).

3.3.6.2. Genetic association testing

The association of 36 index SNPs with MTX response (as measured by DAS28-CRP at 6 months) was performed using logistic regression analysis assuming an additive model. The covariates used in the association were sex, age, smoking and disease duration. The association was reported as an odds ratio (OR) with 95% confidence interval (CI). The testing model was also used to find associations of 7821 SNPs (occurring in +/- 25 kb of the 36 SNPs) with MTX response was analysed between the R and NR in a logistic regression with an additive model. To address multiple-hypothesis testing, a P-value of 0.05, after the 1000 permutation test, was considered significant. All statistical analyses were performed using Plink v1.9 (Chang et al. 2015; Purcell et al. 2007). A LocusZoom plot (http://locuszoom.org/) was generated to assess the regional association for each of the non-overlapping regions.

3.3.7. *In silico* function prediction annotation

The likely biological function of statistically associated SNPs was predicted using publicly and freely available web-based genomic annotation tools such as Ensembl Variant Effect Predictor (VEP) (GRCh37) (http://grch37.ensembl.org/Tools/VEP) (McLaren et al. 2016). VEP predicts the gene feature functional effect of variants in both coding and non-coding regions. VEP achieves this by incorporating variant pathogenicity prediction tools, among others, SIFT, PolyPhen, CADD, and LofTool, which provide output scores based on the strength of functional evidence. VEP also provides output results for the different mRNA transcripts of a specific gene. Concordance from two or more of the four prediction tools was accepted as likely to have a biological functional impact. Apart from CADD scores, non-coding variants were further assessed for regulatory activity based on functional evidence of their role in the expression of genes mapped in cis. To achieve this, RegulomeDB (http://www.regulomedb.org/) (Boyle et al. 2012), an integrated regulatory database which incorporates data from the ENCODE project (CHIP-seq peaks, DNase I hypersensitivity peaks, DNase I footprints, transcription factors binding sites, biochemically validated promoter regions, chromatin state), Roadmap Epigenomics project, GEO data, eQTL data and published literature, were used to categorise non-coding variants. The functional impact was based on how likely the variant is to alter a transcription factor binding site based on the weight of experimental evidence for a regulatory role. RegulomeDB classifies variants into one of six different categories and subcategories, where category 1 has strong regulatory influence confidence compare to category 6. Within category 1, 1a has more confidence than 1f. HaploReg v4 (http://compbio.mit.edu/HaploReg) (Ward and Kellis 2012; 2016), which uses the same data sources with the addition of evolutionarily conserved genome sequences, was used to further annotate SNPs overlapping to regulatory histone marks (promoter, enhancer, DNase I hypersensitivity).

3.4. Results

3.4.1. Demographic and clinical characteristics of study participants

In total, 247 RA patients with baseline and 6 months MTX response data were included in the study. Patient demographic and clinical characteristics are presented in Table 3.1. This study included more women (n = 201; 80%) and 72% were non-smokers with an age range between 27 to 86 years (SD+-12.1), at the start of MTX monotherapy. After 6 months on MTX, 75% (n = 186) of all cases showed a response to MTX therapy (43% good response (n = 106) and 32% moderate response (n = 80)), while 25% (n = 61) showed no response (non-responders) according to the EULAR criteria.

Table 3.1: Demographic and clinical characteristics of the study participants

Parameter		Responders (n=186)	Non-responders (n=61)
Age at diagnosis - Mean±SD (Years)		53.97 ±11.3	59.73±13.4
Sex (Male/Female)		32/154	9/52
Smoking (Ever/Never)		42/144	15/46
Disease duration - Mean±SD (Years)		2.9±1.6	3.7±1.7
ESR (mm/h)	Baseline	33.5±24.1	39.0±28.9
	Six months	25.4±20.8	26.7±20.0
CRP (mg/L)	Baseline	24.0±29.7	28.9±46.7
	Six months	12.7±14.9	13.9±15.7
DAS28(4)-CRP	Baseline	5.22±1.3	5.83±1.4
	Six months	3.49±1.4	3.85±1.5

ESR – erythrocyte sedimentation rate, CRP – C-reactive protein, DAS28(4)-CRP – Disease activity score.

3.4.2. SNP extraction in H3Africa array data

A total of 40 out of the 52 index SNPs were found to be present in H3A imputed data. However, four SNPs were found to be below the MAF cut-off of 0.05 for common SNP and were removed from further analysis. Therefore 36 common SNPs were available for exact association testing for association with MTX response. For regional association, 7812 SNP within 25kb of index SNPs were extracted and tested for MTX response.

3.4.3. Genotype and allele frequencies

The genotype and allele frequencies of the 36 previously identified SNPs were compared in the four groups. This was done to assess whether there was a statistically significant difference that could indicate that a genotype or allele was likely to influence MTX response in BSA. The results showed that six SNPs in five genes had significant difference in allele or genotype frequencies when compared between groups and these results are described further below. A summary of the result is presented in Table 3.2. and a complete table with all the results in Appendix (Appendix G).

Table 3. 2: Summary showing Allele and genotype frequencies between pair groups of good responders (GR), moderate responders (MR), nonresponders (NR) and responders (R)

Gene	Geno	GR (n=106)	MR (n=80)	P	GR (n=106)	NR (n=61)	P	MR (n=80)	NR (n=61)	P	R (n=186)	NR (n=61)	P
ABCB1 rs1045642 (A/G)	A	26	26	0.293	26	6	**0.033**	26	6	**0.004**	52	6	**0.006**
	G	186	134		186	116		134	116		320	116	
	AA	1	2		1	1		2	1		3	1	
	AG	24	22	0.513	24	4	**0.010**	22	4	**0.002**	46	4	**0.004**
	GG	81	56		81	56		56	56		137	56	
ADORA3 rs3394 (T/C)	T	53	49	0.242	53	21	0.103	49	21	**0.012**	102	21	**0.029**
	C	159	111		159	101		111	101		270	101	
	TT	7	8		7	2		8	2		15	2	
	TC	39	33	0.512	39	17	0.305	33	17	**0.036**	72	17	0.094
	CC	60	39		60	42		39	42		99	42	
MTHFD1 rs2236225 (A/G)	A	45	40	0.454	45	30	0.498	40	30	1.000	85	30	0.712
	G	167	120		167	92		120	92		287	92	
	AA	3	9		3	6		9	6		12	6	
	AG	39	22	**0.048**	39	18	0.139	22	18	0.937	61	18	0.597
	GG	64	49		64	37		49	37		113	37	
SLC19A1 rs1051266 (C/T)	C	52	56	**0.029**	52	39	0.161	56	39	0.613	108	39	0.569
	T	160	104		160	83		104	83		264	83	
	CC	6	9		6	7		9	7		15	7	
	CT	40	38	0.084	40	25	0.298	38	25	0.717	78	25	0.703
	TT	60	33		60	29		33	29		93	29	
ATIC rs7563206 (C/T)	C	83	77	0.091	83	64	**0.022**	77	64	0.548	160	64	0.075
	T	129	83		129	58		83	58		212	58	
	CC	19	16		19	17		16	17		35	17	
	CT	45	45	0.065	45	30	0.066	45	30	0.543	90	30	0.202
	TT	42	19		42	14		19	14		61	14	

One SNP in the *ABCB1* gene (rs1045642) had a significantly different allele frequency when GR was compared with NR (P = 0.03) (Figure 3.1b), GR and NR (P = 0.003) (Figure 3.1c), and between R and NR (P = 0.005) (Figure 3.1d). Similarly, genotype distribution was significantly different in those groups with P-values of 0.009, 0.001 and 0.003 respectively (Figure 3.1).

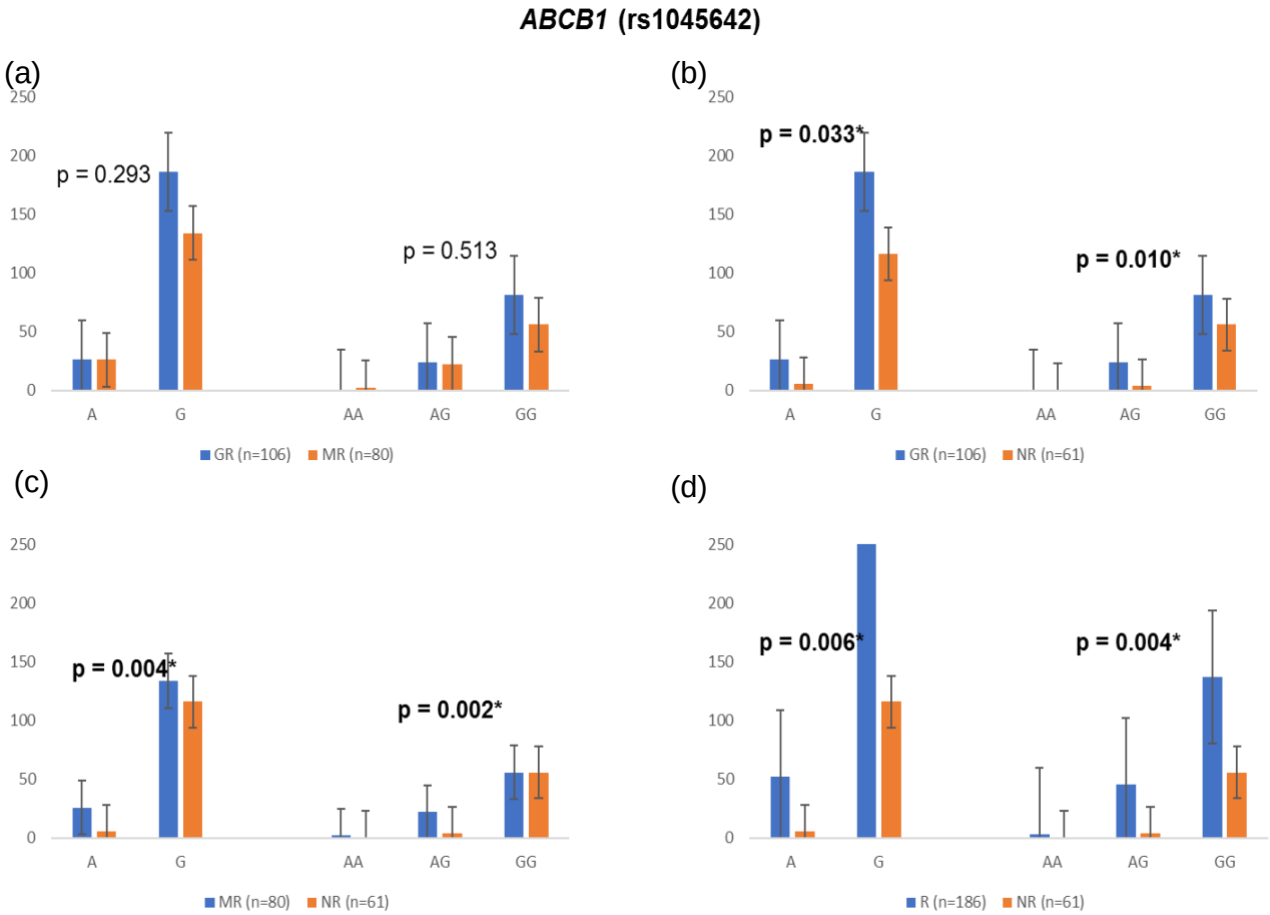


Figure 3.1: The genotype and allele frequencies comparisons of a SNP in the *ABCB1* gene, rs1045642, previously associated with MTX response. The allele and genotype frequencies were not significantly different between (a) GR and MR, but significantly different when b) GR was compared with NR, c) MR and NR and d) R compared with NR at P-value < 0.05. The standard errors were also calculated and are represented by the thin black lines on the vertical bars. The number of individuals in each group is represented by n. The allele and genotype are represented on the X-axis and the frequency is represented on the Y-axis. The comparisons showing statistically significant allele and genotype frequency differences are represented by bold numbers and an asterisk.

Another SNP (rs3394) in the *ADORA3* gene had significant different allele frequency when MR was compared with NR (P = 0.01) (Figure 3.2c) and when R was compared to NR (P = 0.03) (Figure 3.2d). No statistically different allele frequency was observed when GR was

compared with MR (Figure 3.2a) and GR compared with NR (Figure 3.2b). When the genotypes were compared, only the genotype frequencies between MR and NR were significantly different (P = 0.04) (Figure 3.2b). There was no significant difference when the genotype frequency between GR and MR were compared (Figure 3.2a), GR and NR (Figure 3.2b), and R and NR (Figure 3.2d)

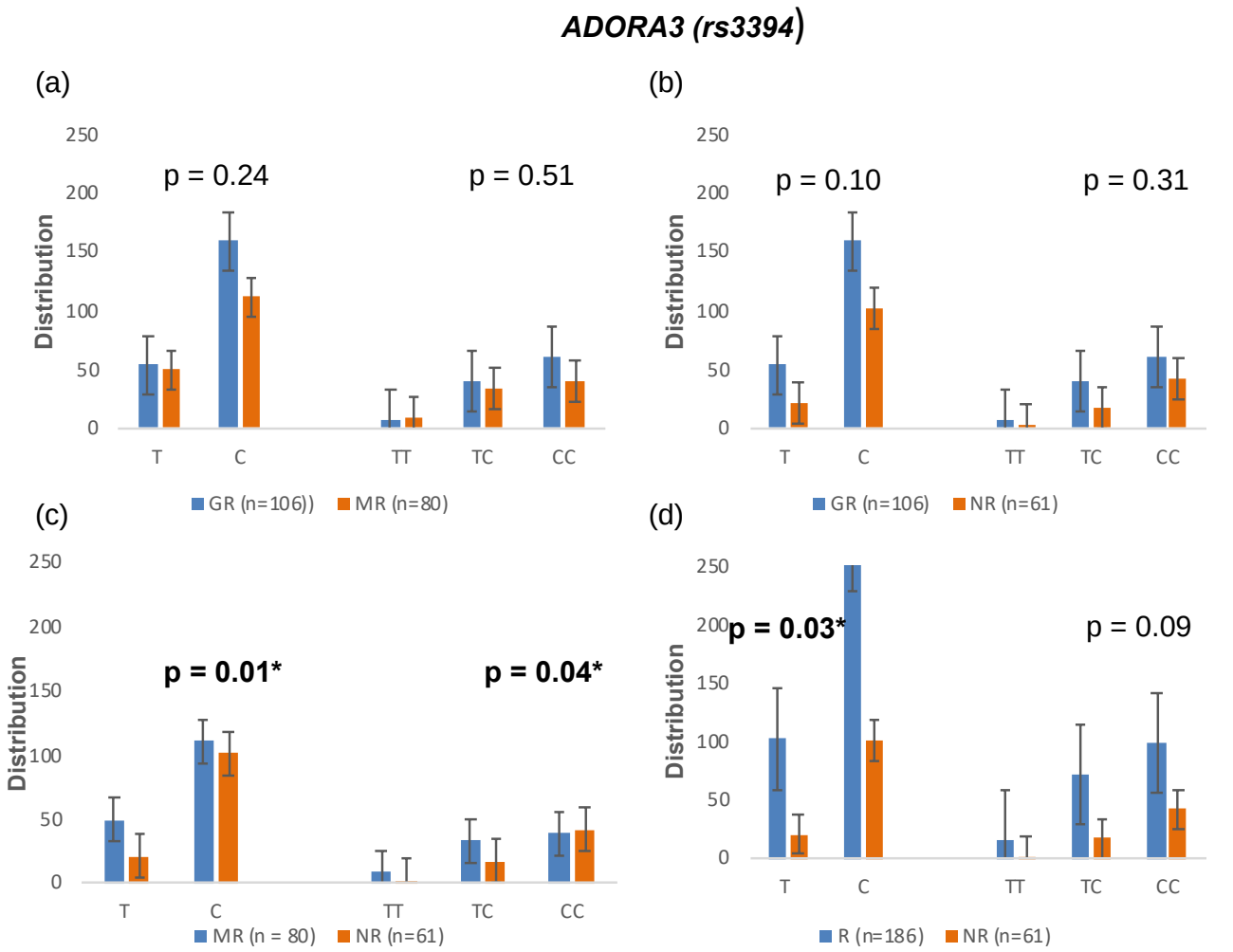


Figure 3.2: The genotype and allele frequencies comparisons of a SNP in the *ADORA3* gene, rs3394, previously associated with MTX response. The allele and genotype frequencies were not significantly different between (a) GR and MR and b) GR was compared with NR but was significant when c) MR were compared with NR. Only the allele frequency was statistically significantly different when d) R compared with NR at P-value < 0.05. The standard errors were also calculated and are represented by the thin black lines on the vertical bars. The number of individuals in each group is represented by n. The allele and genotype are represented on the X-axis and the frequency is represented on the Y-axis. The comparisons showing statistically significant allele and genotype frequency differences are represented by bold numbers and an asterisk.

When the SNP (rs2236225) on *MTHFD1* gene was assessed between the four groups, only the genotype frequencies GR and MR was significantly different between (P = 0.05) (Figure 3.3a). No statistical difference was observed between the allele and genotype frequency when the other groups were compared (Figure 3.3).

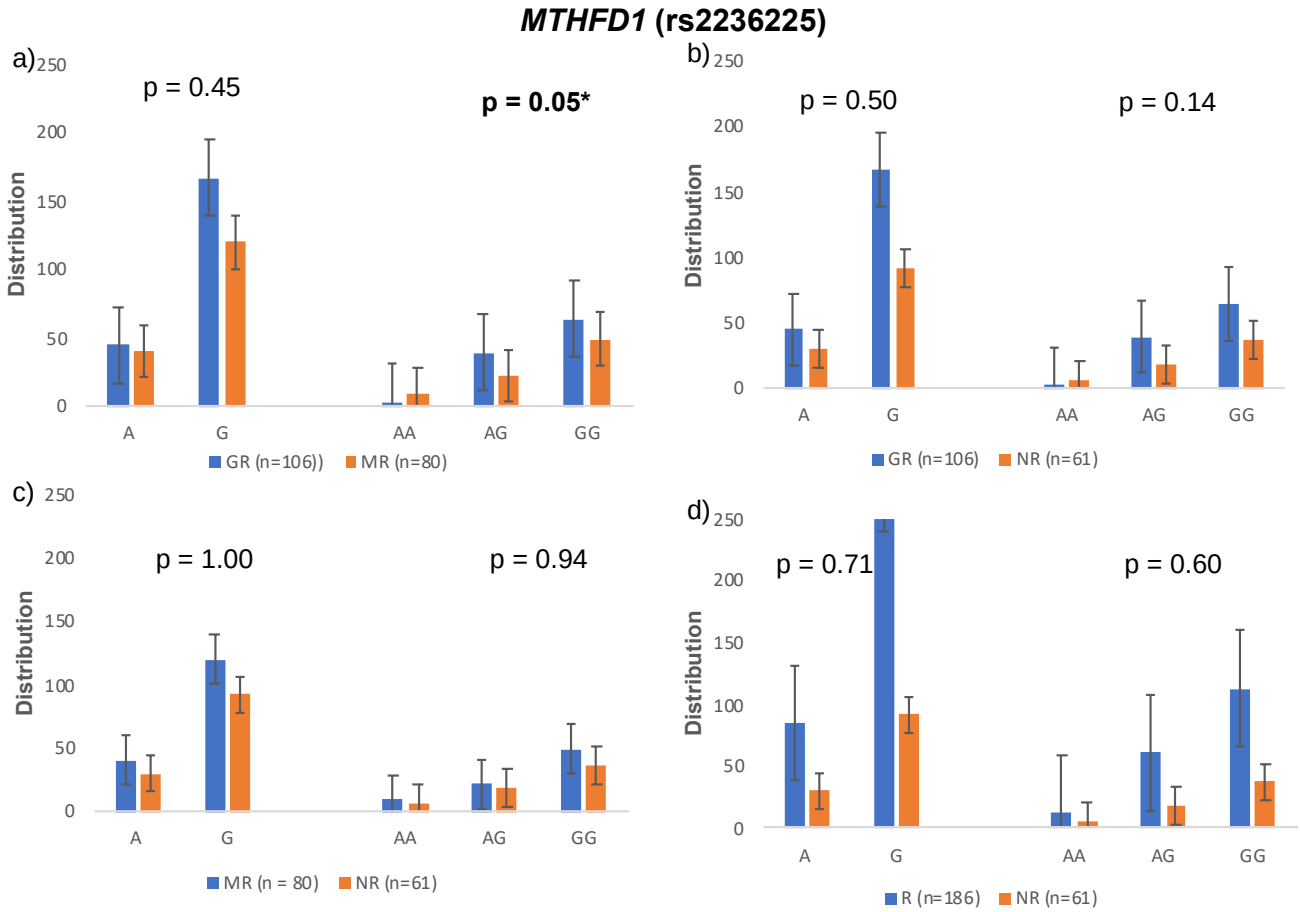


Figure 3.3: The genotype and allele frequencies comparisons of a SNP in the *MTHFD1* gene, rs2236225, previously associated with MTX response. Only the allele frequency between (a) GR and MR was significantly different (Figure 3.3a). There was no allele and genotype statistically significantly different when b) GR were compared with NR, (c) MR and NR, and (d) R compared with NR at P-value < 0.05. The standard errors were also calculated and are represented by the thin black lines on the vertical bars. The number of individuals in each group is represented by n. The allele and genotype are represented on the X-axis and the frequency is represented on the Y-axis. The comparisons showing statistically significant allele and genotype frequency differences are represented by bold numbers and an asterisk.

When the *SLC19A1* SNP, rs1051266, was compared between the groups, only the allele distribution between GR and MR (P = 0.03) (Figure 3.4a) was statistically significant. No

statistical difference was observed between the allele and genotype frequency when the other groups were compared (Figure 3.4).

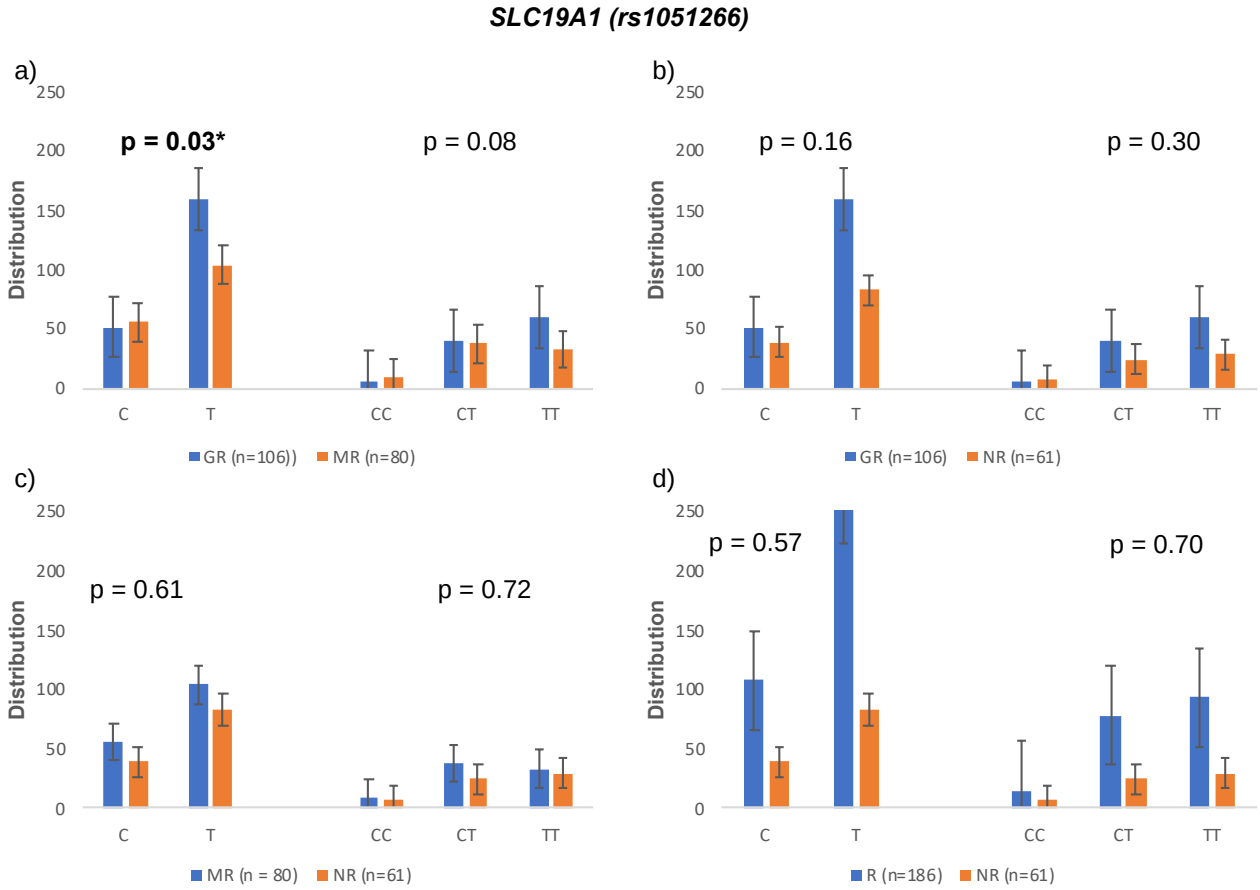


Figure 3.4: The genotype and allele frequencies comparisons of a SNP in the *SLC19A1* gene, rs1051266, previously associated with MTX response. Only the allele frequency was significantly different between (a) GR and MR (Figure 3.4a). There was no allele or genotype statistically significantly different when b) GR were compared with NR, (c) MR and NR, and (d) R compared with NR at P-value < 0.05. The standard errors were also calculated and are represented by the thin black lines on the vertical bars. The number of individuals in each group is represented by n. The allele and genotype are represented on the X-axis and the frequency is represented on the Y-axis. The comparisons showing statistically significant allele and genotype frequency differences are represented by bold numbers and an asterisk.

When SNPs in *ATIC* were assessed, the allele frequency of rs7563206 was significantly different between GR and NR (P = 0.02) (Figure 3.5d). No statistically significant difference was observed between the allele and genotype frequency when the other groups were compared (Figure 3.4).

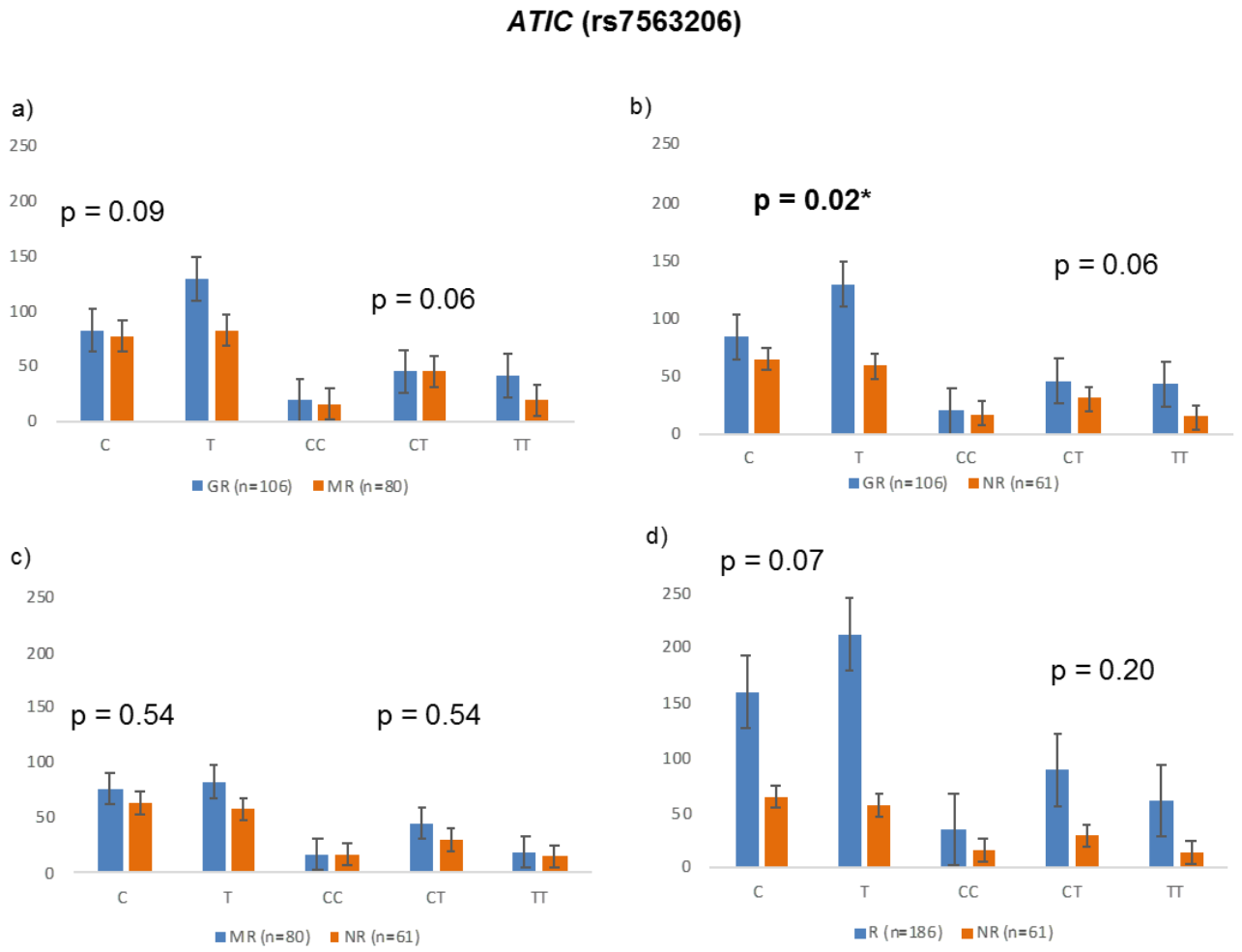


Figure 3.5: The genotype and allele frequencies comparisons of *ATIC* SNP (rs7563206) gene previously associated with MTX response. Only the allele frequency of rs7563206 was significantly different when (b) GR were compared with NR. There was no allele and genotype frequency statistically significant difference when (c) MR was compared with NR, and when (d) R were compared with NR for both SNPs at P < 0.05. The standard errors were also calculated and are represented by the thin black lines on the vertical bars. The number of individuals in each group is represented by n. The allele and genotype are represented on the X-axis and the frequency is represented on the Y-axis. The comparisons showing statistically significant allele and genotype frequency differences are represented by bold numbers and an asterisk.

3.4.4. Exact replication of genetic association of 36 index SNP association with MTX response

A total of 36 SNPs previously associated with MTX response were tested in a logistic regression with an additive effect model controlling for sex, age, smoking and disease duration (Table 3.3). Two of the 36 SNPs that mapped to *ABCB1* (rs1045642, OR = 3.03 (0.45 – 1.26), P unadj = 0.01) and *ADORA3* (rs3394, OR = 1.78 (0.27 – 1.05), P unadj =

0.03) appeared to show association with MTX response at a P < 0.05. However, after adjusting for multiple testing by using 1000 permutations, none of the SNPs were significantly associated with MTX response (*ABCB1* rs1045642, P adj = 0.33 and *ADORA3* rs3394 P adj = 0.64).

Table 3.3: Exact replication of 36 previously associated SNPs with methotrexate response

Genes	SNP	A1	A2	MAF		OR (95% CI)	P unadj	P adj
				R	NR			
ADORA3	rs3394	T	C	0.27	0.17	1.78 (1.05-3.03)	0.03	0.64
MTR	rs1805087	G	A	0.25	0.30	0.73 (0.46-1.15)	0.18	1
MTHFR	rs1801133	A	G	0.09	0.11	0.70 (0.36-1.35)	0.29	1
	rs1801131	G	T	0.11	0.12	0.97 (0.50-1.87)	0.92	1
	rs7563206	C	T	0.43	0.48	0.66 (0.44-1.01)	0.06	0.83
ATIC	rs2372536	G	C	0.05	0.08	0.63 (0.28-1.40)	0.25	1
	rs16853826	A	G	0.18	0.14	1.30 (0.72-2.36)	0.39	1
	rs12995526	C	T	0.27	0.28	0.84 (0.53-1.32)	0.46	1
	rs4673993	C	T	0.09	0.10	0.97 (0.49-1.91)	0.92	1
	rs3821353	T	G	0.22	0.19	0.99 (0.59-1.67)	0.96	1
MTRR	rs1801394	G	A	0.33	0.30	1.08 (0.67-1.74)	0.75	1
	rs1045642	A	G	0.14	0.05	3.03 (1.26-7.32)	0.01	0.33
ABCB1	rs1128503	A	G	0.10	0.06	1.96 (0.81-4.73)	0.13	1
	rs10280623	C	T	0.26	0.24	1.12 (0.67-1.88)	0.67	1
GGH	rs3758149	A	G	0.15	0.20	0.70 (0.41-1.22)	0.21	1
	rs1544105	C	T	0.36	0.30	1.36 (0.85-2.16)	0.20	1
FPGS	rs10760502	A	G	0.13	0.10	1.48 (0.73-2.98)	0.28	1
	rs10106	T	C	0.40	0.39	1.12 (0.71-1.76)	0.62	1
ABCC2	rs2273697	A	G	0.17	0.14	1.27 (0.70-2.30)	0.43	1
	rs4148396	T	C	0.22	0.22	1.18 (0.70-2.00)	0.54	1
SLC22A11	rs11231809	A	T	0.05	0.07	0.95 (0.43-2.12)	0.90	1
KLRC1	rs7301582	C	T	0.41	0.38	1.14 (0.74-1.75)	0.55	1
KLRD1	rs2302489	T	A	0.26	0.26	1.11 (0.66-1.87)	0.69	1
MTHFD1	rs2236225	A	G	0.23	0.25	1.01 (0.63-1.63)	0.97	1
ABCC1	rs3784864	A	G	0.09	0.08	1.39 (0.63-3.04)	0.42	1
	rs246240	G	A	0.07	0.07	0.90 (0.40-2.04)	0.80	1
SHMT1	rs1979277	A	G	0.40	0.37	1.24 (0.80-1.93)	0.33	1
TYMS	rs2853539	A	G	0.12	0.17	0.76 (0.41-1.43)	0.40	1
ADA	rs244076	C	T	0.49	0.45	1.19 (0.78-1.76)	0.45	1
SLC19A1	rs1131596	A	G	0.24	0.23	1.11 (0.68-1.81)	0.67	1
	rs1051266	C	T	0.29	0.32	0.97 (0.61-1.54)	0.90	1

	rs2267076	T	C	0.26	0.20	1.32 (0.77-2.27)	0.32	1
	rs3761422	T	C	0.26	0.20	1.28 (0.74-2.20)	0.38	1
ADORA2A	rs5751876	C	T	0.34	0.39	0.83 (0.53-1.31)	0.42	1
	rs5760410	G	A	0.37	0.31	1.21 (0.75-1.94)	0.43	1
	rs2298383	T	C	0.30	0.30	0.97 (0.61-1.56)	0.91	1

SNP- single nucleotide polymorphisms, CHR – chromosome, BP – base position, A1 – minor allele, A2 – major allele, MAF – minor allele frequency, R – responders to methotrexate treatment, NR – methotrexate non-responders, OR – odd ratio, 95% CI – confidence interval at 95%, P – P-value measure of statistical significance.

3.4.5. Assessment of regional association with MTX response

After evaluation of 7812, only 17 SNPs (all imputed at quality score >0.8) mapping to a single region within 50 kb on chromosome 9 region were significantly associated with MTX response (Table 3.4). Fourteen of these SNPs remained significant after correcting for multiple testing with 1000 permutations (Table 3.4). Two of these SNPs mapped to *CDK9* (rs114465623, OR = 0.12 (0.05 – 0.29), P = 4.338 X 10^{-06}; rs80266961, OR = 0.12 (0.05 – 0.29), P = 4.338 X 10^{-06}), three mapped to *FPGS* (rs115707377, OR = 0.12 (0.05 – 0.29), P = 4.338 X 10^{-06}; rs41358947, OR = 0.12 (0.05 – 0.29), P = 4.338 X 10^{-06}; rs78191768, OR = 0.12 (0.05 – 0.29), P = 4.338 X 10^{-06}) and one mapped on *ENG* (rs150775138, OR = 0.12 (0.05 – 0.29), P = 4.338 X 10^{-06}). The minor allele frequency was higher in non-responders than among responders for all the significant SNP, suggesting these variants may be associated with an increased risk of reduced response to MTX treatment. A LocusZoom plot showed that the significantly associated SNPs were in strong LD with the lead SNP (rs114465623) across the region but mapped to different adjacent genes (Figure 3.6).

Table 3.4: SNPs significantly associated with methotrexate response in the regional analysis. Multiple testing was corrected for using 1000 permutations (P adj)

Gene	CHR	SNP	BP	A1	A2	MAF*		OR (95% CI)	P unadj	P adj
						R	NR			
FPGS	9	rs115707377	130556190	A	T	0.03	0.14	0.12 (0.05-0.29)	4.30 X 10^{-06}	0.004
FPGS	9	rs41358947	130578280	G	A	0.03	0.14	0.12 (0.05-0.29)	4.30 X 10^{-06}	0.004
FPGS	9	rs78191768	130568160	T	C	0.03	0.14	0.12 (0.05-0.29)	4.30 X 10^{-06}	0.004
FPGS	9	rs41483150	130576512	C	A	0.04	0.14	0.15 (0.06-0.35)	1.76 X 10^{-05}	0.018
ENG	9	rs150775138	130583449	T	G	0.03	0.14	0.12 (0.05-0.29)	4.30 X 10^{-06}	0.004
ENG	9	rs76326400	130592992	G	A	0.03	0.13	0.12 (0.05-0.31)	1.08 X 10^{-05}	0.012
ENG	9	rs41409546	130588169	T	C	0.03	0.13	0.13 (0.05-0.33)	1.41 X 10^{-05}	0.015
CDK9	9	rs114465623	130551120	C	A	0.03	0.14	0.12 (0.05-0.29)	4.30 X 10^{-06}	0.004
CDK9	9	rs80266961	130546204	T	C	0.03	0.14	0.12 (0.05-0.29)	4.30 X 10^{-06}	0.004
CDK9	9	rs114028917	130546601	C	G	0.03	0.13	0.12 (0.05-0.31)	1.10 X 10^{-05}	0.012
CDK9	9	rs1002095	130545950	A	G	0.06	0.16	0.19 (0.08-0.42)	4.57 X 10^{-05}	0.05
CDK9	9	rs2297214	130551153	A	G	0.06	0.16	0.19 (0.08-0.42)	4.57 X 10^{-05}	0.05
CDK9	9	rs3217740	130550044	T	C	0.06	0.16	0.19 (0.08-0.42)	4.57 X 10^{-05}	0.05
CDK9	9	rs3217751	130552915	G	A	0.06	0.16	0.19 (0.08-0.42)	4.57 X 10^{-05}	0.05
CDK9	9	rs5019766	130544671	C	T	0.06	0.16	0.20 (0.09-0.45)	8.05 X 10^{-05}	0.08
CDK9	9	rs10819306	130553108	C	T	0.06	0.16	0.21 (0.09-0.45)	8.30 X 10^{-05}	0.09
CDK9	9	rs10987729	130553436	T	C	0.06	0.16	0.21 (0.09-0.45)	8.30 X 10^{-05}	0.09

SNP- single nucleotide polymorphisms, CHR – chromosome, BP – base position, A1 – minor allele, A2 – major allele, MAF – minor allele frequency, R – responders to Methotrexate treatment, NR – methotrexate non-responders, OR – odd ratio, 95% CI – confidence interval at 95%, P – P-value measure of statistical significance.
*MAF calculates the allele frequency of the minor allele (A1) in both R and NR

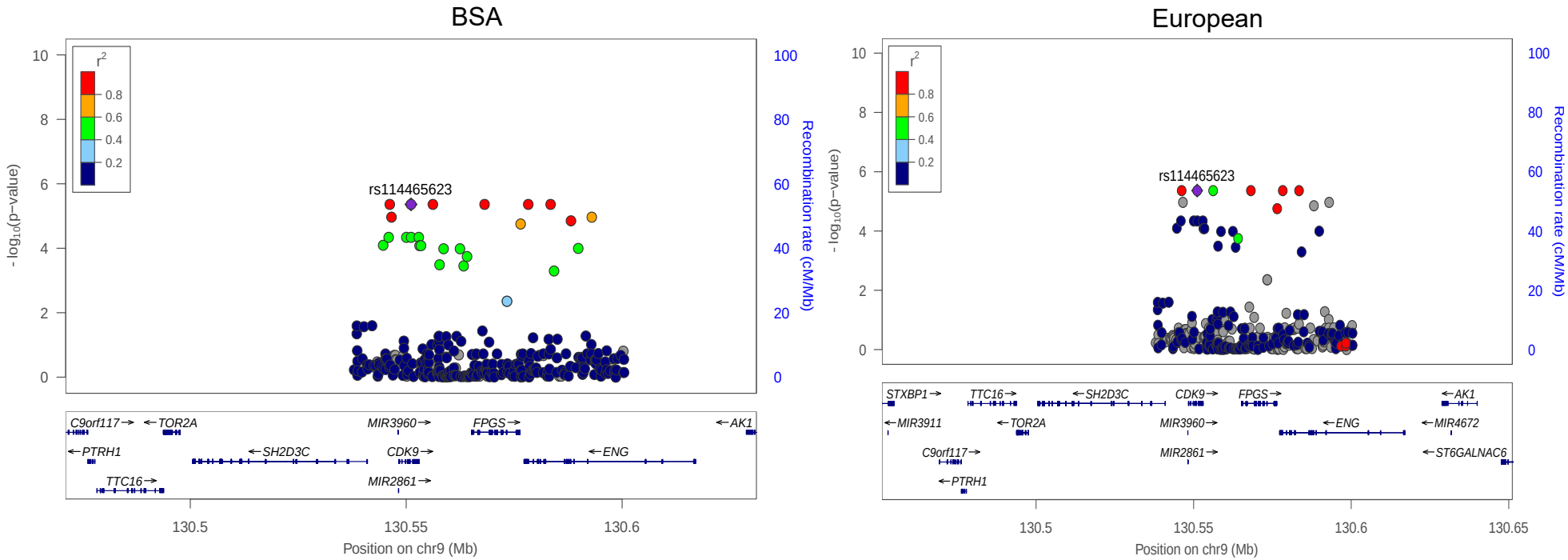


Figure 3.6: Regional plot on chromosome 9 showing the significantly associated imputed SNPs and their positions relative to the genes in BSA compared to in European ancestry. The plot also denotes the degree of LD based on the 1000 Genome Project African ancestry and European population, shown by the different colours. Red indicating an almost complete LD (r_2 > 0.8) and dark blue indicating a very low degree of LD. Each SNP is plotted as a -log10 of the P-value produced by LocusZoom**.**

3.4.6. Functional annotation predictions of significant SNP

To functionally predict and annotate the SNP that were significantly associated with MTX response, publicly available annotation tools were used. Using the experimental evidence-based functional annotation RegulomeDB database, four of the 14 SNPs mapping to *CDK* (rs1002095, score = 1f), *FPGS* (rs115707377, score = 2b), *CDK9* (rs80266961, score = 2b) and *CDK* (rs114465623, score = 3a) had higher probability of exerting a gene regulatory effect as indicated by a score of <3 (Table 3.5). Further analysis of the four SNPs to asses regulatory histone chromatin marks tested in whole blood and immune cells showed that two (rs1002095 and rs80266961) overlapped a promoter region. Although all four overlapped DNase I hypersensitivity sites, only three also overlapped enhancer histone marks. Consequently, the SNP reported having high evidence to likely affect transcription factor binding (rs1002095) by RegulomeDB also correlated with eQTL effects in transformed fibroblasts cells.

Table 3.5: Functional annotation prediction of SNP significantly associated with methotrexate response (sorted by base position)

SNP	Gene	Genomic location	RegulomeDB	Regulatory Histone Marks			eQTL target gene
				Promoter	Enhancer	DNase	
rs1002095	***CDK9***	5' upstream	1f	Yes	No	Yes	*FPGS*
rs80266961	***CDK9***	5' upstream	2b	Yes	Yes	Yes	-
rs114028917	***CDK9***	5' upstream	5	Yes	Yes	Yes	-
rs3217740	***CDK9***	Intronic	5	Yes	Yes	Yes	*FPGS*
rs114465623	***CDK9***	Intronic	3a	No	Yes	Yes	-
rs2297214	***CDK9***	Intron	4	No	Yes	Yes	*FPGS*
rs3217751	***CDK9***	3’ UTR	5	No	Yes	Yes	*FPGS*
rs10987729	***CDK9***	3' downstream	4	No	No	Yes	*FPGS*
rs115707377	***FPGS***	5' upstream	2b	No	Yes	Yes	-
rs78191768	***FPGS***	Intronic	4	No	Yes	Yes	-
rs41483150	***FPGS***	3' downstream	4	No	Yes	Yes	-
rs41358947	***FPGS***	3' downstream	4	No	Yes	Yes	-
rs150775138	***ENG***	Intronic	5	No	Yes	Yes	-
rs41409546	***ENG***	Intronic	4	Yes	Yes	Yes	-
rs76326400	***ENG***	Intronic	5	No	Yes	Yes	-

SNP- single nucleotide polymorphisms, Yes – evidence suggests SNP overlaps that regulatory region, No – evidence suggest SNP does not overlap that region.

3.5. Discussion

Methotrexate remains the anchor drug in the treatment of RA. Although largely effective, interpatient variability in clinical response presents a challenge in achieving and maintaining widespread remission in RA patients. Genetic polymorphisms have been shown to influence MTX outcome, which may, at least in part, explain the variation in clinical response. However, there is a dearth of data on the pharmacogenetics of MTX response in people of African ancestry. This study was undertaken to firstly assess the allele and genotype frequency of 36 previously identified MTX metabolism associated SNPs with MTX response in BSA RA patients. Secondly, to evaluate the association between the 36 SNPs with MTX response. Thirdly, to evaluate the regional association of 7821 SNPs surrounding the 36 index SNP (25kb in both directions), with MTX response after adjusting for potential confounders. Finally, to functionally predict and annotate the significantly associated SNP using publicly available data.

3.5.1. SNP frequency and exact replication

This study conducted allele and genotype frequency of 36 SNPs previously associated with MTX response in Eurasian populations. Our assessment revealed that six SNPs across five genes were significantly different when comparing four groups of GR, MR, NR, and R. However, after logistic regression and adjustment of confounders, the previously identified index SNP were not significantly associated with MTX response in this study population. This finding highlights the difference in genetic architecture between Eurasian population and African ancestry population, which affect the allele frequency of a SNPs between responders and non-responders. For example, rs3394 mapping to *ADORA3* and rs1045642 in *ABCB1* had a frequency difference of 10% and 9% between responders and non-responders. For SNPs that were statistically different between the compared groups, the lack of association may suggest that the study was not powered to detect the association in a logistic regression model. Similar results have recently been reported for *ADORA* (rs3394) with 271 patients (Kobold et al. 2019) and for *ABCB1* (rs1045642) with 212 patients (Jenko et al. 2017) of European ancestry, where the SNP had significant different frequencies between responders and non-responders, but were not associated with MTX response when fitted to a logistic regression model. Therefore, larger studies are warranted to confirm findings.

Our study corroborated a finding from a study which included 30 AA participants and 193 Caucasians, which reported that there was no association between the two extensively studied SNPs (rs1801133 and rs1801131) on *MTHFR* and MTX response (Hughes et al. 2006). Interestingly, the allele frequencies for the two SNPs rs1801133 and rs1801131) on *MTHFR* are similar between AA (0.11 and 0.13) and BSA (0.09 and 0.11) but high in Caucasian (0.30, 0.32), which may explain the lack of association between both groups (Hughes *et al*. 2006). Furthermore, the low SNP frequency and lack of significantly different allele frequency in our study population could partly explain the lack of toxicity observed by in a longitudinal retrospective observational study with 200 BSA RA patients on MTX Zannettou et al. 1997 personal communication). Cumulatively, these findings further confirm the heterogeneity and complexity of MTX treatment response in different ethnic groups. To mitigate the challenges of genetic architecture and allele frequency difference, a regional or local replication was performed.

3.5.2. Regional replication

Regional-based analysis revealed a statistically significant association of 50kb region comprised of 14 imputed SNPs on chromosome 9 overlapping three gene*s* (*CDK, FPGS, ENG*). Furthermore, the strongly associated SNPs were found to be in strong LD across *CDK, FPGS, ENG* using LocusZoom. These findings lend support to the hypothesis that regional replication has the added power to capture closely linked potential risk-associated SNP (Kohler *et al*. 2014; Pan 2011; Wu *et al*. 2011). This means that with the difference in genetic architecture between ethnic groups, a different SNP in a previously associated region could be associated with the trait of interest. Furthermore, an association within a region is accepted as confirmation of the role of the gene or region in the trait of interest (Clarke et al. 2007). We observed in this study that the significantly associated SNPs had a higher minor allele frequency in non-responders than in responders, which may suggest that the variant allele is correlated with a reduced response to MTX treatment in BSA RA patients. Of interest are SNPs overlapping the *FPGS* gene as variation in this gene have been previously shown to be associated with MTX response in RA patients (Dervieux et al. 2005).

FPGS gene

FPGS gene encodes an enzyme very crucial for activating MTX by adding glutamates to make it MTX-GLU, increasing intracellular retention of polyglutamated MTX (MTX-PG) and eliciting MTX anti-inflammatory activity (Yamamoto et al. 2016). A recent study has shown

that *FPGS* variants influence intracellular MTX-PGs retention and thus pharmacological activity (Yamamoto *et al*. 2016; Moya *et al*. 2016). In contrast, inefficient polyglutamation has been associated with MTX resistance (Rots et al. 1999). This study identified four SNPs in *FPGS* (rs41483150, rs41358947, rs115707377, rs78191768), which were associated with MTX treatment response in this study. Two SNPs (rs41358947 and rs41483150) are 437bp and 2.2kb from *FPGS* rs10106, which has previously been associated with MTX reduced response in Dutch (van der Straaten *et al*. 2007), Japanese (Hashiguchi *et al*. 2016) Indian cohorts (Senapati *et al*. 2014) Caucasian and non-Caucasian (Chen et al. 2016). Although the SNPs are relatively close to the previously associated SNP (rs10106), LD analysis with all four SNPs identified in this study in Ensembl (http://grch37.ensembl.org/index.html) using the KGP African ancestry dataset revealed that the SNPs are not in complete LD ($r_2 < 0.2$ for both SNP). This finding may highlight heterogeneity inherent in different ancestries and suggest that the SNPs identified in *FPGS* are independent of the previously reported rs10106, thus explaining the lack of association for the index SNP.

Gene expression of *FPGS* in peripheral blood mononuclear cells (PBMCs) reported an association with poor response to MTX in RA patients (Stark *et al*. 2009; Stranzl *et al*. 2003). Interestingly, a SNP (rs115707377, RegulomeDB = 2b) in *FPGS* was predicted to likely to affect the binding of transcription factors to target gene/s. This finding further supports an association between the identified locus in *FPGS* with modulation of expression of target genes. Further studies should aim to into validate this association in larger studies of African ancestry.

CDK9 gene

Our study also revealed eight SNPs (rs1002095, rs80266961, rs114028917, rs3217740, rs114465623, rs2297214, rs3217751, rs10987729) mapping to the *CDK9* gene. Interestingly, one of the associated SNP, rs1002095, is in LD with the previously identified rs10106 in European ancestry. Furthermore, rs1002095, upstream of *CDK9*, is in a varying degree of LD ($r_2 > 0.6$) with all four SNPs, particularly rs115707377 (*FPGS*), which was in almost complete LD with rs1002095 (*CDK9*) in our BSA cohort. Interestingly, the LocusZoom in European ancestry also shows SNPs mapping to *CDK9* in LD with SNPs mapping to *FPGS* (Figure 3.6). Furthermore, due to the similarity of the strength of association and allele frequency, it is difficult to ascertain which SNP or SNPs could be driving the observed association in BSA.

CDK9 encodes an enzyme involved in the regulation of cell cycle by transcriptional elongation through RNA polymerase II (Larochelle et al. 2012). Functional annotation revealed that *CDK9* SNPs (rs1002095; rs80266961 and rs114465623), were predicted by RegulomeDB to have a high probability of altering transcription factor binding. Furthermore, *CDK9* SNPs (rs10987729, rs2297214, rs3217740, rs3217751) were predicted to be a *cis*-eQTL of *FPGS* in transformed fibroblast cells. Transcriptomics studies with PBMCs from collagen-induced RA mouse models revealed that inhibition of this *CDK9* gene induced apoptosis and modulated immune cell function (Hellvard et al. 2016; Wang et al. 2012), which is one of MTX's pharmacological activity in RA therapy. Taken together, these results support a possible association of the regulation of gene expression of target genes including *FPGS*, as a mechanism of MTX response in BSA RA patients.

The major limitation of the study was the small sample size and thus it may not have been powered to detect SNPs with small to moderate effect sizes. In addition, the study was conducted retrospectively with some data being extracted from patient files, which depends on uncontrolled variables such as physician experience and preference in data collection, which could influence the accuracy and quality of data. Furthermore, the significantly associated SNPs are from imputed data and were not actually genotyped on the array. Although SNPs had a good imputation quality score of above 0.8, these findings should be validated with directly genotyped SNPs. It is for these reasons that the study should be considered a pilot study that requires replication in a larger sample size of participants of sub-Saharan ancestry.

In conclusion, despite the limited sample size, this study replicated an association between a locus on chromosome 9 with MTX response in BSA RA patients on MTX monotherapy for 6 months. Furthermore, this study reports significantly associated novel SNP associations on the previously MTX-response associated genes *FPGS* and *CDK9,* with predicted functional annotation evidence in relevant tissues and cells. Our findings suggest that MTX response appears to be influenced or at least partly influenced, by regulation of *FPGS* expression in BSA. These preliminary but novel SNP associations should be validated in a larger cohort and the molecular mechanism of *FPGS* regulation should also be confirmed experimentally. These findings, although preliminary, are hypothesis-generating and warrant a large-scale study to establish biomarkers for MTX response in BSA.

CHAPTER 4: OVERALL DISCUSSION AND CONCLUSION

The aims of this study were two-fold. Firstly, to use a case-control GWAS approach to investigate genetic variants associated with susceptibility to seropositive RA in BSA. Secondly, to evaluate exact and regional association for SNP previously associated with MTX response in BSA.

4.1. GWAS study

After GWAS genotype data imputation, a total of 7,986,485 SNPs were tested for association in 577 RA cases compared to 1612 population controls in a logistic regression analysis assuming an additive model and adjusting for sex, age, smoking and the first three principal components. The GWAS, as expected, confirmed that variants in the HLA region confer the strongest increased risk to RA in BSA. This is in accordance with previous RA genetic studies across all populations, including BSA. The widely reported difficulty of imputing the HLA region was also observed in this study, where a total of 42 SNPs were lost during imputation and post-imputation QC. Although not completely understood, factors such as the high polymorphism, gene homology and LD between genes and SNPs in the HLA locus region have been implicated. It was for this reason that the HLA region pre-imputation data was used in subsequent analysis.

The precise independent localisation and the number of contributing alleles to the HLA effect on RA pathogenesis remains an active area of research. There is an ongoing debate in the science community about the exact region of highest association, with overwhelming genotype and amino acid evidence suggesting the *HLA-DRB1* locus is primarily responsible for most of the observed associations, but consensus is lacking and there may be ethnic-specific differences. Previous findings have also suggested that multiple independent SNPs may synergistically be responsible for the majority of the association observed in the HLA region. Interestingly, a study in Europeans showed that correcting for the *HLA-DRB1* locus in RA cases completely removed the association observed in the HLA region, suggesting that the observed associations were not independent of the *HLA-DRB1* risk locus. However, our study identified two independent SNPs, rs617578 intergenic between *HLA-DRB1* and *HLA-DQA1*, and rs28377109 intergenic to *HLA-DRB9* and *HLA-DRB5,* when conditioned for the top SNP (rs617578), which supports the multiple independent HLA region risk locus hypothesis. Furthermore, the regional visualisation LocusZoom showed that although there were few SNPs in strong LD with the top SNP in European and Asian populations in the KGP,

there were none with the African populations in the same dataset. Cumulatively, these findings appear to support that the most significant HLA locus association in this study was an independent signal. However, the HLA region is highly polymorphic and not all the variation was captured in our study. Furthermore, HLA genes and consequently embedded SNPs are in very strong LD, making exclusion of the influence of the *HLA-DRB1* locus very difficult.

The GWAS with imputed data identified a novel non-HLA risk locus in intron 12 and three other risk loci in intron 14 of *CPT1A*. Previous studies have reported an association between RA related traits such as obesity and bone density, and SNPs mapping to the *CPT1A* gene. In addition, gene expression studies in cell culture and mouse studies have all shown that the *CPT1A* gene is differentially expressed in CD4+ T-cells from patients with autoimmune diseases, including RA. This study offers insight into the genetic association of *CPT1A* with RA and demonstrates a potential association between a locus in *CPT1A* and RA in Africans. In addition, functional analysis revealed that the identified *CPT1A* locus overlaps regulatory histone marks and had evidence suggesting a high probability of influencing binding to transcription factors and eQTL. Cumulatively, these results suggested that the *CPT1A* risk locus may be potentially associated with a role in the regulation and expression of a target gene or genes involved in RA pathogenesis. Moreover, chromatin interaction mapping studies using high-throughput chromatin conformational capture (HI-C) data revealed that, apart from being a cis-QTL to *CPT1A*, the independent risk locus, rs3019593, also showed evidence of regulating the gene expression of the nearby genes, *MTL, GAL, and PPP6R3* through chromatin interactions. *MTL* and *GAL* have previously been reported to be dysregulated in RA and associated with modulation of oxidative stress, which is a risk factor for RA. Taken together, these results demonstrate the potential role of the dysregulated expression of the associated genes and their potential role in RA pathogenesis. This study, therefore, identified non-HLA SNPs in *CPT1A* that appear to be associated with the regulation and expression of several genes implicated in the pathogenesis of RA.

This GWAS also replicated 10 non-HLA SNPs associations previously identified largely in European and Asian ancestry populations. This study further replicated a non-HLA SNP previously identified in a study of participants of sub-Saharan ancestry (a sub-group of the present study). The SNP replication in this study suggests that although there is heterogeneity among different population ancestries that lead to differences in minor allele

frequency, there appears to be a significant concordance of RA risk loci, albeit with lower effect sizes in BSA.

The primary limitation is the moderate sample size which is not sufficiently powered to detect rare associated SNPs. Furthermore, the sample size may not be enough to detect SNPs with low and modest effect sizes. However, despite the small sample size, this study identified novel SNP associations and replicated a previously identified associated region. Another limitation is that most of the identified SNPs overlap non-coding regions, which makes an accurate elucidation of SNP functionality, conclusively inferring true causality and linking variation to phenotype difficult. Furthermore, the prediction annotation databases were mostly generated using healthy individuals, which may not necessarily represent a diseased cellular state. However, the integration of GWAS findings with other biological functional data such as tissue-specific regulatory and expression prediction data has, in part at least, assisted to explain the potential disease pathways, prioritise genes of interest, and contributed to hypothesis generation.

In conclusion, this study confirmed the association of the HLA locus with increased risk of RA and identified a potentially novel independent association in the intergenic region of HLA-*DRB1* and *HLA-DQA1* in people of sub-Saharan African ancestry. Taking into consideration the findings of other empirical studies that highlight a potential role for the *CPT1A* locus in autoimmune diseases, this study showed that genetic variants in an intronic region of *CPT1A* contribute to RA pathogenesis.

4.2. Pharmacogenetics

The pharmacogenetics of MTX study was conducted in two parts. First, 36 index SNPs previously associated with MTX response in Eurasian ancestry were compared in 247 (186 responders and 61 non-responders) BSA RA patients on MTX monotherapy for at least six months. Next, a regional genomic association of 7821 SNPs surrounding the 36 previously identified SNPs were tested in the same population. There was no association observed with the 36 previously identified SNPs after correcting for multiple testing. Although limited by sample size to detect previously associated SNPs, the lack of association did highlight potential heterogeneity of genetic association between Eurasians and BSA. The regional analysis with 7821 SNPs identified a risk locus within 50kb of a previously associated SNPs,

rs10106 (*FPGS*), which was associated with MTX treatment response in BSA. The identified locus overlaps a region containing three genes (*CDK9, FPGS, ENG)*. The minor allele frequencies of the SNPs were higher in non-responders compared to responders, suggesting an association of the variant alleles with a reduced MTX treatment response. *In silico* functional prediction and annotation revealed that the risk loci contain amongst other variants four SNPs in *CDK9* that are cis-eQTL to the *FPGS* gene in fibroblast-like synoviocyte cells and whole blood tissue. One of the SNPs, rs1002095, had experimental evidence of influencing binding to transcription factors, thus suggests its role in modulation of transcription and expression of *FPGS*. This gene is involved in polyglutamation, which has previously been shown to influence response to MTX treatment.

The major limitation of this component of the study was the small sample size and very limited power. In addition, the study was conducted retrospectively with some data being extracted from patient files, which depends on uncontrolled variables such as physician experience and differences in data collection, which could influence the accuracy and quality of data. Furthermore, the significantly associated SNPs are from imputed data and were not actually genotyped on the array. Although SNPs had a good imputation quality score of above 0.8, these findings should be validated with directly genotyped SNPs. It is for these reasons that the study should be considered a pilot study that requires replication in a larger study conducted in sub-Saharan Africa.

In conclusion, despite the limited sample size, this study replicated an association between a locus on chromosome 9 with MTX response in BSA RA patients on MTX monotherapy for 6 months. Furthermore, this study reports significantly associated novel SNP associations on the previously MTX-response associated genes *FPGS* and *CDK9* with relevant predicted functional annotation. Our findings suggest that MTX response appears to be influenced or at least partly influenced, by regulation of *FPGS* expression in BSA. These preliminary but novel SNP associations should be validated in a larger cohort and the molecular mechanism of *FPGS* regulation should also be confirmed experimentally. These findings, although preliminary, are hypothesis-generating and warrant a large-scale study to establish biomarkers for MTX response in BSA.

4.3. Future studies

The associated genetic variants were identified using modest sample sizes in this study. Future studies should look into validating these findings in larger cohorts. In addition, with the difficulty of imputing SNPs in the HLA region being well reported across populations, future studies should consider imputing with an HLA enriched population-specific reference panel or using high throughput sequencing of the HLA region to validate the associations. These suggested methods have the potential to bridge the gap between genotyped SNPs missing in the imputed data because they were not in the imputing reference panel. Furthermore, along with advanced statistical analysis methods, the suggested methods have the potential to identify independent signals that drive the association observed with the HLA locus.

The *in silico* functional annotation and prediction of GWAS results using disease-specific tissue biological data has allowed prioritization of putative causal risk loci. Further studies should consider validating the predicted function using experimental approaches, including cellular and animal models. Because of the challenges of gene expression and chromatin interaction being tissue/cell-specific and time-dependent, the findings in this study should ideally be validated in disease-relevant cell cultures isolated from affected patients. The polymorphisms identified in this study will then be assessed in the experimental cells and exposed to the potential triggers of specific pathways. For example, to test the predicted cis-eQTL of *CPT1A* in synovial cells from RA patients by exposing the cells to fluorescently marked fatty acids. At different time points the activity of *CPT1A*, rate of FAO and the immune response will be measured. Cells can be exposed to heavy metals to assess the binding affinity of *MTL* in the presence of the other polymorphisms in the other genes. Gene expression and binding capacity can then be measured at different time points, along with levels of oxidative stress. This has the potential to replicate the biological cell behaviour in a disease state and reveal potential biological pathways.

A potential goal of performing GWAS is for application in personalised medicine to enable the development of polygenic risk scores (PRS) in individuals in order to stratify the population into low, moderate and high genetic risk. However, predictive potential for an individual remains relatively low. In a recent study, PRSs were used to illustrate their predictive value for coronary artery disease, showing that about 8% of European ancestry individuals from the UK biobank had a three-fold increased risk of developing coronary artery disease (Khera

et al. 2018). In addition, Jones and colleagues (2018) used PRS to confirm an association between RA and neural phenotypes in 7977 individuals. However, transferability of PRS to other non-European populations remains a challenge, thus exacerbating health disparities. PRS have been shown to perform poorly if developed in one population and then applied to an unmatched ancestry target population. For example, when PRS are constructed based on data from individuals of European ancestry, their prediction accuracy was reduced by 4.5-fold in non-Europeans (Martin et al. 2019) and were worst in individuals of African ancestry (Duncan et al. 2019). Although improving, there is currently a limited number of sufficiently powered genomic research studies for the clinical utility of PRS to be realised in individuals of African ancestry. Improving the of number well powered GWASs will allow improvement of technological and statistical tools that include significant amount of data from individuals of African ancestry, and thus the potential clinical utility of PRS.

The *CDK9* loci are predicted to influence the expression of the *FPGS* gene. As proof of concept, future studies should first ascertain the level of MTX-PG in BSA RA patients on MTX for six months. This can be done by collecting blood samples from BSA patients on MTX for six months and measure the red blood cells methotrexate polyglutamate concentrations using chromatography techniques (i.e. Liquid-chromatography mass-spectrometry or high-pressure liquid chromatography). After evaluating the methotrexate polyglutamate concentration, the influence of the polymorphic variant could be validated using cell culture as described for *CPT1A*.

REFERENCES

Reference list:

Abdel-Nasser AM, Rasker JJ and Valkenburg HA. (1997). Epidemiological and clinical aspects relating to the variability of rheumatoid arthritis. Semin Arthritis Rheum. 27:123–40. https://doi.org/10.1016/S0049-0172(97)80012-1.

Abreu L, Kucher F, Eckstein V, et al. (2019). P001 Differential expression of key metabolic genes in antigen-specific B cell subsets in rheumatoid arthritis and systemic lupus erythematosus Annals of the Rheumatic Diseases, 78:A1.

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, *Chapter 7*, Unit7.20. https://doi.org/10.1002/0471142905.hg0720s7

Afridi HI, Kazi TG, Brabazon D, et al. (2012). Interaction between zinc, cadmium, and lead in scalp hair samples of Pakistani and Irish smokers rheumatoid arthritis subjects in relation to controls. Biol Trace Elem Res. 2012 Aug;148(2):139-47. https://doi.org/10.1007/s12011-012-9352-6

Afridi HI, Kazi TG, Kazi N, et al. (2013). Evaluation of status of arsenic, cadmium, lead and zinc levels in biological samples of normal and arthritis patients of age groups (46– 60) and (61 – 75) years. Clin Lab. 59(1-2):143-53.

Aletaha D, Neogi T, Silman AJ, et al. (2010). 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Ann Rheum Dis 69: 1580-1588. http://doi.org/10.1002/art.27584.

Alghamdi J and Padmanabhan S. (2014). Fundamentals of Complex Trait Genetics and Association Studies. Handbook of Pharmacogenomics and Stratified Medicine 2014, Pages 235-257. https://doi.org/10.1016/B978-0-12-386882-4.00012-8.

Alpizar-Rodriguez D, Lesker TR, Gronow A, et al. (2019). Prevotella copri in individuals at risk for rheumatoid arthritis. Annals of the Rheumatic Diseases 78:590-593. http://dx.**doi**.org/10.1136/annrheumdis-2018-eular.5899

Alpízar-Rodríguez D, Pluchino N, Canny G, et al. (2017). The role of female hormonal factors in the development of rheumatoid arthritis, *Rheumatology*, Volume 56, Issue 8, August 2017, Pages 1254–1263, https://doi.org/10.1093/rheumatology/kew318

Anaya J, Correa PA, Mantilla RD, et al. (2001). Rheumatoid arthritis in African Colombians from Quibdo. Seminars in Arthritis and Rheumatism, Volume 31, Issue 3, December 2001, Pages 191-198. https://doi.org/10.1053/sarh.2001.27737

Anderson CA, Pettersson FH, Clarke GM, et al. (2010). Data quality control in genetic case-control association studies. Nature Protocols, 5(9), 1564–1573. http://doi.org/10.1038/nprot.2010.116

Ardlie KG, Kruglyak L, and Seielstad M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3, 299–309. https://doi.org/10.1038/nrg777

Arshad M, Bhatti A, and John P. (2018). Identification and in silico analysis of functional SNP of human TAGAP protein: A comprehensive study. PloS ONE 13(1): e0188143. https://doi.org/10.1371/journal.pone.0188143.

Attar SM. (2010). Adverse effects of low dose methotrexate in rheumatoid arthritis patients. A hospital-based study. Saudi Med J. 31(8):909-15.

Avina-Zubieta JA, Choi HK, Sadatsafavi M, et al. (2008). Risk of Cardiovascular Mortality in Patients WithRheumatoid Arthritis: A Meta-Analysis ofObservational Studies. Arthritis & Rheumatism (Arthritis Care & Research) Vol. 59, No. 12, December 15, 2008, pp 1690 –1697. http://doi.org/10.1002/art.24092.

Axmann R, Bohm C, Kronke G, et al. (2009). Inhibition of interleukin-6 receptor directly blocks osteoclast formation in vitro and in vivo. Arthritis & RheumatismVolume 60, Issue 9. https://doi.org/10.1002/art.24781.

Balandraud N and Roudier J. (2018). Epstein-Barr virus and rheumatoid arthritis. Joint Bone Spine. 2018 Mar;85(2):165-170. http://doi.org/10.1016/j.jbspin.2017.04.011.

Balsa A, Cabeson A, Orozco G, et al. (2010). Influence of HLA DRB1 alleles in the susceptibility of rheumatoid arthritis and the regulation of antibodies against citrullinated proteins and rheumatoid factor. Arthritis Research & Therapy. https://doi.org/10.1186/ar2975.

Bang SY, Lee HS, Lee KW, et al. (2013). Interaction of HLA-DRB1*09:01 and *04:05 with smoking suggests distinctive mechanisms of rheumatoid arthritis susceptibility beyond the shared epitope. J Rheumatol. (2013 Jul;40(7):1054-62. http://doi.org/10.3899/jrheum.121280.

Barbarroja N, Arias De La Rosa I, Pena M, et al. (2016). Effect of systemic and local inflammation on the insulin resistance and glucose/lipid metabolism in rheumatoid arthritis: Humans and CIA mouse models. Arthritis Rheumatol. 68(suppl 10).

Bartok B and Firestein GS. (2010). Fibroblast-like synoviocytes: key effector cells in rheumatoid arthritis. Immunol Rev. 2010 Jan;233(1):233-55. http://doi.org/10.1111/j.0105-2896.2009.00859.x.

Bartold MP, Marino V, Cantley M, et al. (2010). Effect of Porphyromonas gingivalis-induced inflammation on the development of rheumatoid arthritis. Journal of Clinical PeriodontologyVolume 37, Issue 5. https://doi.org/10.1111/j.1600-051X.2010.01552.x.

Begovich AB, Carlton VEH, Honigberg LA, 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 75:330–337. https://doi.org/10.1086/422827

Beighton P, Solomon L, Valkenburg HA. (1975). Rheumatoid arthritis in a rural South African Negro population. Ann Rheum Dis. 1975;34(2):136–41.

Bergström B, Carlsten H, and Ekwall A H. (2018). Methotrexate inhibits effects of platelet-derived growth factor and interleukin-1β on rheumatoid arthritis fibroblast-like synoviocytes. Arthritis research & therapy, 20(1), 49. http://doi.org/10.1186/s13075-018-1554-7.

Bjorkman PJ, Saper MA, Samraoui B, et al. (1987). The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature. 329:512–518.

Bongartz T, Cantaert T, Atkins SR, et al. (2007). Citrullination in extra-articular manifestations of rheumatoid arthritis. Rheumatology (Oxford);46(1):70-5. Epub 2006 Jun 16. https://doi.org/10.1093/rheumatology/kel202

Bose P, Simmons GL and Grant S. (2013). Cyclin-dependent kinase inhibitor therapy for hematologic malignancies. Expert Opin Investig Drugs. 22(6):723-38. http://doi.org/10.1517/13543784.2013.789859.

Bottini N and Firestein GS. (2013). Epigenetics in Rheumatoid Arthritis: A Primer for Rheumatologists. Curr Rheumatol Rep. 15: 372. https://doi.org/10.1007/s11926-013-0372-9.

Boyce EG, Rogan EL, Vyas D, et al. (2018). Sarilumab: Review of a Second IL-6 Receptor Antagonist Indicated for the Treatment of Rheumatoid Arthritis. Annals of Pharmacotherapy, Volume: 52 issue: 8, page(s): 780-791. https://doi.org/10.1177/1060028018761599.

Boyle AP, Hong EL, Hariharan M, et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22(9): 1790–1797. http://doi.org/10.1101/gr.137323.112.

Braun J. (2011). Methotrexate: optimizing the efficacy in rheumatoid arthritis. Therapeutic advances in musculoskeletal disease, 3(3), 151–158. http://doi.ord/10.1177/1759720X11408635.

Brennan-Olsen SH, Cook S, Leech MT, et al. (2017). Prevalence of arthritis according to age, sex and socioeconomic status in six low and middle income countries: analysis of data from the World Health Organization study on global AGEing and adult health (SAGE) Wave 1. BMC Musculoskelet Disord. 2017; 18: 271. Published online 2017 Jun 21. http://doi.org/10.1186/s12891-017-1624-z.

Brighton SW, de la Harpe AL, van Staden DJ, et al. (1988). The prevalence of rheumatoid arthritis in a rural African population. J Rheumatol 15(3):405–8.

Budzik GP, Colletti LM and Faltynek CR. (2000). Effects of methotrexate on nucleotide pools in normal T cells and the CEM T cell line. Life Sci. 2000;66(23):2297-307. https://doi.org/10.1016/S0024-3205(00)00559-2

Cantley MD, Haynes DR, Marino V, et al. (2011). Pre-existing periodontitis exacerbates experimental arthritis in a mouse model. J Clin Periodontol, 38(6):532–541. http://doi.org/10.1111/j.1600-051X.2011.01714.x.

Cao M, Guo M, Wu DQ, et al. (2018). Pharmacogenomics of Methotrexate: Current Status and Future Outlook. Curr Drug Metab. 2018;19(14):1182-1187. http://doi.org/10.2174/1389200219666171227201047.

Cates M and Hutchinson D. (2015). Cadmium the missing link between smoking and increased rheumatoid disease activity? Annals of the Rheumatic Diseases 2015;74:e1. http://doi.org/10.1136/annrheumdis-2014-206538.

Chan ESL and Cronstein BN. (2002). Molecular action of methotrexate in inflammatory diseases. Arthritis Research & Therapy20024:266. https://doi.org/10.1186/ar419.

Chang CC, Chow CC, Tellier LCAM, et al. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets, *Gigascience,* 4: 7. http://doi.org/10.1186/s13742-015-0047-8.

Chang X and Wang K. (2012). wANNOVAR: annotating genetic variants for personal genomes via the web Journal of Medical Genetics 2012;49:433-436. http://doi.org/10.1136/jmedgenet-2012-100918.

Chen JC, Wright K, Mavis JM, et al. (2016). An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. Genome Medicine, 8:43 https://doi.org/10.1186/s13073-016-0299-7.

Chen SH, Lue JH, Hsiao YJ, et al. (2018). Elevated galanin receptor type 2 primarily contributes to mechanical hypersensitivity after median nerve injury. PloS one, 13(6), e0199512. http://doi.ord/10.1371/journal.pone.0199512.

Choy EH, Isenberg DA, Garrood T, et al. (2002). Therapeutic benefit of blocking interleukin-6 activity with an anti-inter-leukin-6 receptor monoclonal antibody in rheumatoid arthritis: a randomized, double-blind, placebo-controlled, dose-escalation trial. Arthritis Rheum 46: 3143–3150. https://doi.org/10.1002/art.10623

Choy EHS, Smith CM, Farewell V, et al. (2008). Factorial randomised controlled trial of glucocorticoids and combination disease modifying drugs in early rheumatoid arthritis. Annals of the Rheumatic Diseases, 67:656-663.

Clarke GM, Carter KW, Palmer LJ, et al. (2007). Fine mapping versus replication in whole-genome association studies. AJHG. Volume 81, Issue 5, November 2007, Pages 995-1005. https://doi.org/10.1086/521952

Cojocaru M, Cojocaru IM, Silosi I, et al. (2010). Extra-articular Manifestations in Rheumatoid Arthritis. Maedica (Buchar), 5:286–29.1.

Cole BC and Griffiths MM. (1993). Triggering and exacerbation of autoimmune arthritis by the Mycoplasma arthritidis superantigen MAM. Arthritis Rheum, 36(7):994-1002.

Comuzzie AG, Cole SA, Laston SL, et al. (2012). Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS One. 7:e51954. http://doi.org/10.1371/journal.pone.0051954.

Cooper GM, Stone EA, Asimenos G, et al. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15: 901-913. http://doi.org/10.1101/gr.3577405.

Cronstein BC. (2005). Low-dose Methotrexate: A mainstay in the treatment of rheumatoid arthritis. Pharmacological Reviews, 57 (2) 163-172. https://doi.org/10.1124/pr.57.2.3.

Cronstein BN and Sitkovsky M. (2017). Adenosine and adenosine receptors in the pathogenesis and treatment of rheumatic diseases Nat Rev Rheumatol, 13: pp. 41-51. http://doi.org/10.1038/nrrheum.2016.178.

Cronstein BN. (1996). Molecular therapeutics. Methotrexate and its mechanism of action. Arthritis & amp; Rheumatism Volume 39, Issue 12. Pp1951-1960. http://doi.org/10.1002/art.1780391203.

Cronstein BN. (1997). The Mechanism Of Action Of Methotrexate. Rheumatic Disease Clinics of North America, volume 23, Issue 4, 1, Pages 739-755. https://doi.org/10.1016/S0889-857X(05)70358-6

Cross M, Smith E, Hoy D, et al. (2014). The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. Ann Rheum Dis. 73(7):1323-30. http://doi.org/10.1136/annrheumdis-2013-204763.

Crostein BN. 2007. Interleukin-6 – a key mediator of systemic and local symptoms in rheumatoid arthritis. Bull NYU Hosp J Dis, 65(Suppl 1):S11–S15.

Crowson CS, Matteson EL, Myasoedova E, et al. (2011). The lifetime risk of adult-onset rheumatoid arthritis and other inflammatory autoimmune rheumatic diseases. Arthritis Rheum, 63(3):633-9. http://doi.org/10.1002/art.30155.

Cutolo M, Sulli A, Pizzorni C, et al. 2001. Anti-inflammatory mechanism of methotrexate in rheumatoid arthritis. Ann. Rheum. Dis. 60:729-735. doi: 10.1136/ard.60.8.729

Da Rocha Sobrinho HM, Jerach R, da Silva NA, et al. (2011). Mycoplasmal lipid-associated membrane proteins and Mycoplasma arthritidis mitogen recognition by serum antibodies from patients with rheumatoid arthritis. Rheumatol Int, 31(7):951-7. http://doi.org/10.1007/s00296-010-1612-1.

Danila MI, Laufer VA, Reynolds RA, et al. (2017). Dense Genotyping of Immune-Related Regions Identifies Loci for Rheumatoid Arthritis Risk and Damage in African Americans. Mol Med. 2017 Sep;23:177-187. http://doi.org/10.2119/molmed.2017.00081.

De Bakker PI, McVean G, Sabeti PC, et al. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC, Nat. Genet. Vol. 38 (pg. 1166-1172). http://doi.org/10.1038/ng1885.

DeChristopher LR, Uribari J, and Tucker KL. (2016). Intake of high-fructose corn syrup sweetened soft drinks, fruit drinks and apple juice is associated with prevalent arthritis in US adults, aged 20-30 years. Nutr Diabetes, 7;6:e199. http://doi.org/10.1038/nutd.2016.7.

Deighton CM, Walker DJ, Griffiths ID, et al. (1989). The contribution of HLA to rheumatoid arthritis. Clin Genet, 36(3):178-82.

Dervieux T, Furst D, Lein DO, et al. (2005). Pharmacogenetic and metabolite measurements are associated with clinical status in patients with rheumatoid arthritis treated with methotrexate: Results of a multicentred cross sectional observational study. Ann Rheum Dis 64: 1180–1185. doi:10.1136/ard.2004.033399

Dervieux T, Greenstein N, Kremer J. (2006). Pharmacogenomic and metabolic biomarkers in the folate pathway and their association with methotrexate effects during dosage escalation in rheumatoid arthritis. Arthritis & RheumatismVolume 54, Issue 10. https://doi.org/10.1002/art.22129.

Di Guiseppe D, Discacciati A, Orsini N, et al. (2014). Cigarette smoking and risk of rheumatoid arthritis: a dose-response meta-analysis. Arthritis Res Ther, 5;16(2):R61. http://doi.org/10.1186/ar4498.

Diaz-Gallo L, Ramsköld D, Shchetynsky K, et al. (2018). Systematic approach demonstrates enrichment of multiple interactions between non-HLA risk variants and *HLA-DRB1* risk alleles in rheumatoid arthritis Annals of the Rheumatic Diseases, 77**:**1454-1462. http://doi.org/10.1136/annrheumdis-2018-213412.

Dilthey A, Leslie S, Moutsianas L, et al. (2013). Multi-Population Classical HLA Type Imputation. PloS Comput Biol 9(2): e1002877. https://doi.org/10.1371/journal.pcbi.1002877.

Ding B, Padyukov L, Lundstrom E, 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(1): 30–38. http://doi.org/10.1002/art.24135.

Dowman B, Campbell RM, Zgaga L, et al. (2012). Estimating the burden of rheumatoid arthritis in Africa: A systematic analysis. J Glob Health, 2(2):020406. http://doi.org/10.7189/jogh.02.020406.

Drosos AA, Alamanos I, Voulgari PV, et al. (1997). Epidemiology of adult rheumatoid arthritis in northwest Greece 1987–1995. J. Rheumatol. 24: 2129-2133.

Duncan L, Shen H, Gelaye B, et al. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nature Communication, 10:3328. https://doi.org/10.1038/s41467-019-11112-0.

Edgar R, Domrachev M, Lash AE. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Research, Volume 30, Issue 1:1 Pages 207–210, https://doi.org/10.1093/nar/30.1.207.

Edno L, Bressolle F, Gomeni R, et al. (1996). Total and free methotrexate pharmacokinetics in rheumatoid arthritis patients. her Drug Monit. 18(2):128-34. http://doi.org/10.1093/rheumatology/34.5.421

Edwards SL, Beesley J, French JD, et al. (2013). Beyond GWAS: Illuminating the Dark Road from Association to Function. The American Journal of Human Genetics, 93: 779–797. http://doi.org/10.1016/j.ajhg.2013.10.012.

Elshabrawy HA, Chen Z, Volin MV, et al. (2015). Angiogenesis, 18: 433. https://doi.org/10.1007/s10456-015-9477-2

Elshafie AI, Elkhalisa AD, Elbargir S, et al. (2016). Active Rheumatoid Arthritis in Central Africa: A Comparative Study Between Sudan and Sweden. J Rheumatol, 43(10):1777-1786. http://doi.org/10.3899/jrheum.160303.

Emery P, Pope JE, Kruger K, et al. (2018). Efficacy of Monotherapy with Biologics and JAK Inhibitors for the Treatment of Rheumatoid Arthritis: A Systematic Review. Adv Ther, 35: 1535. https://doi.org/10.1007/s12325-018-0757-2.

Entezami P, Fox DA, Claham PJ, et al. (2011). Historical Perspective on the Etiology of Rheumatoid Arthritis. Hand Clin. 2011 Feb; 27(1): 1–10. doi: 10.1016/j.hcl.2010.09.006.

Eyre S, Bowes J, Diogo D, et al. (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nature genetics, 44(12), 1336–1340. http://doi.org/10.1038/ng.2462.

Fadista J, Oskolkov N, Hansson O, et al. (2017). LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals, Bioinformatics, Volume 33, Issue 4, 15, Pages 471–474. https://doi.org/10.1093/bioinformatics/btv602.

Farh KK, Marson A, Zhu J, et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature, 518:337–343. https://doi.org/10.1038/nature13835.

Finckh A, Liang MH, van Herckenrode CM, et al. (2006). Long-term impact of early treatment on radiographic progression in rheumatoid arthritis: a meta-analysis. Arthritis Rheum20065586472. http://doi.org/10.1002/art.22353.

Finckh A, Choi HK, Wolfe F. (2006). Progression of radiographic joint damage in different eras: trends towards milder disease in rheumatoid arthritis are attributable to improved treatment. Annals of the Rheumatic Diseases, 65:1192-1197. http://doi.org/ 10.1136/ard.2005.049338.

Firestein GS. (2003). Evolving concepts of rheumatoid arthritis. Nature, 423: 356–361. http://doi.org/ 10.1038/nature01661.

Firestein GS. (2017). Etiology and Pathogenesis of Rheumatoid Arthritis. Kelley and Firestein's Textbook of Rheumatology, 1115–1166.e7. http://doi.org/10.1016/b978-0-323-31696-5.00069-3.

Forbes SA and Trowsdale J. (1999). The MHC quarterly report. Immunogenetics, 50:152-159. http://doi.org/10.1007/s002510050591.

Foulquier C, Sebbag M, Clavel C, et al. (2007). Peptidyl arginine deiminase type 2 (PAD-2) and PAD-4 but not PAD-1, PAD-3, and PAD-6 are expressed in rheumatoid arthritis synovium in close association with tissue inflammation. Arthritis & Rheumatism, 56(11). https://doi.org/10.1002/art.22983.

Fransen J and van Riel PLCM. (2005). The disease Activity score and the EULAR response criteria. Clin Exp Rheumatol. 23 (Suppl. 39): S93-S99. http://doi.org/10.1016/j.rdc.2009.10.001.

Freudenberg J, Lee H, Han B, et al. (2011). Genome-Wide Association Study of Rheumatoid Arthritisin Koreans: Population-Specific Loci as Well as Overlap With European Susceptibility Loci. ARTHRITIS & RHEUMATISM, 63(4): pp 884–893. http://doi.org/10.1002/art.30235.

Friedman B and Cronstein B. (2018). Methotrexate mechanism in treatment of rheumatoid arthritis. Joint Bone Spine, 86(3): 301-307. https://doi.org/10.1016/j.jbspin.2018.07.004.

Frisell T, Holmqvist M, Kallberg H, et al. (2013). Familial risks and heritability of rheumatoid arthritis: role of rheumatoid factor/anti–citrullinated protein antibody status, number and

type of affected relatives, sex, and age. Arthritis Rheum, 65: 2773– 82. http://doi.org/10.1002/art.38097.

Fujiwa S and Takei M. (2015). Epstein-Barr virus and autoimmune diseases. Clinical and Experimental Neuroimmunology 6 (Suppl. 1): 38–48. https://doi.org/10.1111/cen3.12263.

Furukawa H, Oka S, Shimada K, Hashimoto A, et al. (2018). Independent association of HLA-DPB1*02:01 with rheumatoid arthritis in Japanese populations. PloS one, 13(9), e0204459. http://doi.org/10.1371/journal.pone.0204459.

Gabriel SB, Schaffner SF, Nguyen H, et al. (2002). The Structure of Haplotype Blocks in the Human Genome. Science, 296(5576): 2225-2229. http://doi.org/10.1126/science.1069424.

Gabriel SE, Crowson CS, Kremers HM, et al. (2003). Survival in rheumatoid arthritis: a population-based analysis of trends over 40 years. Arthritis Rheum. 48(1):54-8. http://doi.org/10.1002/art.10705.

Garrod AB. (1859). Treatise on nature and treatment of gout and rheumatic gout. London: Walton and Maberly.

Gilani ST, Khane DA, Khane FA, et al. (2012). Adverse effects of low dose methotrexate in rheumatoid arthritis patients. J Coll Physicians Surg Pak, 22(2):101-4. http://doi.org/02.2012/JCPSP.101104.

Glossop JR, Dawes PT, and Mattey DL. (2006). Association between cigarette smoking and release of tumour necrosis factor α and its soluble receptors by peripheral blood mononuclear cells in patients with rheumatoid arthritis, Rheumatology, 45(10): Pages 1223–1229, https://doi.org/10.1093/rheumatology/kel094.

Gonzalez-Gay MA, Gonzalez-Juanatey C, Lopez-Diaz MJ, et al. (2007). HLA-DRB1 and persistent chronic inflammation contribute to cardiovascular events and cardiovascular mortality in patients with rheumatoid arthritis. Arthritis Rheum, 57:125–32. http://doi.org/10.1002/art.22482.

Gough AG, Faint J, Salmon M, et al. (1994). Genetic typing of patients with inflammatory arthritis at presentation can be used to predict outcome. Arthritis & Rheumatism, 37(8). https://doi.org/10.1002/art.1780370809.

Govind N, Choudhury A, Hodkinson B, et al. (2014). Immunochip identifies novel, and replicates known, genetic risk loci for rheumatoid arthritis in black South Africans. Molecular medicine (Cambridge, Mass.), 20(1): 341–349. http://doi.org/10.2119/molmed.2014.00097.

Govind N, Reynolds RJ, Hodkinson B, et al. (2019). HLA-DRB1 Amino Acid Positions and Residues Associated with Antibody-positive Rheumatoid Arthritis in Black South Africans. J Rheumatol. 46(2):138-144. http://doi.org/10.3899/jrheum.180107.

Grant SF, et al. (2001). The inheritance of rheumatoid arthritis in Iceland. Arthritis Rheum. 44, 2247–2254. http://doi.org/10.1002/1529-0131(200110)44:10<2247::aid-art387>3.0.co;2-y.

Gregersen PK, Amos CI, Lee AT, et al. (2009). REL, a member of the NF-κB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat Genet, 41(7): 820–823. http://doi.org/10.1038/ng.395.

Gregersen PK, Silver J, Winchester RJ. (1987). The shared epitope hypothesis – an approach to understanding the molecular genetics of rheumatoid arthritis susceptibility. Arthritis Rheum. 30:1205–1213. http://doi.org/10.1002/art.1780301102.

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. Nature, 550: 204–213. https://doi.org/10.1038/nature24277.

Guo Q, Wang Y, Xu Q, et al. (2018). Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. Bone Res. 6: 15. http://doi.org/10.1038/s41413-018-0016-9.

Gurdasani D, Carstensen T, et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature volume 517: pages 327–332. http://doi.org/10.1038/nature13997.

Han B, Diogo D, Eyre S, et al. (2014). Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity. Am J Hum Genet. 94(4):522-32. http://doi.org/10.1016/j.ajhg.2014.02.013. Epub 2014 Mar 20.

Han B, Diogo D, Eyre S, et al. (2014). Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity. Am J Hum Genet. 94(4):522-32. http://doi.org/10.1016/j.ajhg.2014.02.013. Epub 2014 Mar 20.

Harris ED. (1990). Rheumatoid arthritis: pathophysiology and implications for therapy. N Engl J Med. 322:1277-1289. http://doi.org/10.1056/NEJM199005033221805.

Hashiguchi M, Tsuru T, Miyawaki K, et al. (2016). Preliminary study for predicting better methotrexate efficacy in Japanese patients with rheumatoid arthritis. J Pharm Health Care Sci. 2:13. http://doi.org/10.1186/s40780-016-0047-6.

Hayer S, Bauer G, Willburger M, et al. (2016). Cartilage damage and bone erosion are more prominent determinants of functional impairment in longstanding experimental arthritis than synovial inflammation. Disease Models & Mechanisms 9, 1329-1338. http://doi.org/10.1242/dmm.025460.

Hazes JM, Dijkmans BA, Vandenbroucke JP, et al. (1990). Lifestyle and the risk of rheumatoid arthritis: cigarette smoking and alcohol consumption. Ann Rheum Dis. 49(12):980–2. http://doi.org/10.1136/ard.49.12.980.

Hedstrom AK, Stawiarz L, Klareskog L, et al. (2018). Smoking and susceptibility to rheumatoid arthritis in a Swedish population-based case-control study. Eur J Epidemiol. 33(4):415-423. http://doi.org/10.1007/s10654-018-0360-5.

Heliövaara M, Aho K, Knekt P, et al. (2000). Coffee consumption, rheumatoid factor, and the risk of rheumatoid arthritis. Annals of the Rheumatic Diseases, 59**:**631-635. http://doi.org/10.1136/ard.59.8.631.

Hellvard A, Zeitlmann L, Heiser U, et al. (2016). Inhibition of CDK9 as a therapeutic strategy for inflammatory arthritis. *Scientific Reports* volume 6, Article number: 31441. https://doi.org/10.1038/srep31441.

Helmick CG, Felson DT, Lawrence RC, et al. (2008). Estimates of the prevalence of arthritis and other rheumatic conditions in the United Stated. Part I. Arthritis Rheum. 58(1):15-25. http://doi.org/10.1002/art.23177.

Hifinger M, Putrik P, RamiroS, et al. (2015). In rheumatoid arthritis, country of residence has an important influence on fatigue: results from the multinational COMORA study. Rheumatology, 55(4): 735–744, https://doi.org/10.1093/rheumatology/kev395.

Hitchon F, Ferucci ED, et al. (2010). Antibodies to porphyromonas gingivalis are associated with anticitrullinated protein antibodies in patients with rheumatoid arthritis and their relatives. J Rheumatol. 37(6):1105-12. http://doi.org/10.3899/jrheum.091323.

Hodkinson B, Musenge E, Ally M, et al. (2012). Response to traditional disease-modifying anti-rheumatic drugs in indigent South Africans with early rheumatoid arthritis. Clin Rheumatol. 31(4):613-9. http://doi.org/10.1007/s10067-011-1900-5.

Holers VM. (2013). Autoimmunity to citrullinated proteins and the initiation of rheumatoid arthritis. Curr Opin Immunol. 25(6):728-35. http://doi.org/10.1016/j.coi.2013.09.018.

Hong H, Zeng Y, Jian W, et al. (2018). CDK7 inhibition suppresses rheumatoid arthritis inflammation *via* blockage of NF-κB activation and IL-1β/IL-6 secretion. J Cell Mol Med. 22(2): 1292–1301. http://doi.org/10.1111/jcmm.13414.

Honorati MC, Neri S, Cattini L, et al. (2006). Interleukin-17, a regulator of angiogenic factor release by synovial fibroblasts. Osteoarthritis and Cartilage, 14(4): 345-352. http://doi.org/ 10.1016/j.joca.2005.10.004.

Hou S, Gao GP, Zhang XJ, et al. (2013). PADI4 polymorphisms and susceptibility to rheumatoid arthritis: a meta-analysis. Mod Rheumatol. 23:50–60. http://doi.org/10.1007/s10165-012-0639-4.

Houtman M, Shchetynsky K, Chemin K et al. (2018). T cells are influenced by a long non-coding RNA in the autoimmune associated PTPN2 locus. Journal of Autoimmunity, 90: 28-38. https://doi.org/10.1016/j.jaut.2018.01.003.
https://doi.org/10.1007/s00439-016-1734-y

Hudson M, Bernatsky S, Colmegna I, et al. (2017). Novel insights into systemic autoimmune rheumatic diseases using shared molecular signatures and an integrative analysis, Epigenetics, 12:6, 433-440, http://doi.org/10.1080/15592294.

Hughes LB, Beasley TM, Patel H, et al. (2006). Racial or ethnic differences in allele frequencies of single-nucleotide polymorphisms in the methylenetetrahydrofolate reductase gene and their influence on response to methotrexate in rheumatoid arthritis. Ann Rheum Dis. 65: 1213-1218. http://doi.org/10.1136/ard.2005.046797.

Hughes LB, Morrison D, Kelley JM, et al. (2008). The HLA-DRB1 shared epitope is associated with susceptibility to rheumatoid arthritis in African Americans through European genetic admixture. Arthritis Rheum. 58(2):349-58. http://doi.org/10.1002/art.23166.

Hughes LB, Reynolds RJ, Brown EE, 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(12):3547-53. http://doi.org/10.1002/art.27732.

Hutchinson D, Shepstone L, Moots R, et al. (2001). Heavy cigarette smoking is strongly associated with rheumatoid arthritis (RA), particularly in patients without a family history of RA. Ann Rheum Dis, 60: 223-227. http://doi.org/10.1136/ard.60.3.223.

Irfan S, Rani A, Riaz N, et al. (2017). Comparative Evaluation of Heavy Metals in Patients with Rheumatoid Arthritis and Healthy Control in Pakistani Population. Iran J Public Health, 46(5): 626–633.

Isaacs JD. (2010). The changing face of rheumatoid arthritis: sustained remission for all? Nat Rev Immunol. 10(8):605-11. http://doi.org/10.1038/nri2804.

Jawaheer D, Li W, Graham RR, et al. (2002). Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. Am J Hum Genet. 71(3):585-94. https://doi.org/10.1086/342407.

Jenko B, lusa L, and Tomsic M. (2017). Clinical–pharmacogenetic predictive models for MTX discontinuation due to adverse events in rheumatoid arthritis. The Pharmacogenomics Journal, 17: 412–418. https://doi.org/10.1038/tpj.2016.36.

Jiang L, Jiang D, Han Y, et al. (2018). Association of HLA-DPB1 polymorphisms with rheumatoid arthritis: A systemic review and meta-analysis. Int J Surg. 52:98–104. http://doi.org/10.1016/j.ijsu.2018.01.046.

Jones MA, Silman AJ, Whiting S, et al. (1996). Occurrence of rheumatoid arthritis is not increased in the first degree relatives of a population based inception cohort of inflammatory polyarthritis. Ann Rheum Dis. 55: 89–93.

Jones HJ, Hubbard L, Mitchell RE, et al. (2018). Association of genetic rick of rheumatoid arthritis with cognitive and psychiatric phenotypes across childhood and adololescence. JAMA network Open, 2(6):e196118. http://doi.org/10.1001/jamanetworkopen.2019.6118

Joo SH, Lee J, Hutchinson D, et al. (2019). Prevalence of rheumatoid arthritis in relation to serum cadmium concentrations: cross-sectional study using Korean National Health and Nutrition Examination Survey (KNHANES) data BMJ, 9:e023233. http://doi.org/10.1136/bmjopen-2018-023233.

Joshi VR. (2012). Rheumatology, Past, Present and Future. The Journal of the Association of Physicians of India, 60(1):21-4.

Julia A, Ballina J, Canete JD, 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(8):2275-86. http://doi.org/10.1002/art.23623.

Kalla AA and Tikly M. (2003). Rheumatoid arthritis in the developing world. Best Pract Res Clin Rheumatol. 17:863–75. http://doi.org/10.1016/S1521-6942(03)00047-0.

Kallberg H, Ding B, Padyukov 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(3): 508–511. http://doi.org/10.1136/ard.2009.120899.

Kalpakcioglu B and Senel K. (2008). The interrelation of glutathione reductase, catalase, glutathione peroxidase, superoxide dismutase, and glucose-6-phosphate in the pathogenesis of rheumatoid arthritis. Clinical Rheumatology, 27(2): 141–145. http://doi.org/ 10.1007/s10067-007-0746-3.

Kaltsonoudis E, Papagoras C and Drosos AA. (2012). Current and future role of methotrexate in the therapeutic armamentarium for rheumatoid arthritis. Int. J. Clin. Rheumatol. 7(2): 179–189. http://doi.org/10.2217/IJR.12.10.

Kaminsky ZA, Tang T, Wang SC, et al. (2009). DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet. 41(2):240–5. http://doi.org/10.1038/ng.286.

Karlson EW, Mandl LA, Aweh GN, et al. (2003). Coffee consumption and risk of rheumatoid arthritis. Arthritis Rheum. 48(11): 3055–3060. http://doi.org/10.1002/art.11306.

Kawahito Y, Ichinose S, Sano H, et al. (2008). Mycoplasma fermentans glycolipid-antigen as a pathogen of rheumatoid arthritis. Biochem Biophys Res Commun. 369(2):561-6. http://doi.org/10.1016/j.bbrc.2008.02.079.

Kay J, and Upchurch KS. (2012). ACR/EULAR 2010 rheumatoid arthritis classification criteria. Rheumatology, 51(6): vi5–vi9, https://doi.org/10.1093/rheumatology/kes279.

Kettunen J, Tukiainnen T, Sarin AP, et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat Genet. 29:44(3):269-276. http://doi.org/10.1038/ng.1073.

Keyszer G. et al. (1998). Differential expression of cathepsins B and L compared with matrix metalloproteinases and their respective inhibitors in rheumatoid arthritis and osteoarthritis: a parallel investigation by semiquantitative reverse transcriptase-polymerase chain reaction and immunohistochemistry. Arthritis Rheum. 41: 1378–1387. http://doi.org/10.1002/1529-0131(199808)41:8<1378::AID-ART6>3.0.CO;2-J.

Khera AV, Chaffin M, Aragam KG, et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. 50: 1219–1224.

Kichaev G, Bhatia G, Loh PR, et al. (2019). Leveraging polygenic functional enrichment to improve GWAS power. Am J Hum Genet. 104(1):65-75. http://doi.org/10.1016/j.ajhg.2018.11.008.

Kidd BA, Ho PP, Sharpe O, et al. (2008). Epitope spreading to citrullinated antigens in mouse models of autoimmune arthritis and demyelination. Arthritis Research & Therapy, 10:R119. https://doi.org/10.1186/ar2523.

Kim HJ and Berek C. (2000). B cells in rheumatoid arthritis. Arthritis Res. 2(2):126-31. http://doi.org/ 10.1186/ar77.

Kinloch AJ, Alzabin S, Brintnell W, et al. (2011). Immunization with Porphyromonas gingivalis enolase induces autoimmunity to mammalian α-enolase and arthritis in DR4-IE-transgenic mice. Arthritis Rheum. 63(12):3818-23. http://doi.org/10.1002/art.30639.

Kivity S, Zafrir Y, Loebstein R, et al. (2014). Clinical characteristics and risk factors for low dose methotrexate toxicity: a cohort of 28 patients. Autoimmun Rev. 13(11):1109-13. http://doi.org/10.1016/j.autrev.

Klareskog L, Stolt P, Lundberg K, et al. (2005). 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 & Rheumatism, 54(1). https://doi.org/10.1002/art.21575.

Klareskog L, Stolt P, Lundberg K, 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(1):38-46. http://doi.org/ 10.1002/art.21575.

Klatt T, Ouyang Q, Flad T, et al. (2005). Expansion of Peripheral CD8+ CD28– T Cells in Response to Epstein-Barr Virus in Patients with Rheumatoid Arthritis. The Journal of Rheumatology, 32:2.

Klein J and Hořejší V. (1998). Immunology. 2nd ed. Oxford, England: Blackwell Scientific.

Klein J and Sato A. (2000). The HLA System. N Engl J Med, 343:702-709 http://doi.org/10.1056/NEJM200009073431006.

Kobold N, Jenko B, Tomsic M, et al. (2019). ADORA2A Polymorphisms Influence Methotrexate Adverse Events in Rheumatoid Arthritis. The Israel Medical Association Journal, 5(21):333-338.

Kochi, Y. et al. (2010). A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. Nat. Genet. 42: 515–519. https://doi.org/10.1038/ng.583.

Kohler JR, Guennel T and Marshall SL. (2014). Analytical strategies for discovery and replication of genetic effects in pharmacogenomic studies. Pharmacogenomics and personalized medicine, 7:217-225**.** http://doi.org/10.2147/PGPM.S66841.

Konig M, Fabusleme L, Reinholdt J, et al. (2016). Aggregatibacter actinomycetemcomitans-induced hypercitrullination links periodontal infection to autoimmunity in rheumatoid arthritis. Sci Transl Med, 8:369ra176. http://doi.org/10.1126/scitranslmed.aaj1921.

Konttimen Y, Ceponis A, Takagi A, et al. (1998). New collagenolytic enzymes/cascade identified at the pannus-hard tissue junction in rheumatoid arthritis: destruction from above. Matrix Biology, 17(8–9): 585-601. https://doi.org/10.1016/S0945-053X(98)90110-X.

Koumantaki, Y. et al. (1997). Family history as a risk factor for rheumatoid arthritis: a case–control study. J. Rheumatol. 24: 1522–1526.

Kremer JM. (2006). Methotrexate pharmacogenomics. Ann Rheum Dis. 65(9): 1121–1123. http://doi.org/10.1136/ard.2006.051789.

Kumar P, Leong J, Saidin S, et al. (2018). THU0030 Molecular mechanisms of autophagic memory in pathogenic t cells in human rheumatoid arthritis. Annals of the Rheumatic Diseases, 77:242-243. http://doi.org/10.1136/annrheumdis-2018-eular.6060.

Kumar V, Westra HJ, Karjalainen J, et al. (2013). Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PloS Genet. 9(1):e1003201. http://doi.org/10.1371/journal.pgen.1003201.

Kung TN, Dennis J, Ma Y, et al. (2014). RFC1 80G>A is a genetic determinant of methotrexate efficacy in rheumatoid arthritis: a human genome epidemiologic review and meta-analysis of observational studies. Arthritis Rheumatol. 66(5):1111-20. http://doi.org/10.1002/art.38331.

Lajas C, Abasolo L, Bellajdel B, et al. (2003). Costs and predictors of costs in rheumatoid arthritis: a prevalence-based study. Arthritis Rheum. 49: 64–70. http://doi.org/10.1002/art.10905.

Lamichhane D, Collins C, Constatinesca F, et al. (2019). Coffee and Tea Consumption in Relation to Risk of Rheumatoid Arthritis in the Women's Health Initiative Observational Cohort. J Clin Rheumatol. 25: 127–132. http://doi.org/10.1097/RHU.0000000000000788.

Lang R, Gundlach AL, Holmes FE, et al. 2015. Physiology, Signaling, and Pharmacology of Galanin Peptides and Receptors: Three Decades of Emerging Diversity. Pharmacological Reviews, 67(1): 118-175. https://doi.org/10.1124/pr.112.006536.

Langdon K and Haleagrahara N. (2018). Regulatory T-cell dynamics with abatacept treatment in rheumatoid arthritis. Nt Rev Immunol. 37(4):206-214. http://doi.org/10.1080/08830185.

Larochelle S, Amat R, Clover-Cutter K, et al. (2012). Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II. Nat Struct Mol Biol. 19(11):1108-15. http://doi.org/10.1038/nsmb.

Larsen JM. (2017). The immune response to Prevotella bacteria in chronic inflammatory disease. Immunology. 151(4): 363–374. Published online 2017 Jun 20. http://doi.org/10.1111/imm.12760.

Laufer VA, Tiwari HK, Reynolds RJ, et al. (2019). Genetic Influences on Susceptibility to Rheumatoid Arthritis in African-Americans. Hum Mol Genet. http://doi.org/10.1093/hmg/ddy395.

Lee H, Korman BD, Le JM, et al. (2009). Genetic risk factors for rheumatoid arthritis differ in Caucasian and Korean populations. Arthritis & Rheumatism, 60(2). https://doi.org/10.1002/art.24245.

Lee HS, Lee AT, Criswell LA, 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(5-6):293-300. http://doi.org/10.2119/2007-00123.Lee.

Lee Y, Bae SC and Song GZ. (2016). Association of the ABCB1 C3435T polymorphism with responsiveness to and toxicity of DMARD in rheumatoid arthritis. Rheumatol. 75: 707. https://doi.org/10.1007/s00393-015-1618-x.

Lee YH and Bae SC. (2016). Association of the ATIC 347 C/G polymorphism with responsiveness to and toxicity of methotrexate in rheumatoid arthritis: a meta-analysis. Rheumatol Int 36: 1591. https://doi.org/10.1007/s00296-016-3523-2.

Lee YH and Song GG. (2010). Associations between the C677T and A1298C Polymorphisms of MTHFR and the Efficacy and Toxicity of Methotrexate in Rheumatoid Arthritis. Clin. Drug Investig. 30: 101. https://doi.org/10.2165/11531070-000000000-00000.

Lee YH, Woo J, Choi SJ, et al. (2010). Association between the rs7574865 polymorphism of STAT4 and rheumatoid arthritis: a meta-analysis. Rheumatol Int. 30:661–666. http://doi.org/10.1007/s00296-009-1051-z.

Lefevre S, Knedla A, Tennie C, et al. (2009). Synovial fibroblasts spread rheumatoid arthritis to unaffected joints. Nature Medicine, 15: 1414–1420. http://doi.org/10.1038/nm.2050.

Lefevre S, Meier FMP, Neumann E, et al. (2015). Role of Synovial Fibroblasts in Rheumatoid Arthritis. Current Pharmaceutical Design, 21(2):130-141(12). http://doi.org/ 10.2174/1381612820666140825122036.

Lekpa FK, Ndongo S, Tiendrebeogo J, et al. (2012). Rheumatoid arthritis in Senegal: a comparison between patients coming from rural and urban areas, in an urban tertiary health care center in Senegal. Clin Rheumatol. 31(11):1617-20. http://doi.org/10.1007/s10067-012-2054-9.

Lewis CM and Knight J. (2012). Introduction to Genetic Association Studies. Genetics of Complex Human Diseases (ed. Al-Chalabi and Almasy). CSHL Press, Cold Spring Harbor, NY, USA, 2009. Cold Spring Harbor Protoc. http://doi.org/10.1101/pdb.top068163.

Li X, Hu M, Li W, et al. (2016). The association between reduced folate carrier-1 gene 80G/A polymorphism and methotrexate efficacy or methotrexate related-toxicity in rheumatoid arthritis: a meta-analysis. Int Immunopharmacol. 38:8–15. http://doi.org/10.1016/j.intimp.2016.05.012.

Lim J, and Kim K. (2019). Genetic variants differentially associated with rheumatoid arthritis and systemic lupus erythematosus reveal the disease-specific biology. Scientific Reports 9, Article number: 2739. https://doi.org/10.1038/s41598-019-39132-2.

Lin L, Chen Y, Xiao Z, et al. (2007). The association of HLA-DRB1 alleles with rheumatoid arthritis in the Chinese Shantou population: a follow-up study. Biochemistry and Cell Biology, 85(2): 227-238. https://doi.org/10.1139/o06-204.

Ling SF, Viatte S, Lunt M, et al. (2016). HLA–DRB1 Amino Acid Positions 11/13, 71, and 74 Are Associated With Inflammation Level, Disease Activity, and the Health Assessment Questionnaire Score in Patients With Inflammatory Polyarthritis. Arthritis Rheumatol. 68(11): 2618–2628. http://doi.org/10.1002/art.39780.

Linos A, Kaklamani VG, Kaklamani E, et al. (1999). Dietary factors in relation to rheumatoid arthritis: a role for olive oil and cooked vegetables? Am J Clin Nutr. 70:1077–82. http://doi.org/10.1093/ajcn/70.6.1077.

Lins TC, Vieira RG, Grattapaglia D, et al. (2010). Allele and haplotype frequency distribution in PTPN22 gene across variable ethnic groups: Implications for genetic association studies for autoimmune diseases. Autoimmunity, 43: 308–316. http://doi.org/10.3109/08916930903405883.

Liu K, Tan LJ, Wang P, et al. (2017). Functional relevance for associations between osteoporosis and genetic variants. PloS one, 12(4), e0174808. http://doi.org/10.1371/journal.pone.0174808.

López-Rodríguez R, Ferreiro-Iglesias A, Lima A, Bernardes M, et al. (2018). Replication study of polymorphisms associated with response to methotrexate in patients with rheumatoid arthritis. Scientific reports, 8(1): 7342. https://doi.org/10.1038/s41598-018-25634-y.

Lu B, Solomon DH, Constenbader KH, et al. (2014). Alcohol consumption and risk of incident rheumatoid arthritis in women: a prospective study. Arthritis Rheumatol. 66(8):1998-2005. http://doi.org/10.1002/art.38634.

Lunemann JD, Frey O, Eidner T, et al. (2008). Increased Frequency of EBV-Specific Effector Memory CD8+ T Cells Correlates with Higher Viral Load in Rheumatoid Arthritis. J Immunol. 81(2): 991–1000. http://doi.org/10.4049/jimmunol.181.2.991.

MacGregor AJ, Snieder H, Rigby AS, et al. (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum. 43:30–7. 2. http://doi.org/10.1002/1529-0131(200001)43:1<30::AID-ANR5>3.0.CO;2-B.

MacGregor AJ. (1995). Classification criteria for rheumatoid arthritis. Baillieres Clin Rheumatol. 9(2):287-304. http://doi.org/10.1016/s0950-3579(05)80191-8.

Machold KP, Stamm TA, Eberl GJM, et al. (2002). Very recent onset arthritis–clinical, laboratory, and radiological findings during the first year of disease. The Journal of Rheumatology, 29 (11) 2278-2287.

Malemba JJ and Mbuyi-Muamba JM. (2008). Clinical and epidemiological features of rheumatic diseases in patients attending the university hospital in Kinshasa. Clin Rheumatol. 15:47–54. http://doi.org/10.1007/s10067-008-0865-5.

Malemba JJ, Mbuyi-Muamba JM, Mukaya J, et al. (2012). The epidemiology of rheumatoid arthritis in Kinshasa, Democratic Republic of Congo—a population-based study. Rheumatology (Oxford). 51(9):1644–7. http://doi.org/10.1093/rheumatology/kes092.

Martin AR, Kanai M, Kamatani Y, et al. (2019). Clinical use of current polygenic risk scores may exacerbate health disparaties. Nature Genetics, 51:584-591. https://doi.org/10.1038/s41588-019-0379-x.

Maresz KJ, Hellvard A, Stroka A, et al. (2013). Porphyromonas gingivalis Facilitates the development and progression of destructive arthritis through its unique bacteria peptidylarginine deiminase (PAD). PLoS Patholog. V.9(9). http://doi.org/10.1371/journal.ppat.1003627.

Martell RW, Du Toit ED and Kalla AA, et al. (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.

Martell RW, Stein M, Davis P, et al. (1990). The association between HLA and rheumatoid arthritis in Zimbabwean blacks. Tissue Angitens, 36: 125-126.

Martinez-Martinez RE, Abud-Mendoza C, Patino-Marin N, et al. (2009). Detection of periodontal bacterial DNA in serum and synovial fluid in refractory rheumatoid arthritis patients. J Clin Periodontol. 36:1004–10. http://doi.org/10.1111/j.1600-051X.2009.01496.x.

Mateen S, Moin S, Khan AQ, et al. (2016). Increased Reactive Oxygen Species Formation and Oxidative Stress in Rheumatoid Arthritis. PloS ONE 11(4): e0152925. https://doi.org/10.1371/journal.pone.015292.

Mattey DL, Dawes PT, Clarke S, et al. (2002). Relationship among the HLA–DRB1 shared epitope, smoking, and rheumatoid factor production in rheumatoid arthritis. Arthritis Care & Research, 47(4). https://doi.org/10.1002/art.10514.

McCarthy DJ, Humburg P, Kanapin A, et al. (2014). Choice of transcripts and software has a large effect on variant annotation. Genome Medicine, 6:26. https://doi.org/10.1186/gm543.

McDaniel DO, Alarcon GS, Pratt PW, et al. (1995). Most African-American patients with rheumatoid arthritis do not have the rheumatoid antigenic determinant (epitope). Ann Intern Med. 123(3):181-7. http://doi.org/10.7326/0003-4819-123-3-199508010-00004.

McIntosh LA, Marion MC, Sudman M, et al. (2017). Genome-Wide Association Meta-Analysis Reveals Novel Juvenile Idiopathic Arthritis Susceptibility Loci. Arthritis Rheumatol. 69(11):2222-2232. http://doi.org/10.1002/art.40216.

McLaren W, Gil L, Hunt SE, et al. (2016). The Ensembl Variant Effect Predictor. Genome Biology, 17:122. https://doi.org/10.1186/s13059-016-0974-4.

Medina-Gomez C, Kemp JP, Trajanoska K, et.al. (2018). Life-Course Genome-wide association study meta-analysis of total body BMD and assessment of age-specific effects. Am J Hum Genet. 4:102(1):88-102. http://doi.org/10.1016/j.ajhg.2017.12.005.

Meyer P W, Hodkinson B, Ally M, 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 research & therapy, 13(5), R160. http://doi.org/10.1186/ar3479.

Meyers OL, Daynes G, Beighton P. (1977). Rheumatoid arthritis in a tribal Xhosa population in the Transkei, Southern Africa. Ann Rheum Dis. 36(1):62–5. Epub 1977/02/01.

Mikuls TR, Cerhan JR, Criswell LA et al. (2002). Coffee, tea, and caffeine consumption and risk of rheumatoid arthritis: results from the Iowa Women's Health Study. Arthritis Rheum.

46(1): 83–91. https://doi.org/10.1002/1529-0131(200201)46:1<83::AID-ART10042>3.0.CO;2-D

Mikuls TR, Payne JB, Reinhardt RA, et al. (2009). Antibody responses to Porphyromonas gingivalis (P. gingivalis) in subjects with rheumatoid arthritis and periodontitis. Int Immunopharmacol. 9(1):38-42. http://doi.org/10.1016/j.intimp.2008.09.008.

Mikuls TR, Sayles H, Yu F, et al. (2010). Associations of cigarette smoking with rheumatoid arthritis in African Americans. Arthritis Rheum. 62(12): 3560–3568. http://doi.org/10.1002/art.27716.

Mikuls TR, Thiele Gm, Deane KD, et al. (2012). Porphyromonas gingivalis and disease-related autoantibodies in individuals at increased risk of rheumatoid arthritis. Arthritis Rheum. 64(11):3522-30. http://doi.org/10.1002/art.34595.

Miller SA, Dykes DD and Polesky HF. (1988). A simple salting out DNA extraction from paraffin procedure for extracüng DNA from human nucleated cells. Nucleic Acids Res. 16, 12-15.

Mirowsky JE, Dailey LA and Devlin RB. (2016). Differential expression of pro-inflammatory and oxidative stress mediators induced by nitrogen dioxide and ozone in primary human bronchial epithelial cells. Inhalation toxicology, 28(8), 374–382. http://doi.org/10.1080/08958378.2016.1185199.

Mody GM and Hammond MG. (1994). Differences in HLA-DR association with rheumatoid arthritis among migrant Indian communities in South Africa. Br J Rheumatol. 33(5):425-7.

Mody GM and Meyers OL. (1988). Resorptive arthropathy in rheumatoid arthritis. Journal of Rheumatology, 15 (1988), pp. 1075-1077.

Mody GM, Hammond MG and Naidoo PD. (1989). HLA association with rheumatoid arthritis in African blacks. Journal of Rheumatology, 16: 1326-1328.

Mody GM, Shaw J and Ramchurren A. (1988). Rheumatoid arthritis impact survey. South African Medical Journal, 74: 409-410.

Mody GM. (1995). Rheumatoid arthritis and connective tissue disorders: sub-Saharan Africa. Baillieres Clin Rheumatol. 9:31–44. http://doi.org/10.1016/S0950-3579(05)80141-4.

Mody GM. (2017). Rheumatology in Africa—challenges and opportunities. Arthritis Res Ther. 19: 49. http://doi.org/10.1186/s13075-017-1259-3.

Mohamed BM, Verma NK, Davies AM, et al. (2012). Citrullination of proteins: a common post-translational modification pathway induced by different nanoparticles in vitro and in vivo. Nanomedicine. 7:1181–95. http://doi.org/10.2217/nnm.11.177.

Moolenburgh JD, Valkenburg HA and Fourie PB. (1984). Rheumatoid arthritis in Lesotho. Annuls of the Rheumatic diseases. Ann Rheum Dis. 43:40–3. http://doi.org/10.1136/ard.43.1.40.

Morgan MD, Al-Shaarawy N, Martin S, et al. (2014). MTHFR functional genetic variation and methotrexate treatment response in rheumatoid arthritis: a meta-analysis. Pharmacogenomics, 15(4). https://doi.org/10.2217/pgs.13.235.

Morgan SL, Anderson AM, Hood SM, et al. (1997). Nutrient intake patterns, body mass index, and vitamin levels in patients with rheumatoid arthritis. Arthritis Care and Research, 10(1):9-17. http://doi.org/10.1002/art.1790100103.

Morvan J, Berthelot J, Devauchelle-Pensec V, et al. (2009). Changes Over Time in the Diagnosis of Rheumatoid Arthritis in a 10-year Cohort. The Journal of Rheumatology, 36(11): 2428-2434. http://doi.org/10.3899/jrheum.090072.

Moya P, Salazar J, Arranz J, et al. (2016). Methotrexate pharmacokinetic genetic variants are associated with outcome in rheumatoid arthritis patients. Pharmacogenomics. 17(1):25-9. http://doi.org/10.2217/pgs.15.150.

Muller IB, Hebing RF, Jansen G, et al. (2018). Personalized medicine in rheumatoid arthritis: methotrexate polyglutamylation. Personalized medicine in drug development and clinical practice. 3:331-334. http://doi.org/10.1080/23808993.2018.1517025.

Muller S, Hider SL, Raza K, et al. (2015). An algorithm to identify rheumatoid arthritis in primary care: a Clinical Practice Research Datalink study BMJ, 5:e009309. http://doi.org/10.1136/bmjopen-2015-009309.

Muller-Ladner U. et al. (1996). Synovial fibroblasts of patients with rheumatoid arthritis attach to and invade normal human cartilage when engrafted into SCID mice. Am. J. Pathol. 149, 1607–1615. https://doi.org/10.1080/23808993.2018.1517025.

Murphy D, Mattey D and Hutchinson D. (2017). Anti-citrullinated protein antibody positive rheumatoid arthritis is primarily determined by rheumatoid factor titre and the shared epitope rather than smoking per se. PloS ONE 12(7): e0180655. https://doi.org/10.1371/journal.pone.0180655.

Murphy D, Sinha-Royle E, Bellis K, et al. (2019). Nodular rheumatoid arthritis (RA): A distinct disease subtype, initiated by cadmium inhalation inducing pulmonary nodule formation and subsequent RA-associated autoantibody generation. Med Hypotheses. 122:48-55. http://doi.org/10.1016/j.mehy.2018.10.021.

Myasoedova E, Davis JM, Crowson CS, et al. (2010). Epidemiology of Rheumatoid Arthritis: Rheumatoid Arthritis and Mortality. Curr Rheumatol Rep 12: 379. https://doi.org/10.1007/s11926-010-0117-y.

Myozen K, Kochi Y, Okada Y, et al. (2012). Functional variants in NFKBIE and RTKN2 involved in activation of the NF-κB pathway are associated with rheumatoid arthritis in Japanese. PloS Genet. 8(9):e1002949. http://doi.org/10.1371/journal.pgen.1002949.

Naranjo A, Toloza S, Guimaraes da Silveira I, et al. (2010). Smokers and non-smokers with rheumatoid arthritis have similar clinical status: data from the multinational QUEST-RA database. Clin Exp Rheumatol. 28(6):820-7. Epub 2011 Jan 3.

Neogi T, Aletaha D, Silman AJ, et al. (2010). The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for

rheumatoid arthritis: phase 2 methodological report. Arthritis Rheum. 62: 2582-2591. http://doi.org/10.1002/art.27580.

Newkirk MM, Goldbach-Mansky R, Senior BW, et al. (2005). Elevated levels of IgM and IgA antibodies to Proteus mirabilis and IgM antibodies to Escherichia coli are associated with early rheumatoid factor (RF)-positive rheumatoid arthritis. Rheumatology (Oxford). 44(11):1433-41. http://doi.org/10.1093/rheumatology/kei036.

Newton, J., Harney, S., Wordsworth, B. et al. 2004. A review of the MHC genetics of rheumatoid arthritis. Genes Immun. 5: 151–157. https://doi.org/10.1038/sj.gene.6364045

Ngo ST, Steyn FJ, McCombe PA. (2014). Gender differences in autoimmune disease. Front Neuroendocrinol. 35 (3): 347-69. https://doi.org/10.1016/j.yfrne.2014.04.004.

Ohagwu KA, Olaosebikan H,Oba RB, et al. (2017). Pattern of rheumatoid arthritis in Nigeria; Study of patients from a Teaching Hospital. African Journal of Rheumatology, 5(2).

Okada Y, et al. (2012). Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet. 44:511–516. http://doi.org/10.1038/ng.2231.

Okada Y, Eyre S, Suzuki A, et al. (2019). Genetics of rheumatoid arthritis: 2018 status Annals of the Rheumatic Diseases, 78:446-453. http://doi.org/ 10.1136/annrheumdis-2018-213678.

Okada Y, Kim K, Han B, et al. (2014). Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. Hum Mol Genet. 23: 6916-6926. http://doi.org/10.1093/hmg/ddu387.

Orellana C, Saevarsdottir S, Klareskog L, et al. (2017). Oral contraceptives, breastfeeding and the risk of developing rheumatoid arthritis: results from the Swedish EIRA study. Ann Rheum Dis. 76(11):1845–52. http://doi.org/10.1136/annrheumdis-2017-211620.

Østensen M and Villiger PM. (2007). The remission of rheumatoid arthritis during pregnancy. Semin Immunopathol 29: 185. https://doi.org/10.1007/s00281-007-0072-5.

Padyukov L, Seielstad M, Ong RT, et al. (2011). A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. Ann Rheum Dis. 70:259–65. http://doi.org/10.1136/ard.2009.126821.

Pan W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genet Epidemiol. 35(4):211–216. http://doi.org/10.1002/gepi.20567.

Pappas DJ, Lizee A, Paunic V, et al. (2018). Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. The Pharmacogenomics Journal, 18: 367–376. http://doi.org/10.1038/tpj.2017.7.

Pastore S, Stocco G, Favretto D, et al. (2015). Genetic determinants for methotrexate response in juvenile idiopathic arthritis. Front Pharmacol. 6:52. Published 2015 Mar 23. http://doi.org/10.3389/fphar.2015.00052.

Patterson N, Price AL, and Reich D. (2006). Population structure and eigenanalysis. PloS Genet. 2: e190. http://doi.org/10.1371/journal.pgen.0020190.

Pedersen L, Permin H. (1988). Rheumatic disease, heavy-metal pigments, and the great masters. The Lancet, 331(8597): 1267-1269. http://doi.org/10.1016/s0140-6736(88)92082-x.

Peng SL. (2006). Fas (CD95)-related apoptosis and rheumatoid arthritis. Rheumatology (Oxford). 45(1):26-30. http://doi.org/10.1093/rheumatology/kei113.

Pianta A, Arvikar S, Strle K, et al. (2017). Evidence for Immune Relevance of Prevotella copri, a Gut Microbe, in Patients with Rheumatoid Arthritis. Arthritis Rheumatol. 69(5): 964–975. http://doi.org/10.1002/art.40003.

Pile KD, Tikly M, Bell JI, et al. 1992. HLA-DR antigens and rheumatoid arthritis in black South Africans: a study of ethnic groups. Tissue Antigens, 39: 138-40.

Pincus T, Yazici Y, Sokka T, et al. (2003). Methotrexate as the “anchor drug” for the treatment of early rheumatoid arthritis. Clin Exp Rheumatol. 21(5 Suppl 31):S179-85.

Plenge RM, Seielstad M, Padyukov L, et al. (2007). TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. N Engl J Med. 357(12):1199-209. http://doi.org/10.1056/NEJMoa073491.

Pollard L, Choy EH, and Scott DL. (2005). The consequences of rheumatoid arthritis: quality of life measures in the individual patient. Clin Exp Rheumatol. 23(5 Suppl 39):S43-52.

Pruim RJ. Et al. 2010. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics, 26: 2336–2337. http://doi.org/10.1093/bioinformatics/btq419.

Puleston DJ, Zhang H, Powell TJ, et al. (2014). Autophagy is a critical regulator of memory CD8(+) T cell formation. Elife. 11:3. http://doi.org/10.7554/eLife.03706.

Puntis D, Malik S, Saravanan V, et al. (2013). Urinary tract infections in patients with rheumatoid arthritis. Clin Rheumatol. 32(3):355-60. http://doi.org/10.1007/s10067-012-2129-7.

Purcell S, Neale B, Todd-Brown K, et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81:559–75. http://doi.org/10.1086/519795.

Qiu Q, Huang J, Lin Y, et al. 2017. Polymorphisms and pharmacogenomics for the toxicity of methotrexate monotherapy in patients with rheumatoid arthritis: A systematic review and meta-analysis. Medicine (Baltimore). 96(11):e6337. http://doi.org/10.1097/MD.0000000000006337.

Quinonez-flores CM, Gonzalez-Chavez SA, Najera DDR, et al. (2016). Oxidative Stress Relevance in the Pathogenesis of the Rheumatoid Arthritis: A Systematic Review. BioMed Research International, Volume 2016, Article ID 6097417, 14 pages. http://dx.doi.org/10.1155/2016/6097417.

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ramsay M, Crowther N, Tambo E, et al. (2016). H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. Global Health, Epidemiology and Genomics, 1, E20. http://doi.org/10.1017/gheg.

Ranganathan P, Culverhouse R, March S, et al. (2004). Single nucleotide polymorphism profiling across the methotrexate in normal subjects and patients with rheumatoid arthritis. Pharmacogenomics, 5(5):559-69. http://doi.org/10.1517/14622416.5.5.559.

Ranganathan P. (2008). Pharmacogenomics in Rheumatoid Arthritis. In: Yan Q. (eds) Pharmacogenomics in Drug Discovery and Development. Methods in Molecular Biology™, vol 448. Humana Press.

Raposo B, Merky P, Lundqvist C, et al. (2018). T cells specific for post-translational modifications escape intrathymic tolerance induction. Nat Commun. 9(1):353. http://doi.org/10.1038/s41467-017-02763-y.

Rashid T, Jayakumar KS, Binder A, et al. (2007). Rheumatoid arthritis patients have elevated antibodies to cross-reactive and non-cross-reactive antigens from Proteus microbes. Clin Exp Rheumatol. 25: 259– 67.

Raychaudhuri S, Sandor C, and Stahl E. (2012). Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet. 44(3): 291–296. http://doi.org/10.1038/ng.1076.

Raza K, Buckley CE, Salmon M, et al. 2006. Treating very early rheumatoid arthritis. *Best* Practice and Research Clinical Rheumatology, 20(5): 849-863. http://doi.org/10.1016/j.berh.2006.05.005.

Restrepo, Luisa F, et al. (2016). Pharmacogenetics of methotrexate in rheumatoid arthritis: A systematic review. Rev. Colomb. Reumatol. 23(2):102-114. ISSN 0121-8123.

Reveille JD, Alarcon GS, Fowler SE, et al. (1996). HLA—DRB1 genes and disease severity in rheumatoid arthritis. Arthritis & amp; Rheumatism, 39(11). https://doi.org/10.1002/art.1780391105.

Reynolds RJ, Ahmed AF, Danila MI, et al. (2014). HLA-DRB1-associated rheumatoid arthritis risk at multiple levels in African Americans: hierarchical classification systems, amino acid positions, and residues. Arthritis Rheumatol. 66:3274–82. http://doi.org/10.1002/art.38855.

Risch N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet. 46: 222-228.

Roe CR and Ding J. (2001). Mitochondrial fatty acid oxidation disorders. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds). The Metabolic and Molecular Basis of Inherited Disease. McGraw-Hill, New York, pp 2308–2311.

Romanoski CE, Glass CK, Stunnenberg HG, et al. (2015). Epigenomics Roadmap for regulation. Nature, 518: 314–316. http://doi.org/10.1038/518314a.

Rosell M, Wesley AM, Rydin K, et al. (2009). Dietary fish and fish oil and the risk of rheumatoid arthritis. Epidemiology. 20(6):896-901. http://doi.org/10.1097/EDE.0b013e3181b5f0ce.

Rots MG, Pieters R, Peters GJ, et al. (1999). Circumvention of Methotrexate Resistance in Childhood Leukemia Subtypes by Rationally Designed Antifolates. Blood, 94:3121-3128.

Rudan I, Sidhu S, Papana S, et al. (2015). Prevalence of rheumatoid arthritis in low- and middle-income countrie: A systematic review and analysis. J Glob Health. 5(1):010409. http://doi.org/10.7189/jogh.05.010409.

Saag KG, Cerhan JR, Kolluri S, et al. (1997). Cigarette smoking and rheumatoid arthritis severity. Ann Rheum Dis. 56:463–469.

Saevarsdottir S, Wedren S, Seddighzadeh M, et al. (2011). Patients with early rheumatoid arthritis who smoke are less likely to respond to treatment with methotrexate and tumor necrosis factor inhibitors: observations from the Epidemiological Investigation of Rheumatoid Arthritis and the Swedish Rheumatology Register cohorts. Arthritis Rheum. 63(1):26-36. http://doi.org/10.1002/art.27758.

Salliot C and van der Heijde D. (2009). Long-term safety of methotrexate monotherapy in patients with rheumatoid arthritis: a systematic literature research. Ann Rheum Dis. 2009 Jul; 68(7): 1100–1104. http://doi.org/10.1136/ard.2008.093690.

Salomon-Escoto KI, Gravallese EM and Kay J. (2011). Assessment of control of rheumatoid arthritis disease activity. Best Pract Res Clin Rheumatol, 25(4):497–507. http://doi.org/10.1016/j.berh.2011.10.007.

Santos-Moreno PI, de la Hoz-Valle J, Villarreal L, et al. (2015). Treatment of rheumatoid arthritis with methotrexate alone and in combination with other conventional DMARD using the T2T strategy. A cohort study. Clin Rheumatol. 34(2):215-20. http://doi.org/10.1007/s10067-014-2794-9.

Saxena R, Plenge RM, Bjonnes AC, et al. (2017). A Multinational Arab Genome-Wide Association Study Identifies New Genetic Associations for Rheumatoid Arthritis. Arthritis Rheumatol. 69(5):976–85. http://doi.org/10.1002/art.40051.

Schaub MA, Boyle AP, Kundaje A, et al. (2012). Linking disease associations with regulatory information in the human genome. Genome Research 22(9): 1748–1759. http://doi.org/10.1101/gr.136127.111.

Schneider M, Manabile E and Tikly M. (2008). Social aspects of living with rheumatoid arthritis: a qualitative descriptive study in Soweto, South Africa – a low resource context. Health and Quality of Life Outcomes20086:54. https://doi.org/10.1186/1477-7525-6-54.

Scott DL. (2010). Rheumatoid arthritis. The Lancet, 376(9746): P1094-1108. https://doi.org/10.1016/S0140-6736(10)60826-4.

Sedlacek HH. (2001). Mechanisms of action of flavopiridol. Crit. Rev. Oncol. Hematol. 38: 139–170. https://doi.org/10.1016/S1040-8428(00)00124-4.

Seldin MF, Amos CI, Ward R, et al. (1999). The genetics revolution and the assault on rheumatoid arthritis. Arthritis Rheum. 42:1071–1079. http://doi.org/10.1002/1529-0131(199906)42:6<1071::AID-ANR1>3.0.CO;2-8.

Senapati S, Singh S, Das M, et al. (2018). Genome-wide analysis of methotrexate pharmacogenomics in rheumatoid arthritis shows multiple novel risk variants and leads for TYMS regulation. Pharmacogenetics and Genomics, 24:211–219. http://doi.org/10.1097/fpc.0000000000000036.

Senapati S, Singh S, Das M, et al. 2014. Genome-wide analysis of methotrexate pharmacogenomics in rheumatoid arthritis shows multiple novel risk variants and leads for TYMS regulation. Pharmacogenet. Genomics 24: 211–219 http://doi.org/10.1097/FPC.0000000000000036.

Shaik NA, Banaganapalli B, Elango R, et al. (2019). Tools and Methods in Analysis of Complex Sequences. In: Shaik N., Hakeem K., Banaganapalli B., Elango R. (eds) Essentials of Bioinformatics, Volume I. Springer, Cham. https://doi.org/10.1007/978-3-030-02634-9_8.

Shapiro JA, Koepsell TD, Voigt LF, et al. (1996). Diet and rheumatoid arthritis in women: a possible protective effect of fish consumption. Epidemiology. 7:256–263. http://doi.org/10.1097/00001648-199605000-00007.

Silman AJ and Pearson JE. (2002). Epidemiology and genetics of rheumatoid arthritis. Arthritis Res. 4(Suppl 3):S265–S272. http://doi.org/10.1186/ar578.

Silman AJ, et al. (1993). Twin concordance rates for rheumatoid arthritis: results from a nationwide study. Br. J. Rheumatol. 32, 903–907. http://doi.org/10.1093/rheumatology/32.10.903.

Silman AJ, Hochberg MC. (2001). In: In Epidemiology of the Rheumatic Diseases. Edited by AJ Silman, MC Hochberg, editor. Oxford: Oxford University Press. Rheumatoid arthritis. Pp. 31–71.

Silman AJ, Newman J, and Macgregor AJ. (1996). Cigarette smoking increases the risk of rheumatoid arthritis: Results from a nationwide study of disease-discordant twins. Arthritis & Rheumatism, 39(5). https://doi.org/10.1002/art.1780390504.

Simkins HM, Merriman ME, Highton J, et al. (2005). Association of the PTPN22 locus with rheumatoid arthritis in a New Zealand Caucasian cohort. Arthritis Rheum 52:2222–5. http://doi.org/10.1002/art.21126.

Sizova L. (2008). Approaches to the treatment of early rheumatoid arthritis with disease-modifying antirheumatic drugs. Br J Clin Pharmacol. 66(2): 173–178. http://doi.org/10.1111/j.1365-2125.2008.03222.x.

Skoumal M, Haberhauer G, Kolarz G, et al. (2005). Serum cathepsin K levels of patients with longstanding rheumatoid arthritis: correlation with radiological destruction. Arthritis Res Ther. 7(1):R65-70. http://doi.org/10.1186/ar1461

Slimani S, Abbas A, Ben Ammar A, et al. (2017). Prevalence of metabolic syndrome in Algerian rheumatoid arthritis patients. Correlation with disease activity and functional

status. Diabetes Metab Syndr. 11(1):S425-S427. http://doi.org/10.1016/j.dsx.2017.03.029.

Smolen JS, Aletaha D, Barton A, et al. (2018). Rheumatoid arthritis. Nat Rev Dis Primers. 18001. http://doi.org/10.1038/nrdp.2018.1.

Smolen JS, Breedveld FC, Burmester GR, et al. (2016). Treating rheumatoid arthritis to target: 2014 update of the recommendations of an international task force. Annals of the Rheumatic Diseases, 75:3-15. http://doi.org/10.1136/annrheumdis-2015-207524.

Smolik I, Robinson DB, Bernstein CN, et al. (2013). First-degree Relatives of Patients with Rheumatoid Arthritis Exhibit High Prevalence of Joint Symptoms. The Journal of Rheumatology, 40(6): 818-824. https://doi.org/10.3899/jrheum.121016.

Sobacchi C, Menale C and Villa A. (2019). The RANKL-RANK Axis: A Bone to Thymus Round Trip. Front. Immunol. https://doi.org/10.3389/fimmu.2019.00629.

Sokka T, Abelson B and Pincus T. (2008). Mortality in rheumatoid arthritis: 2008 update. Clin Exp Rheumatol. 26 (Suppl. 51): S35-S61.

Sokka T, Kautiainen H, Pincus T, et al. (2010). Work disability remains a major problem in rheumatoid arthritis in the 2000s: data from 32 countries in the QUEST-RA study. Arthritis Res Ther. 12(2):R42. http://doi.org/10.1186/ar2951.

Sokolove J, Bromberg R, Deane KD, et al. (2012). Autoantibody Epitope Spreading in the Pre-Clinical Phase Predicts Progression to Rheumatoid Arthritis. PloS ONE 7(5): e35296. https://doi.org/10.1371/journal.pone.0035296.

Solomon L, Beighton P, Valkenburg HA, et al. (1975). Rheumatic disorders in the South African Negro. Part I. Rheumatoid arthritis and ankylosing spondylitis. S Afr Med J. 49(32):1292–6.

Soukup T, Nekvindova J, Dosedel M, et al. (2016). Methotrexate impact on radiographic progression in biologic-treated rheumatoid arthritis under clinical remission: A case report on monozygotic Caucasian twins. International journal of immunopathology and pharmacology, 29(4), 790–795. http://doi.org/10.1177/0394632016674725.

Spreafico R, Rossetti M, van Loosdregt J, et al. (2016). A circulating reservoir of pathogenic-like CD4+ T cells shares a genetic and phenotypic signature with the inflamed synovial micro-environment. Annals of the Rheumatic Diseases, 75:459-465. http://doi.org/10.1136/annrheumdis-2014-206226.

Stahl EA, et al. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 42:508–514. http://doi.org/10.1038/ng.582.

Stanich, JA, Carter JD, Whittum-Hudson J, et al. (2009). Rheumatoid arthritis: Disease or syndrome? Open access rheumatology : research and reviews, 1: 179–192. http://doi.org/10.2147/oarrr.s7680.

Stark M, Wichman C, Avivi I, et al. (2009). Aberrant splicing of folylpolyglutamate synthetase as a novel mechanism of antifolate resistance in leukemia. Blood, 113: 4362-9. http://doi.org/10.1182/blood-2008-08-173799.

Stastny P. (1978). Association of the B-cell alloantigen DRw4 with rheumatoid arthritis. . N Engl J Med. 298:869–871.

Stranzl, T, Wolf J, Leeb BF, et al. (2003). Expression of folylpolyglutamyl synthetase predicts poor response to methotrexate therapy in patients with rheumatoid arthritis. Clin Exp Rheumatol. 21(1):27-32.

Sugiyama D, Nishimura K, Tamaki K, et al. (2010). Impact of smoking as a risk factor for developing rheumatoid arthritis: a meta-analysis of observational studies. Ann Rheum Dis. 69(1):70-81. http://doi.org/10.1136/ard.2008.096487.

Sulovari A, Chen YH, Hudziak JJ, Li D. 2017. Atlas of human diseases influenced by genetic variants with extreme allele frequency differences. Hum Genet. 136: 39–54.

Suzuki T, Ikari K, Yano K, et al. (2013). PADI4 and HLA-DRB1 Are Genetic Risks for Radiographic Progression in RA Patients, Independent of ACPA Status: Results from the IORRA Cohort Study. PloS ONE 8(4): e61045. https://doi.org/10.1371/journal.pone.0061045.

Taylor JC, Bongartz T, Massey J, et al. (2018). Genome-wide Association Study of Response to Methotrexate in Early Rheumatoid Arthritis Patients. The Pharmacogenomics Journal, 18 (4). ISSN 1470-269X. http://doi.org/https://doi.org/10.1038/s41397-018-0025-5.

Tekola-Ayele, F. and Rotimi, C. N. 2015. Translational genomics in low- and middle-income countries: opportunities and challenges. Public Health Genomics, 18: 242–247. http://doi.org/10.1159/000433518.

Terao C, Ikari K, Nakayamada S, et al. (2016). A twin study of rheumatoid arthritis in the Japanese population. Mod. Rheumatol. 4:1–5. https://doi.org/10.3109/14397595.2015.1135856.

Terao, C. et al. (2011). The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. Hum Mol Genet. 20: 2680–2685. http://doi.org/10.1093/hmg/ddr161

Thalayasingam N, Nair N, Skelton AJ, et al. (2018). CD4+ and B Lymphocyte Expression Quantitative Traits at Rheumatoid Arthritis Risk Loci in Patients With Untreated Early Arthritis: Implications for Causal Gene Identification. Arthritis Rheumatol. 70(3):361-370. http://doi.org/10.1002/art.40393.

The 1000 Genome Project Consortium. (2015). The global reference for human genetics. Nature, 526: 68–74. https://doi.org/10.1038/nature15393.

The GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. Nature, 550: 204–213. https://doi.org/10.1038/nature24277.

The H3Africa Consortium. (2014). Enabling the genomic revolution in Africa. Science, 344(6190): 1346-1348. http://doi.org/10.1126/science.1251546.

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447:661–78. https://doi.org/10.1038/nature05911.

Thomson W, Barton A, Ke X, et al. (2007). Rheumatoid arthritis association at 6q23. Nat. Genet. 39:1431–1433. https://doi.org/10.1038/ng.2007.32.

Tikly M, Govind N, Frost J, et al. (2010). The PTPN22 R620W polymorphism is not associated with systemic rheumatic diseases in South Africans. Rheumatology (Oxford), 49:820–1. http://doi.org/10.1093/rheumatology/kep399.

Tishkoff, SA and Williams SM. (2002). Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet, 3: 611-21. http://doi.org/10.1038/nrg865.

Tiwana H, Wilson C, Alvarez A, et al. (1999). Cross-reactivity between the rheumatoid arthritis-associated motif EQKRAA and structurally related sequences found in Proteus mirabilis Infect. Immun. 67: 2769-2775.

Tobón GJ, Youinou P, Saraux A. (2010). The environment, geo-epidemiology, and autoimmune disease: rheumatoid arthritis. J Autoimmun. 35(1). http://doi.org/10–410.1016/j.jaut.2009.12.009.

Traherne JA, Horton R, Roberts AN, et al. (2006). Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history, PloS Genet. 2: e9. http://doi.org/10.1371/journal.pgen.0020009.

Traylor M, Curtis C, Patel H, et al. (2017). Genetic and environmental risk factors for rheumatoid arthritis in a UK African ancestry population: the GENRA case–control study. Rheumatology, 56(8): 1282–1292, https://doi.org/10.1093/rheumatology/kex048.

Trier NH, Holm BE, Heiden J, et al. (2018). Antibodies to a strain-specific citrullinated Epstein-Barr virus peptide diagnoses rheumatoid arthritis. Scientific Reports 8, Article number: 3684. https://doi.org/10.1038/s41598-018-22058-6.

Truswell AS and Hansen JDL. (1968). Medical and nutritional studies of! Kung Bushmen in north-west Botswana: a preliminary report. S Afr Med J. 42:1338–9.

Trynka G, Sandor C, Han B, et al. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature Genetics, 45: 24–130. http://doi.org/10.1038/ng.2504.

Turesson C, and Jacobsson LTH. (2004). Epidemiology of extra-articular manifestations in rheumatoid arthritis. Scand J Rheumatol. 33:65-72. http://doi.org/10.1080/03009740310004621.

Turesson C, O'Fallon WM, Crowson CS, et al. (2002). Extra-articular disease manifestations in rheumatoid arthritis: incidence trends and risk factors over 46 years. Ann Rheum Dis. 62(8):722-7. http://doi.org/10.1136/ard.62.8.722.

Turner S. 2017. Qqman: Q-Q and Manhattan Plots for GWAS Data, R Package Version 0.1.4. Available at https://cran.r-project.org/web/packages/qqman/. Accessed September 21, 2018.

Usenbo A, Kramer V, Young T, et al. 2015. Prevalence of Arthritis in Africa: A Systematic Review and Meta-Analysis. PloS One. 10(8): e0133858. Published online 2015 Aug 4. http://doi.org/10.1371/journal.pone.0133858.

Van den Hoek J, Boshuizen HC, Roorda LD, et al. (2017). Mortality in patients with rheumatoid arthritis: a 15-year prospective study. Rheumatol Int. 37(4):487-493. http://doi.org/10.1007/s00296-016-3638-5.

Van der Heijde DMFM 1995. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. Baillière's Clinical Rheumatology, 10(3): 435-453. https://doi.org/10.1016/S0950-3579(96)80043-4.

Van der Helm-van Mil AH, Wesoly JZ and Huizinga TW. (2005). Understanding the genetic contribution to rheumatoid arthritis. Curr Opin Rheumatol. 17(3):299-30. http://doi.org/10.1097/01.bor.0000160780.13012.be.

Van der Straaten RJ, Wessels JA, de Vries-Bouwstra JK, et al. (2007). Exploratory analysis of four polymorphisms in human GGH and FPGS genes and their effect in methotrexate-treated rheumatoid arthritis patients. Pharmacogenomics. 8(2):141-50. http://doi.org/10.2217/14622416.8.2.141.

Van der Woude D, Houwing-Deistermaat TT, Toe RE, et al. (2009). Quantitative heritability of anti-citrullinated protein antibody positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. Arthritis Rheum. 60:916. http://doi.org/10.1002/art.24385.

Van der Woude D, Houwing-Duistermaart JJ, Toes REM, et al. (2009). Quantitative heritability of anti–citrullinated protein antibody–positive and anti–citrullinated protein antibody–negative rheumatoid arthritis. https://onlinelibrary.wiley.com/journal/15290131.

van der Woude D, Rantapaa-Dahlqvist S, Ioan-Facsinay A, et al. (2010). Epitope spreading of the anti-citrullinated protein antibody response occurs before disease onset and associated with the disease course of early arthritis Ann Rheum Dis. 69 (8): 1554-1561. http://doi.org/10.1136/ard.2009.124537.

Van Loosdregt J, Rossetti M, Spreaficao R, et al. (2016). Increased autophagy in CD4+ T-cells of rheumatoid arthritis patients results in T-cell hyperactivation and apoptosis resistance. Eur J. Immunol. 46(12):2862-2870. http://doi.org/10.1002/eji.201646375.

Van Oene M, Wintle RF, Liu X, et al. (2005). Association of the lymphoid tyrosine phosphatase R620W variant with rheumatoid arthritis, but not Crohn's disease, in Canadian populations. Arthritis Rheum. 52(7):1993-8. http://doi.org/10.1002/art.21123.

Vaser R, Adusumalli S, Leng SN, et al. (2016). SIFT missense predictions for genomes. Nature Protocols, 11: 1–9. http://doi.org/10.1038/nprot.2015.123.

Vaughan JH, Kouri T, Petersen J, et al. (1988). On the Etiology of Rheumatoid Arthritis, Scandinavian Journal of Rheumatology, 17(74): 19-28, http://doi.org/10.3109/03009748809102935.

Vesperini V, Lukas C, Fautrel B, et al. (2013). Association of tobacco exposure and reduction of radiographic progression in early rheumatoid arthritis: Results from a French multicentre cohort. Arthritis Care Res. (Hoboken) 65:1899–1906. http://doi.org/10.1002/acr.22057.

Viatte S, Plant D and Raychaudhuri S. (2013). Genetics and epigenetics of rheumatoid arthritis. Nat Rev Rheumatol. 9(3):141-53. http://doi.org/10.1038/nrrheum.2012.237.

Viatte S, Plant D, Bowes J, et al. (2012). Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients. Ann Rheum Dis. 71(12):1984-90. http://doi.org/ 10.1136/annrheumdis-2011-201225.

Viatte S, Plant D, Han B, et al. (2015). Association of HLA-DRB1 haplotypes with rheumatoid arthritis severity, mortality, and treatment response. JAMA. 313:1645–1656. http://doi.org/10.1001/jama.2015.3435.

Vignal C, Bansal AT, Balding DJ, et al. (2009). Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-DRB1 loci. Arthritis Rheum. 60(1):53–62. http://doi.org/10.1002/art.24138.

Visentin M, Zhao RB, Goldman ID. (2012). The Antifolates. Hematol Oncol Clin North Am. 26:629–648. http://doi.org/10.1016/j.hoc.2012.02.002.

Voigt LF, Koepsell TD, Nelson JL, et al. (1994). Smoking, obesity, alcohol consumption, and the risk of rheumatoid arthritis. Epidemiology. 5(5):525–32.

Vomero, M, Barbarti C, Colasanti T, et al. (2018). Autophagy and rheumatoid arthritis: Current knowledge and future perspective. Front Immunol. 9:1577. http://doi.org/10/3389/fimmu.2018.01577.

Vyse TJ and Todd JA. (1996). Genetic analysis of autoimmune disease. Cell. 85:311–318. http://doi.org/10.1016/s0092-8674(00)81110-1.

Wakitani S, Murata N, Toda Y, et al. (1997). The relationship between HLA-DRB1 alleles and disease subsets of rheumatoid arthritis in Japanese. Br. J. Rheumatol. 36: 630-636. http://doi.org/10.1093/rheumatology/36.6.630.

Walsh AM, Whitaker JW, Huang CC, et al. (2016). Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. Genome Biology, 17:79. https://doi.org/10.1186/s13059-016-0948-6.

Wang BC, Chen Y, Furnback W, et al. (2016). Comparing The Epidemiology, Cost, And Quality Of Life Of Rheumatoid Arthritis In China With The Rest Of The World. Value in Health, 19(3): A239. https://doi.org/10.1016/j.jval.2016.03.1115.

Wang K, et al. (2012). Cyclin-dependent kinase 9 activity regulates neutrophil spontaneous apoptosis. PloS One 7, e30128. http://doi.org/10.1371/journal.pone.0030128.

Wang K, Li M, Hakonarson H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Research, 38(16): e164, https://doi.org/10.1093/nar/gkq603.

Wang W, Zhou H, and Liu L. (2018). Side effects of methotrexate therapy for rheumatoid arthritis: A systematic review. European Journal of Medicinal Chemistry, 58(5): 502-516. http://doi.org/10.1016/j.ejmech.2018.09.027.

Ward LD and Kellis M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants, Nucleic Acids Research, 40(D1): D930–D934, https://doi.org/10.1093/nar/gkr917.

Ward LD and Kellis M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and diseases. Nucleic Acids Res. 44(D1):D877-81. http://doi.org/10.1093/nar/gkv1340.

Watanabe K, Taskesen E, van Bochoven A. (2017). Functional mapping and annotation of genetic associations with FUMA. Nature Communications 8, Article number: 1826. https://doi.org/10.1038/s41467-017-01261-5.

Wegner N, Wait R, Sroka A, et al. (2010). Peptidylarginine deiminase from Porphyromonas gingivalis citrullinates human fibrinogen and alphaenolase: implications for autoimmunity in rheumatoid arthritis. Arthritis Rheum 62, 2662-2672. http://doi.org/10.1002/art.27552.

Wessels JA, de Vries-Bouwstra JK, Heijmans BT, et al. (2006). Efficacy and toxicity of methotrexate in the early rheumatoid arthritis are associated with single-nucleotide polymorphisms in genes coding for folate pathway enzyme. Arthritis Rheum. 54(4):1087-95. http://doi.org/10.1002/art.21726.

Wessels JA, Huizinga TW and Guchelaar HJ. (2008). Recent insights in the pharmacological actions of methotrexate in the treatment of rheumatoid arthritis. Rheumatology (Oxf.) 47, 249–255. http://doi.org/ 10.1093/rheumatology/kem279.

Westhoff G, Rau R, Zink A. (2008). Rheumatoid arthritis patients who smoke have a higher need for DMARD and feel worse, but they do not have more joint damage than non-smokers of the same serological group. Rheumatology (Oxford), 47:849–854. http://doi.org/ 10.1093/rheumatology/ken057.

Westra HJ, Peters MJ, Esko T, et al. (2013). Systematic identification of trans eQTL as putative drivers of known disease associations. Nat Genet. 45(10):1238-1243. https://doi.org/10.1038/ng.2756.

Weyand CM, Hicok KC, Conn DL, et al. (1992). The influence of HLA-DRB1 genes on disease severity in rheumatoid arthritis. Ann Intern Med. 117(10):801-6. http://doi.org/10.7326/0003-4819-117-10-801.

Whirl-Carrillo M, McDonagh EM, Hebert JM, et al. (2012). Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 92(4):414-417. http://doi.org/10.1038/clpt.2012.96.

Woodman R, Baghdadi L, Shanahan E, et al. (2019). Diets high in n-3 fatty acids are associated with lower arterial stiffness in patients with rheumatoid arthritis: A latent profile analysis – CORRIGENDUM. British Journal of Nutrition, 121(9), 1080-1080. http://doi.org/10.1017/S0007114519000631.

Wordsworth P, Pile KD, Buckely JD, et al. (1992). HLA heterozygosity contributes to susceptibility to rheumatoid arthritis. Am J Hum Genet. 51:585–591.

World Health Organization. (2008). WHO report on the global tobacco epidemic, Geneva, Switzerland: World Health Organization, http://www.who.int/tobacco/mpower/mpower_report_full.

Wu MC, Lee S, Cai T, et al. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 89(1):82–93. http://doi.org/10.1016/j.ajhg.2011.05.029.

Wu Y, Waite LL, Jackson AU, Sheu WH-H, Buyske S, Absher D, et al. (2013). Trans-Ethnic Fine-Mapping of Lipid Loci Identifies Population-Specific Signals and Allelic Heterogeneity That Increases the Trait Variance Explained. PLoS Genet 9(3): e1003379. https://doi.org/10.1371/journal.pgen.1003379

Xu X, Araki K, Li S, et al. (2014). Autophagy is essential for effector CD8+ T cell survival and memory formation. Nature Immunology, 15: 1152–1161. https://doi.org/10.1038/ni.3025.

Yamamoto K, Okada Y, Suzuki A, et al. (2015). Genetic studies of rheumatoid arthritis. Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci. 91:410–422. http://doi.org/10.2183/pjab.91.410.

Yamamoto T, Shikano K, Nanki T, et al. (2016). Folylpolyglutamate synthase is a major determinant of intracellular methotrexate polyglutamates in patients with rheumatoid arthritis. Sci Rep. 6: 35615. http://doi.org/10.1038/srep35615.

Yang TH, Yuan TH, Hwang YH, et al. (2016). Increased inflammation in rheumatoid arthritis patients living where farm soils contain high levels of copper. J Formos Med Assoc. 115(11):991-996. http://doi.org/10.1016/j.jfma.2015.10.001.

Yang XK, Liu J, Liu J, et al. (2015). Associations Between PADI4 Gene Polymorphisms and Rheumatoid Arthritis: An Updated Meta-analysis. Arch Med Res. 46:317–25. https://doi.org/10.1016/j.arcmed.2015.05.011.

Ying YT, Petersen J, Ramarathinam SH, et al. (2018). The interplay between citrullination and HLA-DRB1 polymorphism in shaping peptide binding hierarchies in rheumatoid arthritis. J Biol Chem. 293(9):3236-3251. http://doi.org/10.1074/jbc.RA117.001013.

Zanelli E, Breedveld FC and de Vries RR. (2000). HLA class II association with rheumatoid arthritis: facts and interpretations. Hum Immunol. 61(12):1254-61. http://doi.org/10.1016/s0198-8859(00)00185-3.

Zanelli E, Gonzalez-Gay MA and David CS. (1995). Could HLA-DRB1 be the protective locus in rheumatoid arthritis? Immunology Today, 16(6): 274-278. http://doi.org/10.1016/0167-5699(95)80181-2.

Zannettou N, Hopley M, Tikly M. (1997). Slow-acting antirheumatic drug therapy in rheumatoid arthritis. South African Rheumatism and Arthritis Association 15th Biennial Congress,Durban,1997. Personal communication.

Zeilinger S, Kühnel B, Klopp N, et al. (2013). Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS One. 8:e63812. http://doi.org/10.1371/journal.pone.0063812.

Zhao M, Liang L, Ji L, et al. (2016). MTHFR gene polymorphisms and methotrexate toxicity in adult patients with hematological malignancies: a meta-analysis. Pharmacogenomicsvol. 17:9. https://doi.org/10.2217/pgs-2016-0004.

Zheng Z, Shen J, Cox C, et al. (2014). HIBAG—HLA genotype imputation with attribute bagging. The Pharmacogenomics Journal, 14: 192–200. https://doi.org/10.1038/tpj.2013.18.

Zhu L, Wang H, Wu Y, et al. (2017). The Autophagy level is increased in the synovial tissues of patients with active Rheumatoid arthritis and is correlated with disease severity. Mediators of inflammation. 2017: 7623145. http://doi.org/10.1155/2017/7623145.

APPENDICES:

APPENDIX A: Approval of title and Ethics Certificate

Approval of Title

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Private Bag 3 Wits, 2050
Fax: 027117172119
Tel: 02711 7172076

Reference: Mrs Sandra Benn
E-mail: sandra.benn@wits.ac.za

19 July 2017
Person No: 1701006
PAG

Mr EM Mathebula
209 Smit
Braamfontein Gate
Johannesburg
2001
South Africa

Dear Mr Mathebula

Doctor of Philosophy: Approval of Title

We have pleasure in advising that your proposal entitled *Rheumatoid arthritis in black South Africans: Genomic susceptibility and methotrexate response* has been approved. Please note that any amendments to this title have to be endorsed by the Faculty's higher degrees committee and formally approved.

Yours sincerely

Mrs Sandra Benn
Faculty Registrar
Faculty of Health Sciences

Ethics Clearance Certificate

R14/49 Mr Evans Mantiri Mathebula et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1706109

NAME: (Principal Investigator)	Mr Evans Mantiri Mathebula et al
DEPARTMENT:	Human Genetics University of the Witwatersrand Sydney Brenner Institute for Molecular Bioscience
PROJECT TITLE:	Rheumatoid Arthritis in Black South Africans: Genomic Susceptibility and Methotrexate Response
DATE CONSIDERED:	Adhoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-Study under Primary Study (M10107)
SUPERVISOR:	Prof Michelle Ramsay
APPROVED BY:	Professor P. Cleaton-Jones Chairperson, HREC (Medical)
DATE OF APPROVAL:	13/07/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 10004,10th floor, Senate House/3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review June and will therefore be due in the month of June each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 13/07/2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

APPENDIX B: The 2010 ACR/EULAR classification criteria for rheumatoid arthritis

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

APPENDIX C: RegulomeDB variant classification scheme. Lower scores indicate a high probability for a variant to be located in a functional region

Category	Description
	Likely to affect binding and linked to the expression of a gene target
1a	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
	Likely to affect binding
2a	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	TF binding + any motif + DNase footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
	Less likely to affect binding
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
	Minimal binding evidence
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Motif hit

APPENDIX F: The EULAR response criteria after 6 months on methotrexate monotherapy.

EULAR Response Criteria			
Value of composite measure at endpoint	Improvement in composite measure from baseline		
	> 1.2†	> 0.6 and ≤ 1.2	≤ 0.6
≤ 3.2*	Good		
> 3.2 and ≤ 5.1		Moderate	
> 5.1			None

APPENDIX G: The genotype and allele frequency of the SNP selected for replication in BSA RA patients on Methotrexate for at least 6 months. SNP that were statistically significant at P <0.05 are in bold. The statistical significance of SNP with a cell count of less was estimated using Fisher exact test, and Chi Squared test for all other SNP.

CHR	GENE	Allele / Genotype	Model	GR	MR	P-value	GR	NR	P-value	MR	NR	P-value	R	NR	P-value
1	MTHFR (rs1801131 (G/T))	G	ALLELE	19	22	0.181	19	15	0.351	22	15	0.859	41	15	0.742
		T		193	138		193	107		138	107		331	107	
		GG		0	2		0	1		2	1		2	1	
		GT	GENO	19	18	0.193	19	13	0.341	18	13	1.000	37	13	0.817
		TT		87	60		187	47		60	47		147	47	
1	MTHFR (rs1801133 (A/G))	A	ALLELE	23	9	0.093	23	13	1.000	9	13	0.178	32	13	0.474
		G		189	151		189	109							
		AA		3	1		3	1		1	1		4	1	
		AG	GENO	17	7	0.288	17	11	0.936	7	11	0.185	24	11	0.672
		GG		86	72		86	49		72	49		158	49	
1	ADORA3 (rs3394 (T/C))	T	ALLELE	53	49	0.242	53	21	0.103	49	21	**0.012**	102	21	**0.029**
		C		159	111		159	101		111	101		270	101	
		TT		7	8		7	2		8	2		15	2	
		TC	GENO	39	33	0.512	39	17	0.305	33	17	**0.036**	72	17	0.094
		CC		60	39		60	42		39	42		99	42	
1	MTR (rs1805087 (G/A))	G	ALLELE	54	39	0.904	54	36	0.444	39	36	0.344	93	36	0.343
		A		158	121		158	86		121	86		279	86	
		GG		7	6		7	8		6	8		13	8	
		GA	GENO	40	27	0.842	40	20	0.337	27	20	0.590	67	20	0.328
		AA		59	47		59	33		47	33		106	33	
2	ATIC (rs2372536 (G/C))	G	ALLELE	8	9	0.457	8	10	0.129	9	10	0.474	17	10	0.166
		C		204	151		204	112		151	112		355	112	
		GG	GENO	1	0	0.225	1	2	0.247	0	2	0.319	1	2	0.194

		GC		6	9		6	6		9	5		15	6	
		CC		99	71		99	53		71	53		170	53	
2	**ATIC (rs7563206 (C/T))**	C	ALLELE	83	77	0.091	83	64	**0.022**	77	64	0.548	160	64	0.075
		T		129	83		129	58		83	58		212	58	
		CC		19	16		19	17		16	17		35/	17	
		CT	GENO	45	45	0.065	45	30	0.066	45	30	0.543	90	30	0.202
		TT		42	19		42	14		19	14		61	14	
2	**ATIC (rs3821353 (T/G))**	T	ALLELE	50	30	0.308	50	23	0.339	30	23	1.000	80	23	0.608
		G		162	130		162	99		130	99		292	99	
		TT		5	5		5	2		5	2		10	2	
		TG	GENO	40	20	0.178	40	19	0.653	20	19	0.573	60	19	0.844
		GG		61	55		61	40		55	40		116	40	
2	**ATIC (rs12995526 (C/T))**	C	ALLELE	58	41	0.724	58	34	1.000	41	34	0.685	99	34	0.814
		T		154	119		154	88		119	88		273	88	
		CC		7	8		7	6		8	6		15	6	
		CT	GENO	44	25	0.306	44	22	0.664	25	22	0.851	69	22	0.902
		TT		55	47		55	33		47	33		102	33	
2	**ATIC (rs16853826 (A/G))**	A	ALLELE	39	29	1.000	39	17	0.362	29	17	0.417	68	17	0.333
		G		173	131		153	105		131	105		304	105	
		AA	GENO	1	5	**0.045**	1	0/17 /44	0.614	5	0	0.147	6	0	0.443
		AG		37	19		37	17		19	17		56	17	
		GG		68	56		68	44		56	44		124	44	
2	**ATIC (rs4673993 (C/T))**	C	ALLELE	16	16	0.457	16	12	0.540	16	12	1.000	32	12	0.715
		T		196	144		196	110		114	110		340	110	
		CC	GENO	1	2	0.692	1	2	0.659	2	2	0.931	3	2	0.757

		TC		14	12		14	8		12	8		26	8	
		TT		91	66		91	51		66	51		157	51	
5	**FASTKD3MTRR (rs1801394 (G/A))**	G	ALLELE	75	48	0.317	75	37	0.400	48	37	1.000	123	37	0.656
		A		137	112		137	85		112	85		249	85	
		GG		11	5		11	5		5	5		16	5	
		GA	GENO	53	38	0.501	53	27	0.630	38	27	0.866	91	27	0.774
		AA		42	37		42	29		37	29		79	29	
7	**ABCB1 (rs1045642 (A/G))**	A	ALLELE	26	26	0.293	26	6	**0.033**	26	6	**0.004**	52	6	**0.006**
		G		186	134		186	116		134	116		320	116	
		AA		1	2		1	1		2	1		3	1	
		AG	GENO	24	22	0.513	24	4	**0.010**	22	4	**0.002**	46	4	**0.004**
		GG		81	56		81	56		56	56		137	56	
7	**ABCB1 (rs1128503 (A/G))**	A	ALLELE	19	17	0.600	19	7	0.397	17	7	0.196	36	7	0.201
		G		193	143		193	115		143	115		336	115	
		AA	GENO	0	1	0.619	0	0	0.376	1	0/7/54	0.293	1	0	0.428
		AG		19	15		19	7		15	7		34	7	
		GG		87	64		87	54		64	54		151	54	
7	**ABCB1 (rs10280623 (C/T))**	C	ALLELE	54	41	1.000	54	29	0.793	41	29	0.782	95	29	0.720
		T		158	119		158	93		119	93		277	93	
		CC		7	4		7	1		4	1		11	1	
		CT	GENO	40	33	0.834	40	27	0.364	33	27	0.620	73	102	0.427
		TT		59	43		59	33		43	33		102	33	
8	**GGH (rs719235 (A/C))**	A	ALLELE	1	0	1.000	1	4	0.062	0	4	**0.034**	1	4	**0.014**
		C		211	160		211	118		160	118		371	118	
		AA	GENO	0	0	1.000	0	0	0.060	0	0	**0.033**	0	0	**0.014**

		AC		1	0		1	4		0	4		1	4	
		CC		105	80		105	57		80	4		185	57	
8	**GGH (rs3758149 (A/G))**	A	ALLELE	29	27	0.464	29	25	0.123	27	25	0.443	56	25	0.161
		G		183	133		183	97		133	97		316	97	
		AA		1	3		1	2		3	2		4	2	
		AG	GENO	27	21	0.430	27	21	0.189	21	21	0.567	48	21	0.293
		GG		78	56		78	38		56	38		134	38	
9	**TLR4 (rs4986790 (G/A))**	G	ALLELE	7	2	0.310	7	1	0.266	2	1/12 1	1.000	9	1	0.463
		A		205	158		205	121		158	121		363	121	
		GG		0	0		0	0		0	0		0	0	
		GA	GENO	7	2	0.304	7	1	0.260	2	1	1.000	9	1	0.459
		AA		99	78		99	60		78	60		177	60	
9	**FPGS. (rs1544105 (C/T))**	C	ALLELE	77	57	0.914	77	37	0.283	57	37	0.374	134	37	0.274
		T		135	130		135	85		103	85		238	85	
		CC		18	9		18	4		9	4		27	4	
		CT	GENO	41	39	0.333	41	29	0.136	39	29	0.594	80	29	0.282
		TT		47	32		47	28		32	28		79	28	
9	**FPGS (rs10760502 (A/G))**	A	ALLELE	31	19	0.540	31	12	0.238	19	12	0.702	50	12	0.347
		G		181	141		181	110		141	110		322	110	
		AA		2	1		2	0		1	0		3	0	
		AG	GENO	27	17	0.832	27	12	0.471	17	12	0.907	44	12	0.609
		GG		77	62		77	49		62	49		139	49	
9	**FPGS (rs10106 (T/C))**	T	ALLELE	83	66	0.749	83	47	1.000	66	47	0.713	149	47	0.831
		C		129	94		129	75		94	75		223	75	
		TT	GENO	18	13	0.736	18	6	0.229	13	6	0.525	31	6	0.286

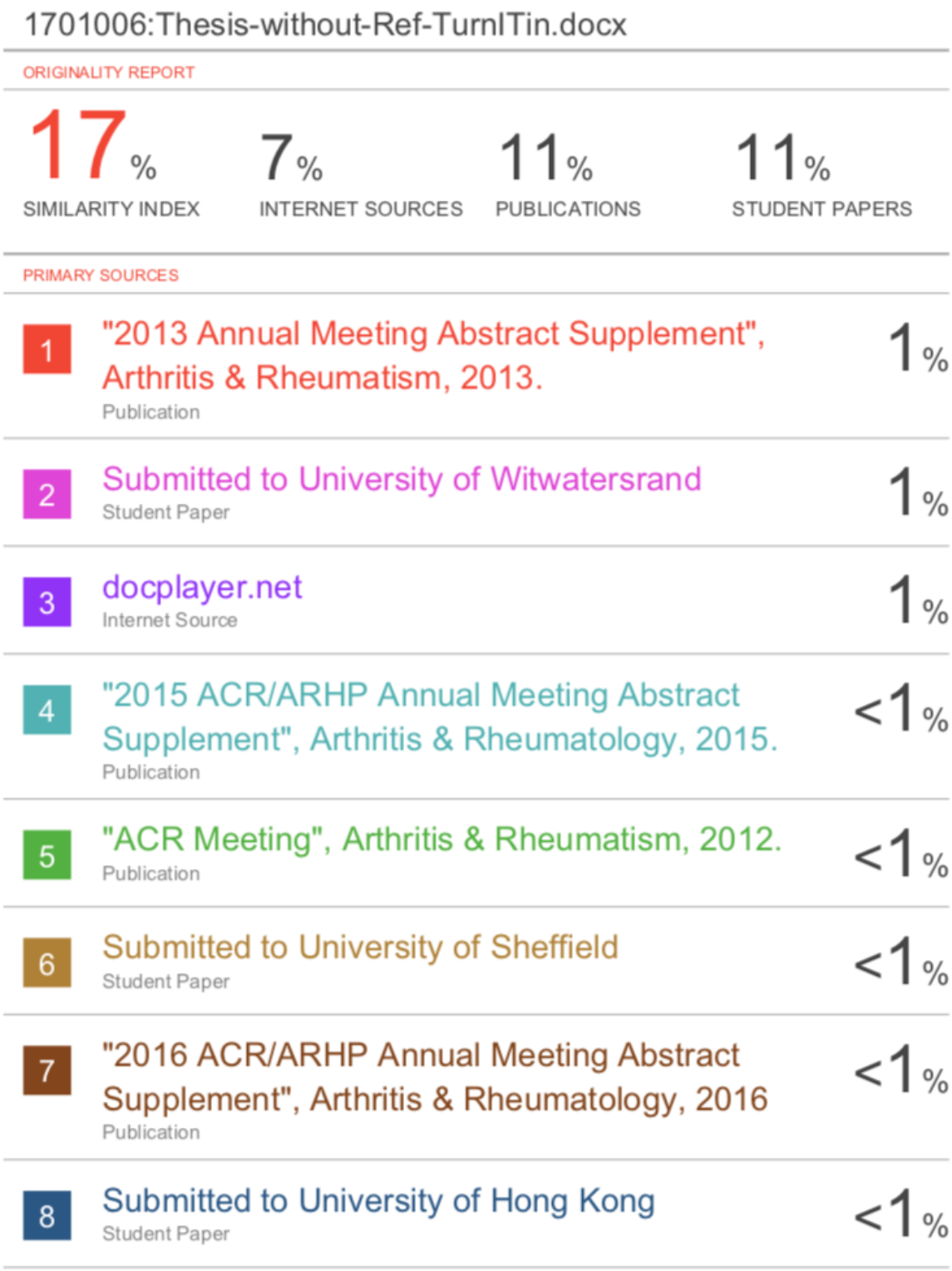
		TC		47	40		47	35		40	35		87	35	
		CC		41	27		41	20		27	20		68	20	
10	**ABCC2 (rs2273697 (A/G))**	A	ALLELE	38	26	0.782	38	17	0.363	26	17	0.620	64	17	0.481
		G		174	134		174	105		134	105		308	105	
		AA		3	2		3	2		2	2		5	2	
		AG	GENO	32	22	0.911	32	13	0.452	22	13	0.653	54	13	0.524
		GG		71	56		71	46		56	46		127	46	
10	**ABCC2 (rs4148396 (T/C))**	T	ALLELE	45	38	0.615	45	27	0.890	38	27	0.777	83	27	1.000
		C		167	122		167	95		122	95		289	95	
		TT		5	3		5	3		3	3		8	3	
		TC	GENO	35	32	0.656	35	21	0.961	32	21	0.758	67	21	0.935
		CC		66	45		66	37		45	37		111	37	
11	**SLC22A11 (rs1123180 9 (A/T))**	A	ALLELE	11	7	0.810	11	8	0.629	7	8	0.435	18	8	0.485
		T		201	153		201	114		153	114		354	114	
		AA		1	2		1	1		2	1		3	1	
		AT	GENO	9	3	0.311	9	6	0.899	3	6	0.381	12	6	0.608
		TT		96	75		96	54		75	54		171	54	
12	**KLRD1 (rs2302489 (T/A))**	T	ALLELE	60	38	0.344	60	32	0.705	38	32	0.677	98	32	1.000
		A		152	122		152	90		122	90		274	90	
		TT		8	2		8	3		2	3		10	3	
		AT	GENO	44	34	0.343	44	26	0.875	34	26	0.786	78	26	1.000
		AA		54	44		54	32		44	32		98	32	
12	**KLRC1 (rs7301582 (C/T))**	C	ALLELE	87	66	1.000	87	46	0.564	66	46	0.623	153	46	0.525
		T		125	94		125	76		94	76		219	76	
		CC	GENO	21	14	0.821	21	8	0.504	14	8	0.826	35	8	0.607
		CT		45	38		45	30		38	30		83	30	

		TT		40	28		40	23		28	23		68	23	
14	**MTHFD1 (rs2236225 (A/G))**	A	ALLELE	45	40	0.454	45	30	0.498	40	30	1.000	85	30	0.712
		G		167	120		167	92		120	92		287	92	
		AA		3	9		3	6		9	6		12	6	
		AG	GENO	39	22	**0.048**	39	18	0.139	22	18	0.937	61	18	0.597
		GG		64	49		64	37		49	37		113	37	
16	**ABCC1 (rs246240 (G/A))**	G	ALLELE	15	10	0.836	15	9	1.000	10	9	0.812	25	9	0.837
		A		197	150		197	113		150	113		347	113	
		GG		1	0		1	0		0	0		1	0	
		GA	GENO	13	10	1.000	13	9	0.774	10	9	0.805	23	9	0.746
		AA		92	70		92	52		70	52		162	52	
16	**ABCC1 (rs3784864 (A/G))**	A	ALLELE	21	13	0.591	21	10	0.698	13	10	1.000	34	10	0.856
		G		191	147		191	112		147	112		338	112	
		AA		2	0		2	0		0	0		2	0	
		AG	GENO	17	13	0.662	17	10	0.774	13	10	1.000	30	10	1.000
		GG		87	67		87	51		67	51		154	51	
17	**SHMT1 (rs1979277 (A/G))**	A	ALLELE	81	66	0.593	81	45	0.907	66	45	0.464	147	45	0.669
		G		131	94		131	77		94	77		225	77	
		AA		16	17		16	6		17	6		33	6	
		AG	GENO	49	32	0.528	49	33	0.522	32	33	0.123	81	33	0.239
		GG		41	31		41	22		31	22		72	22	
18	**TYMS (rs2853539 (A/G))**	A	ALLELE	26	20	1.000	26	21	0.253	20	21	0.308	46	21	0.174
		G		186	140		186	101		140	101		326	101	
		AA		1	1		1	0		1	0		2	0	
		AG	GENO	24	18	1.000	24	21	0.168	18	21	0.156	42	21	0.160
		GG		81	61		81	40		61	40		142	40	

20	**ITPA (rs1127354 (A/C))**	A	ALLELE	5	5	0.751	5	1	0.422	5	1	0.239	10	1	0.308
		C		207	155		207	121		155	121		362	121	
		AA	GENO	0	0		0	0		0	0		0	0	
		AC		5	5	0.747	5	1	0.417	5	1	0.235	10	1	0.302
		CC		101	75		101	60		75	60		176	60	
20	**ADA (rs244076 (C/T))**	C	ALLELE	104	80	0.917	104	55	0.497	80	55	0.471	184	55	0.406
		T		108	80		108	67		80	67		188	67	
		CC		27	25		27	12		25	12		52	12	
		CT	GENO	50	30	0.426	50	31	0.728	30	31	0.198	80	31	0.417
		TT		29	25		29	18		25	18		54	18	
21	**SLC19A1 (rs1051266 (C/T))**	C	ALLELE	52	56	**0.029**	52	39	0.161	56	39	0.613	108	39	0.569
		T		160	104		160	83		104	83		264	83	
		CC		6	9		6	7		9	7		15	7	
		CT	GENO	40	38	0.084	40	25	0.298	38	25	0.717	78	25	0.703
		TT		60	33		60	29		33	29		93	29	
21	**SLC19A1 (rs1131596 (A/G))**	A	ALLELE	44	45	0.111	44	28	0.679	45	28	0.340	89	28	0.903
		G		168	115		168	94		115	94		283	94	
		AA		5	6		5	6		6	6		11	6	
		AG	GENO	34	33	0.264	34	16	0.353	33	16	0.179	67	16	0.266
		GG		67	41		67	39		41	39		108	39	
22	**ADORA2A SPECC1L (rs5760410 (G/A))**	G	ALLELE	80	56	0.664	80	38	0.237	56	38	0.526	136	38	0.326
		A		132	104		132	84		104	84		236	84	
		GG		16	7		16	4		7	4		23	4	
		GA	GENO	48	42	0.398	48	30	0.276	42	30	0.811	90	30	0.453
		AA		43	31		42	27		31	27		73	27	
22		T	ALLELE	62	50	0.732	62	37	0.901	50	37	0.897	112	37	1.000

	ADORA2A (rs2298383 (T/C))	C		150	110		150	85		110	85		260	85	
		TT		8	7		8	5		7	5		15	5	
		CT	GENO	46	36	0.924	46	27	1.000	37	27	1.000	82	27	1.000
		CC		52	37		52	29		37	29		89	29	
22	**ADORA2A (rs3761422 (T/C))**	T	ALLELE	58	38	0.474	58	25	0.189	38	25	0.565	96	25	0.275
		C		154	122		154	97		122	97		276	97	
		TT		9	1		9	1		1	1		10	1	
		TC	GENO	40	36	0.076	40	23	0.185	36	23	0.747	76	23	0.424
		CC		57	43		57	37		43	37		100	37	
22	**ADORA2A (rs2267076 (T/C))**	T	ALLELE	59	39	0.478	59	25	0.151	39	25	0.475	98	25	0.228
		C		153	121		153	97		121	97		274	97	
		TT	GENO	9	1		9	1		1	1		10	1	
		TC		41	37	0.075	41	23	0.185	37	23	0.611	78	23	0.368
		CC		56	42		56	37		42	37		98	37	
22	**ADORA2A (rs2236624 (T/C))**	T	ALLELE	4	1	0.396	4	3	0.709	1	3	0.319	5	3	0.415
		C		208	159		208	119		159	119		367	119	
		TT	GENO	0	0		0	0		0	0		0	0	
		TC		4	1	0.393	4	3	0.707	1	3	0.316	5	3	0.412
		CC		102	79		102	58		79	58		181	58	
22	**ADORA2A (rs5751876 (C/T))**	C	ALLELE	70	58	0.582	70	47	0.341	58	47	0.711	128	47	0.446
		T		142	102		142	75		102	75		244	75	
		CC		9	10		9	8		10	8		19	8	
		CT	GENO	52	38	0.697	52	31	0.568	38	31	0.890	90	31	0.657
		TT		45	32		45	22		32	22		77	22	

APPENDIX H: Plagiarism report



1701006:Thesis-without-Ref-TurnITin.docx

ORIGINALITY REPORT

17%	7%	11%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"2013 Annual Meeting Abstract Supplement", Arthritis & Rheumatism, 2013. Publication	1%
2	Submitted to University of Witwatersrand Student Paper	1%
3	docplayer.net Internet Source	1%
4	"2015 ACR/ARHP Annual Meeting Abstract Supplement", Arthritis & Rheumatology, 2015. Publication	<1%
5	"ACR Meeting", Arthritis & Rheumatism, 2012. Publication	<1%
6	Submitted to University of Sheffield Student Paper	<1%
7	"2016 ACR/ARHP Annual Meeting Abstract Supplement", Arthritis & Rheumatology, 2016 Publication	<1%
8	Submitted to University of Hong Kong Student Paper	<1%