Human leukocyte antigen (HLA) genetic diversity in South African populations

By

Mqondisi Tshabalala

Student no. 15033075

Submitted in partial fulfillment of the requirements for the degree

Doctor of Philosophy (Medical Immunology)

In the Faculty of Health Sciences,
School of Medicine
University of Pretoria

Supervisor: **Professor Michael S Pepper**
Co-supervisor: **Professor Alan Christoffels**

2018

DECLARATION

UNIVERSITY OF PRETORIA

DECLARATION OF ORIGINALITY

This document must be signed and submitted with every essay, report, project, assignment, dissertation and / or thesis.

Full names of student: MQONDISI TSHABALALA

Student number: 15033075

Declaration

1. I understand what plagiarism is and am aware of the University's policy in this regard.

2. I declare that this THESIS/DISSERTATION (eg essay, report, project, assignment, dissertation, thesis, etc) is my own original work. Where other people's work has been used (either from a printed source, Internet or any other source), this has been properly acknowledged and referenced in accordance with departmental requirements.

3. I have not used work previously produced by another student or any other person to hand in as my own.

4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

SIGNATURE OF STUDENT: Mqondisi

SIGNATURE OF SUPERVISOR: Michael Sil

SUMMARY

There is documented evidence of high genetic diversity amongst African populations, but there is limited data on human leukocyte antigen (HLA) diversity in these populations. HLA genes are highly polymorphic, and encode for proteins that are part of the host defence mechanism mediated through antigen presentation to immune system effector cells. The highly polymorphic nature of HLA genes facilitates the presentation of a wide range of antigenic peptides to the immune system leading to an immune response. With the high disease burden in Africa, it is important to fully understand HLA diversity in these populations, to establish HLA-disease associations, and potentially use this data for the informed design of population-specific vaccines against the many diseases, and to improve on donor-recipient matching. The aim of this thesis is to understand HLA diversity in South African populations to support transplantation programs, add knowledge on human diversity and build a potential future resource for disease association and population studies.

There is generally limited HLA data from southern African populations (Chapter 2) to support disease association studies, provide guidance in vaccine design and donor recruitment for transplantation programs. Despite being the only active bone marrow donor registry in Africa supporting transplantation programs, HLA diversity in volunteer bone marrow donors registered at the South African Bone Marrow Registry (SABMR) is largely undocumented. This study documents HLA -A, -B, -C, -DRB1 and -DQB1 allele and haplotype frequencies from a subset of 237 SABMR registered donors with the objective of highlighting HLA diversity in South Africans (Chapter 3). Additionally, mixed resolution HLA data from the National Health Laboratory Services (NHLS) and the South African National Blood Transfusion Service (SANBS) are reported (Chapter 4). A comparison of South African HLA data (NHLS and SANBS) with other global populations including sub Saharan Africans confirm the genetic diversity of South Africans. To counter the paucity of HLA data, *in silico* HLA imputation tools may be used to determine HLA alleles from existing whole genome sequencing (WGS) data. HLA imputation is an economically feasible typing option for resource limited settings. To support the feasibility of HLA imputation, this study describes high resolution (up to 8 digit typing) HLA alleles determined by *in silico*

HLA imputation tools from 24 WGS of South African individuals (chapter 5). Generally, HLA diversity of South African populations is described in detail through literature meta-analysis, documentation of previously typed individuals (SANBS, NHLS and SABMR) and HLA imputation from existing next generation sequencing (NGS) data. Although results reported here are from a small subset of 237 SABMR registered donors (chapter 3), 24 WGS (chapter 5) and mixed resolution typing NHLS and SANBS data (chapter 4), allele and haplotype frequencies generated could be a useful resource for future anthropological and population genetics studies. Furthermore, these findings may better inform donor recruitment strategies for the SABMR, and disease association studies. Future study recommendations include development of an HLA diversity resource for African populations, a comparison of large SABMR dataset with other global registries, and using more robust assembly based computational tools to fully understand the HLA diversity in South Africans.

Key words: HLA, diversity, imputation, mixed resolution, disease burden, population genetics, transplantation

THESIS OUPUTS

Peer reviewed publications

- **Tshabalala M**, Mellet J, Pepper MS. Human Leukocyte Antigen Diversity: A Southern African Perspective. *J Immunol. Res.* 2015;2015:746151. doi: 10.1155/2015/746151. Epub 2015 Aug 12.
- **Tshabalala M**, Ingram C, Schlaphoff T, Borrill V, Christoffels A, Pepper MS. Human Leukocyte Antigen-A, B, C, DRB1, and DQB1 Allele and Haplotype Frequencies in a Subset of 237 Donors in the South African Bone Marrow Registry. *J. Immunol. Res*. 2018 Apr 23;2018:2031571. doi: 10.1155/2018/2031571. eCollection 2018.
- Mellet J, **Tshabalala M**, Agbedare O, Meyer PWA, Gray CM, Pepper MS. Human leukocyte antigen (HLA) diversity and its clinical applications in South Africa. **ACCEPTED** in South African Medical Journal (manuscript number: SAMJ13825)

Manuscript(s) UNDER review

- **Tshabalala M**, Vather K, Nelson D, Mohamed F, Christoffels A, Pepper MS. Mixed resolution HLA~A, ~B, ~C, ~DRB1, ~DQA1, ~DQB1 and ~DPB1 diversity in South African population. **under review** in BMC Medical Genetics (manuscript number MGTC-D-19-0028)

Conference Presentation(s)

- Faculty Day, University of Pretoria, Faculty of Health Sciences. 18-19 August 2015

  POSTER:

  Tshabalala M and Pepper MS. Lack of human leukocyte antigen (HLA) diversity data in South Africa: implications for transplantation and disease association studies.
- International Tissue Banking Conference, SATiBA, African Pride, Irene Country Lodge, Pretoria, 17-18 September 2015

POSTER:

Tshabalala M, Mellet J, Pepper MS. Lack of human leukocyte antigen (HLA) diversity data in South Africa: implications for transplantation and disease association studies.

- Genomics Research Institute Seminar, Auditorium – Plant Science, University of Pretoria 15th October 2015

  POSTER:

  Tshabalala M, Mellet J, Pepper MS. Towards understanding human leukocyte antigen (HLA) diversity in southern African populations.

- Young Researchers Forum for the 9th Congress of the African Society of Human Genetics. May 15-17, 2016 - Dakar, Senegal

  POSTER:

  Tshabalala M, Mellet J, Christoffels A, Pepper MS. Towards understanding human leukocyte antigen (HLA) diversity in southern African populations: implications for transplantation and disease association studies.

- South African Tissue Bank Association (SATiBA) conference. Leriba Hotel, Centurion, South Africa, 05-06 October 2017

  ORAL PRESENTATION:

  Tshabalala M, Ingram C, Schlaphoff T, Borrill V, Christoffels A, Pepper MS. HLA-A, -B, -C, -DRB1, and –DQB1 allele and haplotype frequencies from donors in the South African Bone Marrow Registry (SABMR).

- SAMRC Flagship Conference, Stem Cell Research and Therapy, Innovation Hub, Pretoria, 26-27 October 2017

  ORAL PRESENTATION:

  Tshabalala M, Christoffels A, Pepper MS. HLA diversity in South Africa: Insights from individuals typed at varying resolution by NHLS.

ACKNOWLEDGEMENTS

I deeply thank the following people for their support, encouragement and contributions to this thesis

- My supervisor, Prof M.S. Pepper, for professional guidance and unwavering support throughout my studies. Thank you for your patience and understanding; and giving me the freedom and opportunity to “think outside the box’. Most importantly, I am grateful for your mentorship and guidance.
- My co-supervisor, Prof A Christoffels, I am very grateful for all your critical review of my research, your excellent bioinformatics expertise and guidance.
- Other key players: Prof Fourie Joubert, thank you your bioinformatics skills and assistance, Peter van Heusden thanks for all those sessions and teaching me the ropes at SANBI. Melvin Ambele “my chief” thanks for everything brother. SANBS team (Kuben, Derrick), SABMR team (Charlotte, Charlotte Ingram, Terry Schlaphoff, and Veronica Borrill) a big thanks you to you all.
- My lab mates, thank you for the friendship, support and the ever available coffee (‘sanity juice’)
- To my wife Tracey (TTT), thank you for the support, love and understanding all those days away from home. Thank you for enduring You are my soul mate
- To my granny (Sihupulo), parents eMshengu eMavuso, (Jonathan and MaThobela) eMfidi, eNtandela, brothers (Mnce, Jethro, Skhu), my only sister (Mthunzi), ‘my kids’ Tanatswa, Mxolisi, Mbonisi, Nomaswazi and Nozinhle, this is for you. Thank you for believing in me.
- Above ALL I give thanks to God, from whom I draw strength, knowledge and wisdom,

I also thank the following institutions and funding bodies

- University of Pretoria Postgraduate Research Bursary
- South African Medical Research Council (SAMRC) through the Extramural Stem Cell Unit and University Flagship program
- Institute for Cellular and Molecular Medicine

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AFND	Allele Frequency Net Database
AIDS	acquired immunodeficiency syndrome
BAM	binary version tab delimited txt file with sequence alignment data
BFF	Burkina Faso Fulani
BFM	Burkina Faso Mossi
BFR	Burkina Faso Rimaibe
BMDW	Bone Marrow Donors Worldwide
Bots	Botswana
Bp	base pairs
CaB	Cameroon Bamileke
CARMP	Central African Republic Mbenzele Pygmy
CBkP	Cameroon Bakola Pygmy
CBP	Cameroon Baka Pygmy
CBt	Cameroon Beti
CSw	Cameroon Sawa
CTL	cytotoxic T lymphocyte
CW-EUR	Central and West Europe
CYT	cytoplasmic domain
DNA	deoxyribonucleic acid
EGA	The European Genome-phenome archive
EM	expectation-maximization
Exp Het	expected Heterozygosity
F_{ST}	population differentiation
GGA	Ghana Ga-Adangbe
GVHD	graft versus host disease
Hg19	human reference genome assembly version 19
HIV	human immunodeficiency virus
HLA	human leukocyte antigen
HLA-HD	High-quality Dictionary
HSCT	hematopoietic stem cell transplantation

HWE	Hardy-Weinberg equilibrium
IBD	identity by descent
IMGT HLA	ImMunoGeneTics project/human leukocyte antigen
KEN	Kenya
KENL	Kenya Luo
KENNy	Kenya, Nyanza Province, Luo tribe
KIR	killer-cell immunoglobulin-like receptors
LD	linkage disequilibrium
MAC	multiple allele code
Mb	mega base
MHC	major histocompatibility complex
Moza	Mozambique
NAFR	Northern Africa
NE-EUR	Northeast Europe
NGS	next generation sequencing
NHLS	National Health Laboratory Services
NJ	Neighbour-Joining
NK	natural killer
NMDP	National Marrow Donor Program
NMDS	non-metrical multidimensional scaling
Obs Het	observed heterozygosity
OTH	other European populations of recent origin
PCA	principal component analysis
PCR-SSO	polymerase chain reaction sequence specific oligonucleotide
PCR-SSP	polymerase chain reaction sequence specific primer
p-HWE	*p* value for HWE deviation
PSA	HLA simulated data
RMX	South African Mixed ancestry
RNA	ribonucleic acid
RNAseq	ribonucleic acid sequencing
RSA	Republic of South Africa
RWA	Rwanda

SAB	previously published HLA data from South African Bone Marrow Registry
SABMR	South African Bone Marrow Registry
SAHGP	Southern African Human Genome Program
SAI	South African Indian population
SANBS HREC	South African National Blood Services Human Research Ethics Committee
SANBS	South African National Blood Transfusion Service
SANT	South Africa Natal Tamil
SANZ	South Africa Natal Zulu
SBT	DNA sequencing based HLA typing
SE-EUR	Southeast Europe
SenMAND	Senegal Niokholo Mandenka
SNPs	single nucleotide polymorphisms
SoAB	South Africa Black
SoAC	South Africa Caucasians
SSOP	sequence specific oligonucleotide primer
SSP	sequence specific primer
TA GVHD	transfusion associated graft versus host disease
TB	tuberculosis
TCR	T cell receptor complex
Th	T helper
TM	transmembrane domain
TRALI	Transfusion related lung injury
UgaKam	Uganda Kampala
UgaKam2	Uganda Kampala second population
UTR	untranslated region
WASI	Western Asia
WES	whole exome sequences
WGS	whole genome sequencing
WHO	World Health Organization
W_n	Cramer's V Statistic
WOR	South Africa Worcester
ZaL	Zambia Lusaka HLA data

Zam	Zambia
ZiHS	Zimbabwe Harare Shona
Zim	Zimbabwe

CHAPTER 1

LITERATURE REVIEW

1.1 General Introduction

The African population is genetically diverse[1] with several pointers indicating that the continent is the cradle of humankind[2,3]. Despite this genetic diversity, there is scarce or no information on human leukocyte antigen (HLA) diversity in most African nations, thereby limiting our understanding of human health and susceptibility to disease. In general, genetic diversity of African populations is poorly understood[4]. South Africa has an admixed population giving rise to high genetic diversity[5,6], hence the need for further analysis/evaluation of the national diversity to map disease association and theraupetic gene targets and facilitate vaccine development. Despite the general similarities in culture and shared geographical location, genetic differences exist among populations at every 1000 base pairs[7]. The South Africa human population is predominantly of Bantu ethnicities; additionally, there are populations of mixed ancestry characterised by high diversity in cultural and ethno-linguistic structures (https://en.wikipedia.org/wiki/Bantu_peoples).

The highly polymorphic human leucocyte antigen (HLA) gene region on the short arm of chromosome 6 is divided into class I, II and III gene loci. Classes I and II form the classical (major) HLA molecules while class III are HLA related molecules critical to the human immune system. Figure 1.2 summarizes the genetic structure of classical HLA class I and II. HLA class I molecules, expressed on all nucleated cells, encode membrane bound glycoproteins that bind to endogenous antigenic epitopes and present them to $CD8^+$ T lymphocytes. On the other hand, class II molecules are expressed on all antigen presenting cells, and present antigenic peptides to $CD4^+$ T lymphocytes. The polymorphic nature of HLA genes allows the presentation of a wide range of peptides to the immune system. Each individual has unique HLA alleles inherited from both parents, hence the gene loci can be used in vaccine

development, transplantation and understanding susceptibility, resistance and progression of human diseases.

South Africa has a heterogeneous population, whose HLA genetic diversity has not been well described, despite the immunological significance of HLA. Paximadis and colleagues[8] showed a broad spectrum of distribution of HLA alleles among black South Africans compared to their white counterparts[8]. HLA diversity in South African populations is still not conclusively known, mainly due to the expense in HLA typing methods, a few studies have reported HLA data. There is generally limited high resolution HLA typing from South African individuals which impacts on our understanding of HLA disease association dynamics, and support of transplantation programs through donor-patient HLA matching. Owing to the unknown HLA genetic diversity of South African populations, it is currently difficult to find an HLA match for individuals needing hematopoietic stem cell transplantation. This study seeks to quantify HLA genetic diversity amongst South African populations. The overall study aim is to describe the HLA alleles present and to quantify classical HLA diversity in South Africa with the view to providing a resource for understanding disease pathogenesis, vaccine development and for easier matching of donor-recipient haplomatches, and also as a baseline towards establishment of biobanks for future medical research.

1.2 Problem statement

There is a wide information gap on HLA genetic diversity in South Africans, which this study intended to address. Previous studies are mostly based on disease association datasets[9-14], limited sample size[15], targeted sampling[16,17] and a few high resolution HLA typing studies[8,18-22]. Within South Africa, there is documented evidence of an old human lineage which might be ancestors to modern humans. These founder populations are known to be genetically diverse. Additionally, there is a high infection and disease burden in South Africa, coupled with limited knowledge on genetic diversity in genes coding for the immune system. HLA diversity data from these populations might add to our knowledge on HLA disease association and guide in population specific vaccine design strategies and better inform donor

recruitment strategies into bone marrow registries. It is generally not easy to pinpoint a specific allele (or allele combination) association to a disease especially when data from healthy individuals is not available for inference. There is a need for vaccines for the many diseases/infections in the South African population. Furthermore, population HLA diversity data will help understand immune escape mutants which drive drug resistance infections, and support population genetic studies highlighting evolutionary selection pressures like disease epidemics.

1.3 Literature review

1.3.1 Basic Immunology

The human immune system's ability to recognize 'self' and 'non self' forms a key concept in clinical immunology and host defense against pathogens. Host immune defense can be divided into three broad categories namely mucosal and epithelial barriers, the innate immune and the adaptive immune systems. Mucosal and epithelial barriers offer physical protection through an impermeable layer of cells coupled with antimicrobial secretions and maintain tolerance to commensal microbiome. If a pathogen crosses a physical barrier, the innate immune system is the next line of defense against invading pathogens. The innate system is characterized by a variety of cells circulating in blood (macrophages, neutrophils, mast cells), non-specific killing of pathogens and lack of immunological memory. The adaptive immune system on the other hand is pathogen specific, and has immunological memory. A second encounter with the same pathogen activates the memory cells of the adaptive immune system to elicit an immune response. Both the innate and adaptive immune systems have an antigen recognition phase by antigen presenting cells followed by an effector phase. T-cell based adaptive immune responses are based on antigen presentation to the T cell receptor complex (TCR) by the major histocompatibility complex (MHC), leading to an antigen specific immune response[23].

The MHC genes, also known as the human leukocyte antigen (HLA) loci in humans, are found on chromosome 6, and encode cell surface glycoproteins broadly classified into three classes: HLA class I, II and III. Class III molecules include inflammatory proteins, complement proteins, regulatory receptors and other gene products not directly involved in antigen presentation. Class I and II molecules' primary role is antigen presentation to effector T cells. There are a high number of genetic polymorphisms in class I and II molecules, with multiple alleles at each locus. There are currently 20 088 HLA alleles listed in the IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla/stats.html release 3.34.0 October 2018), of which 14 800 are class I and 5 288 are class II alleles (summarized in Figure 1.1)[20]. The high diversity facilitates presentation of many antigens but is a challenge in matching donors and recipients in transplantation[23]. There is generally an increase in the number of known HLA alleles with time (Figure 1.1) owing to advancement in molecular methods.

1.3.2 HLA class I and II structure

HLA class I consists of glycosylated α1, α2 and α3 chains (encoded on chromosome 6) and non-covalently bound to β2 microglobulin (encoded on chromosome 15) which assemble to form a functional receptor on most nucleated cells. The hyper variable α1 and α2 domains form the antigen binding groove of the HLA class I molecules, which present processed antigens to effector CD8 T lymphocytes. Some HLA class I molecules interact and regulate natural killer (NK) cell function through the killer-cell immunoglobulin-like receptors (KIR)[24]. There are 3 major HLA class I genes (classical HLA class I): HLA-A, HLA-B and HLA-C; minor genes include HLA-E, HLA-F and HLA-G. Figure 1.2 shows the structures of class I and II molecules, including the linear genetic structure showing the number of coding regions (exons). HLA class II molecules are heterodimers of α and β (α1, α2 and β1, β2) chains anchored in the cytoplasm by transmembrane domains in the α2 and β2 chains. The hyper variable α1 and β1 chains of class II molecules form the antigen binding groove of class II molecules (Figure 1.2). HLA Class II α and β heterodimers have the alpha subunit encoded by the "A" or "A1" loci and the beta subunit is encoded by "B" or "B1", resulting in HLA DPA1 and HLA DPB1 for HLA DP and HLA DQA1 and

HLA DQB1 for HLA DQ gene loci. On the other hand, the HLA-DR gene locus is more complex; the alpha chain is encoded by a single HLA-DRA gene (with few minor variants), while the beta subunit is encoded by the HLA-DRB1 locus and other minor loci which are variable amongst individuals (HLA-DRB3, -DRB4, -DRB5). Class II restricted antigens are presented to effector CD4 lymphocytes[23,25]. HLA polymorphisms are highest in the antigen binding grooves of both class I and II molecules[26] (Figure 1.2). MHC restricted antigen presentation was first demonstrated by Zinkernagel and Dougherty in 1974[27], with antigen binding specificities based on amino acid sequences at the antigen binding groove of the HLA molecules.

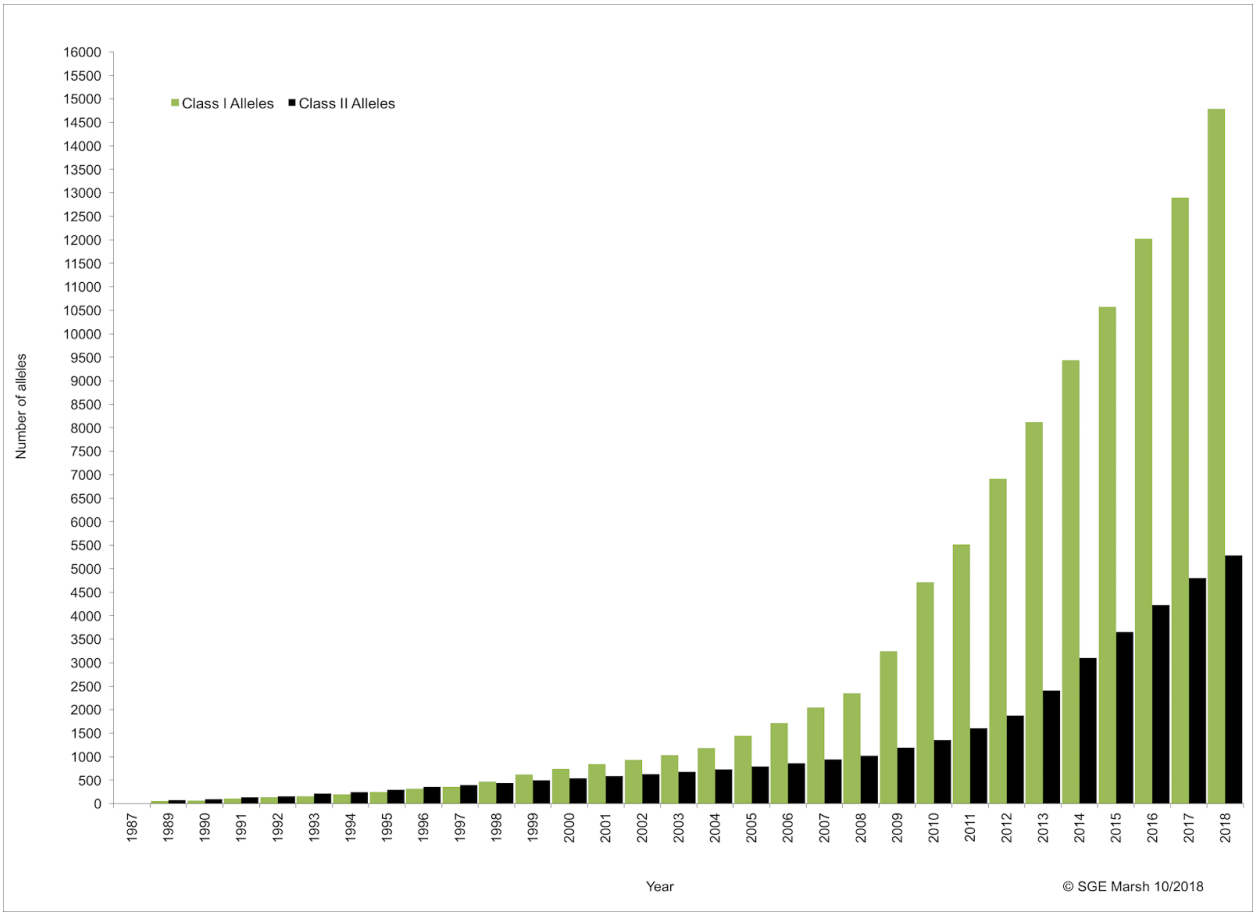

Figure 1.1 The number of known class I and II alleles overtime

The number of HLA alleles has been increasing since 1987 due to advancement in typing methods. There are currently more than 14 000 and 5000 class I (green bars) and II (black bars) alleles respectively in the IMGT HLA database (Figure from[20]).

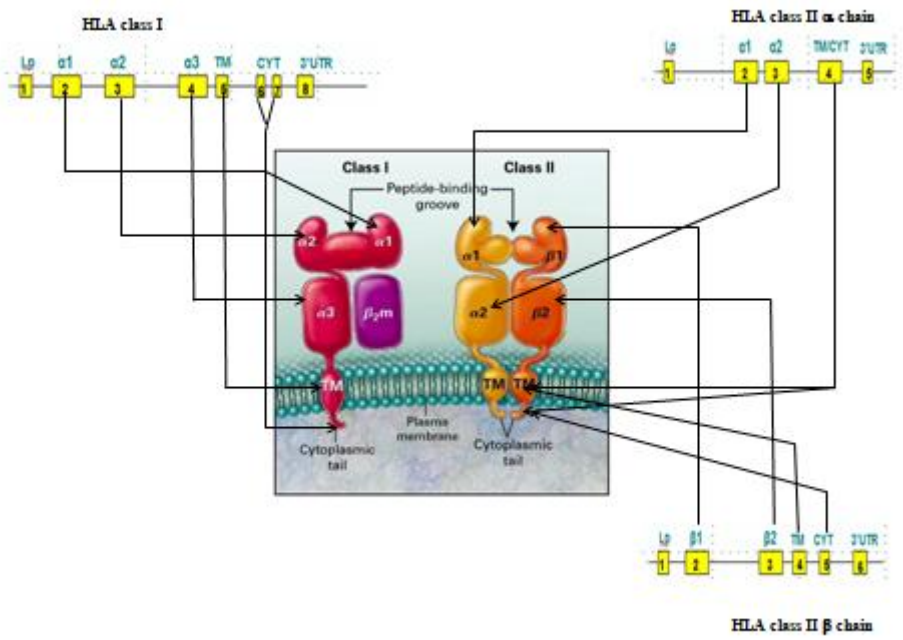

Figure 1.2 HLA class I and class II structures

HLA class I molecules have 8 exons, whilst class II molecules have 5 exons (α chain) and 6 exons (β chains). The general structure includes leader peptide (Lp), α chain, β chains, transmembrane domain(TM), cytoplasmic domain (CYT) and 3' untransalated region (3UTR). Figure was adapted from[26].

1.3.3 HLA nomenclature

The HLA nomenclature uses a unique set of numbers to identify each allele in the IMGT/HLA database[20,21] (Figure 1.3). The naming shows the specific gene locus name (for example HLA A in Figure 1.3), with the first set of digits (Field 1) corresponding to an allotype (antigen level). Field 2 (Figure 1.2) corresponds to the subtype (allele level); the numbers are assigned in order of the DNA sequence discovery within a group. Different allele level numbers correspond to differences in one or more single nucleotide polymorphisms (SNPs) leading to amino acid sequence differences between two related alleles (Field 3). Alleles differing in the

non coding regions including introns, 3′ and 5′ untranslated regions (UTR) have an additional set of numbers (Field 4)[20,21]. Additionally, expression status and level of a protein of a particular allele may be indicated as shown in Figure 1.3.

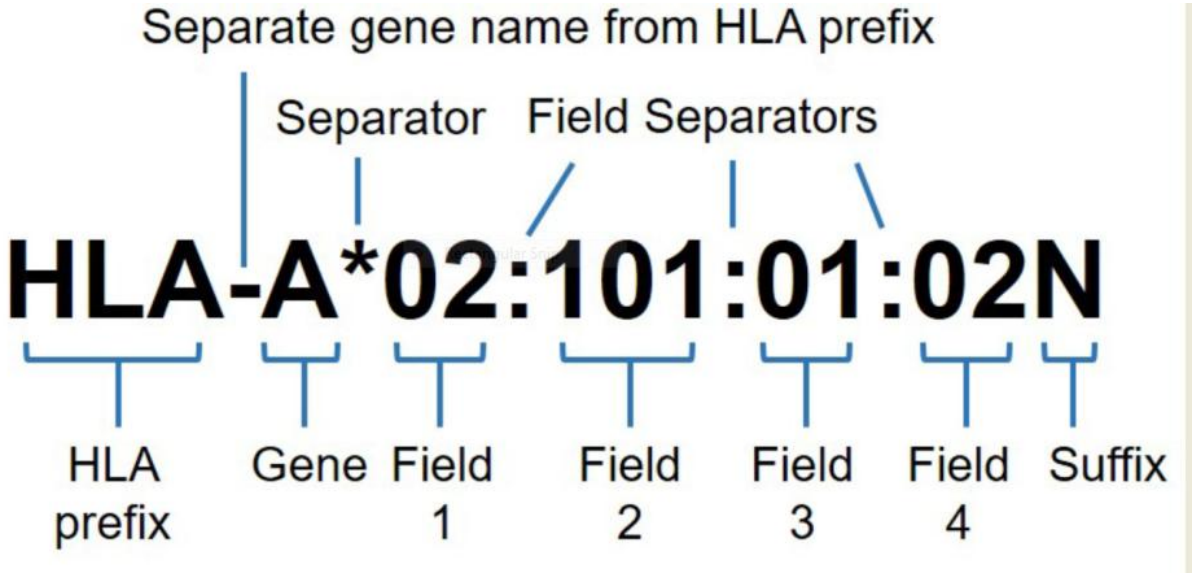

Figure 1.3 HLA nomenclature

HLA prefix (identifies HLA gene region), Gene (specifies the HLA gene locus), Field 1 (antigen group), Field 2 (specific HLA protein/specific HLA allele), Field 3 (Synonymous DNA substitution in coding region), Field 4 (DNA changes in non coding region), Suffix (denotes changes in expression, possible suffices include N=Null, L=Low, S=secreted, A=Aberrant and Q=Questionable). (Adapted from S.G.E Marsh, HLA Informatics Group[20,28].

1.3.4 HLA typing methods

HLA typing methods have evolved from phenotypic identification using serology methods to high resolution DNA sequencing based technologies. Serology based methods identified HLA molecules at the antigen level (Figure 1.3), with DNA methods being able to identify to the protein level as summarized in Figure 1.3. Serology typing methods are based on the detection of expressed HLA molecules on cell surfaces (T cells for HLA class I, and B cells for HLA class II) through use of antisera panels (usually sourced from multiparous women) in a complement-

dependant cytotoxixity test. The compliment-mediated microlymphocytotoxicity method has commonly been used as a serology gold standard in HLA typing[29]. Limitations of serology based HLA typing include i) low resolution results which are applicable for renal but not adequate for bone marrow transplantation, ii) live lymphocytes are needed for the assay, but cell numbers might be very low in some patients, iii) sera cross reactivity, and iv) limited availability of sera.

DNA based HLA typing are polymerase chain reaction (PCR) based molecular methods developed to overcome the low resolution typing of serology methods. There are several DNA based molecular typing methods, with the following being the most common broad categories i) sequence-specific primer (SSP) ii) sequence-specific oligonucleotide probe (SSOP) and iii) sequence-based typing (SBT). The principle of SSP is based on a complete primer matched to a specific HLA allele(s), leading to the amplification of the allele sequence which can be detected by gel electrophoresis. This method is labor intensive and expensive for high throughput HLA typing. Additionally, with the ever increasing number of HLA alleles there is a need to constantly update HLA typing primers. SSOP, more suited for high throughput HLA typing, is based on allele specific panels of synthetic oligonucleotide probes which hybridize HLA allele PCR products. Despite the potential in high throughput HLA typing, SSOP still needs to cope with the ever increasing allele numbers in designing probes. SBT using Sanger sequencing has been a long time gold standard molecular HLA typing method following the discovery of locus and antigen specific polymorphisms in non-coding introns flanking the polymorphic HLA exons (reviewed in[30]). Despite the ability of SBT methods to give high resolution results, limitations include typing certain exons within the HLA loci, thereby giving partial sequences of about 10% of the reported alleles[31]. Clinical HLA typing laboratories rarely sequence exons/introns outside the peptide binding groove for transplantation matching, with the assumption that they are not directly involved in T cell allo-recognition. This assumption is supported by modeling HLA/peptide/T-cell receptor (TCR) interactions[32], and studying allele specific peptide repertoires[33] and other allo-recognition studies[34-37]. Routinely typed exons include exons 2 and 3 for HLA class I and exon 2 for HLA class II (Figure 1.2). Additionally, there is heterogeneity from SBT HLA analysis yielding limited resolution data, making it difficult to correctly assign HLA types. It is possible though to sequence the whole

HLA gene region (coding exons and introns as summarised in Figure 1.2) using current SBT methods, but at a very high cost and requiring expert analysis. Furthermore, as SBT focuses primarily on the selected exons, together with the phasing problem (common in whole-genome assembly), the individual base differences are assigned unambiguously to one of the chromosome (*cis/trans* assignment of DNA bases) in a heterogeneous sample[38].

Advances in next generation sequencing (NGS) HLA typing allow high throughput, with high resolution HLA results in a relatively shorter time frame compared to SBT typing[31,39,40]. NGS HLA typing addresses the inherent phasing ambiguities in SBT Sanger sequencing. With NGS, two chromosomes produce separate reads, and when supported by a strong bioinformatics workflow can separate these reads and assemble them into phased consensuses. The highly polymorphic nature of the HLA gene region together with the high number of pseudogenes and indels contribute to NGS HLA typing challenges. Additionally, the short sequencing reads generated by NGS platforms are difficult to align to reference HLA alleles in the IMGT/HLA database[20,21]. The complex nature of some HLA loci impacts negatively on NGS read alignment to the reference alleles, hence accuracy of typing results becomes less reliable[41]. Most reference HLA allele sequences in the IMGT/HLA database[20,21] have partial sequences[42] making it difficult to accurately call HLA alleles. Quantifying HLA diversity in genetically diverse populations like Africans might contribute to full length reference HLA sequences[39,43].

Despite the advances in HLA typing methods, it is possible to obtain ambiguous results (combinations of several alleles as a result instead of a desired single pair) and inaccurate typing results which impact on HLA clinical applications. PCR forms an integral part of HLA typing including NGS library preparation and the actual sequencing step. Possible PCR sources of HLA genotyping ambiguities are usually the results of i) signal loss due to amplification imbalance or dropout and ii) mixed signals caused by PCR crossover artifacts or PCR stutter that create a mix of artificial alleles in vitro that makes allele selection difficult. Allele dropout can be grouped into three main types: a) complete allele drop out (locus dropout), b) only one allele amplified, with PCR signal for the other allele missing completely (allele dropout) and c) one or both alleles being partially amplified and sequenced (partial

dropout) as reviewed in[44]. PCR primers can unequally amplify HLA alleles leading to an imbalance between the two chromosomes, hence affecting HLA genotyping result. SBT Sanger sequencing methods use a threshold of about 5–20% for the minor signal while NGS-based HLA-typing methods can detect as low as 2% of the minor signal[45]. The high polymorphic nature of the HLA region makes the design of primers difficult; novel variants around the primer binding sites might affect the amplification process. Allele dropout can be due to a technical error, and in some cases due to disease state, for example, false homozygous HLA typing results are common in some cancers due to chromosome 6 loss in cancer affected cells[46]. Additionally, the amplification of short tandem repeats (STRs) in the HLA region results in PCR stutter[47] which might contribute to ambiguity between two alleles that only differ in this STR region.

Generally, SBT Sanger sequencing can produce 1000 base-pair long reads, but the signals from the two chromosomes are mixed leading to an inherent phase ambiguity. On the other hand, most NGS platforms separate reads from different chromosomes to overcome the phasing problem, but with shorter reads than SBT (reviewed in[44]). False homozygous typing is common if an allele pair has a homozygous sequence stretch which is longer than the average NGS read length and the insert between the pairs, leading to unresolved chromosome phasing. Although still under clinical application evaluation, Pacific Biosciences SMRT technology produces longer NGS reads that can cover the whole HLA locus with a single read[48] Based on the codominant expression of HLA alleles, and the Mendelian fashion of HLA haplotype inheritance, family studies can be used to confirm/discard homozygous typing results. Two siblings have a 25% chance of HLA genotype identity, 50% chance of being haploidentical (share one haplotype), and a 25% chance of not sharing a common haplotype[25]. Standardized high quality HLA typing methods form an integral part of the clinical use of HLA results.

1.3.5 HLA imputation

Based on high linkage disequilibrium (LD) within the MHC region, HLA alleles can be determined using *in silico* computational tools by inferring them from surrounding

HLA allele associated SNPs[49]. Additionally NGS generated whole genome sequences (WGS) and whole exome sequences (WES) as well as RNA sequence data (RNAseq) are increasingly used for HLA imputation[50-54]. HLA imputation is a potentially cheaper method for understanding population HLA diversity through the use of existing datasets (SNPs, WES, WGS, RNAseq). Several projects aimed at understanding genetic diversity of African populations [for example Southern African Human Genome Program (SAHGP)[55,56], H3 Africa (https://h3africa.org/), 1000 Genomes project (http://www.internationalgenome.org/)[57], African Genome Variation Project[58]] are potential data sources for HLA imputation. Despite the high imputation accuracy reported by several methods, these tools are good to augment, but not replace routine HLA typing methods in understanding HLA diversity.

1.4 Applications of HLA genetic data

1.4.1 Transplantation and transfusion

Transplantation as a therapeutic intervention requires a match between donor and recipient HLA molecules so as to decrease the chance of rejection[23]. The chance of two individuals having identical HLA molecules on all loci is very low. Siblings have a 25% chance of being HLA-identical due to HLA being codominantly expressed and inherited as haplotypes from both parents. The degree of HLA matching is a predictor of clinical outcome. Acute graft versus host disease (GVHD) is an immunocompetent donor T-cell mediated response against the recipient's immune system which is common in unmatched donor recipient pairs. Acute GVHD can be reduced by donor T-cell depletion, but this increases the risk of rejection, malignant disease relapse and impaired immune recovery[59,60]. In addition to HLA matching, other genes like the killer inhibitory receptors (KIRs) have been documented to affect the clinical outcome of allogeneic transplantation[61-64]. In severely immunocompromised individuals, allogeneic transfusion with immune competent T-cells containing blood products might lead to transfusion associated GVHD (TA GVHD). Transfusion related lung injury (TRALI) is an anti-HLA (mostly class I[65,66]) antibody related complication which might be fatal. Anti-HLA class II antibodies induce TRALI through monocyte and subsequent neutrophil activation[65,67]. Anti-HLA

class I antibodies have been reported to be a cause of neonatal alloimmune thrombocytopenia together with platelet derived specific antigens[68]. Generally, it is critical to know the population HLA diversity to improve donor recipient matching in both transplantation and transfusion, while recruitment of donors from minority populations also helps improve HLA diversity in registries[25]

1.4.2 Disease association

The World Health Organization (WHO) reports a high burden of disease in southern African populations, with human immunodeficiency virus (HIV), tuberculosis (TB) and malaria being the priority problems[69]. Southern African (including South African) populations are documented to be highly genetically diverse[70]. There is however limited information on the genetic diversity in genes coding for immune system including HLA genes[71]. Several autoimmune conditions have been directly associated with specific class I and II HLA alleles, including rheumatoid arthritis, multiple sclerosis, ankylosing spondylitis, Grave's Disease and many more as reviewed by Trowsdale and Knight[72]. HLA association with infectious disease including HIV has been documented, in which several alleles have been associated with varying rates of HIV disease progression[73-76]. HLA in susceptibility, transmission and treatment outcomes in HIV has also been reviewed[77]. The presence or absence of some HLA alleles and their frequencies has been associated with malaria burden in different populations[78]. High HLA -B*53:01:01 and -B*78:01 allele frequencies are reported to be associated with *Plasmodium falciparum* parasitemia, a human malaria causing parasite[79]. Several HLA alleles (mostly class II), have been reviewed to contribute to TB susceptibility and protection in various populations[80], highlighting the role of HLA in TB immunity. Despite the unclear link between HLA alleles and different infectious disease, it is imperative to understand HLA diversity in the highly disease burdened South African populations, particularly to support vaccine development. Identification of HLA restricted epitopes with protective immune correlates is critical in designing T-cell based vaccines against the many pathogens, especially for the South African populations. Furthermore, these epitopes can be analysed as potential vaccine candidates. To refine the identification of HLA restricted cytotoxic T lymphocyte (CTL) escape mutants, knowledge of HLA diversity

of large datasets is needed to statistically increase the power of the currently available CTL escape prediction maps[81]. It is important to map the immune escape pathways of several human pathogens to improve vaccine development strategies.

1.4.3 Population studies

There is a marked difference in HLA diversity distribution globally, with geographically separated regions showing varying amounts of diversity. Most HLA loci, except for HLA-DPB1, show high allele numbers across populations[18,82]. The global distribution of HLA diversity provides insight into human migration patterns, and could help understand past pathogen exposures[83] and other selection pressures. HLA genetic diversity studies have been used to trace the spread of modern humans from East Africa, and model co-evolution of genes and languages in African populations[84]. Interpretation of HLA in population studies can be improved by extensive knowledge of HLA diversity in different populations. Although there have been several efforts to understand global human genetic diversity including the Hap Map Project[85], 1000 Genomes Project[57] and the African Genome Variation Project[58], there is limited information on South African populations. Additionally, previous South African studies targeted populations like hunter gathers[17], some studies with small sample sizes[15]. Diverse and novel HLA alleles have been reported in sub Saharan populations (reviewed in[86]), including some novel HLA alleles from South African populations[8,87], which further supports the presence of high genetic diversity in Africans, and intra African diversity.

1.5 Aims and Objectives

1.5.1 Aim

Despite the documented evidence on genetic diversity of South African populations, there is limited information on HLA diversity. Lack of HLA diversity information impacts on donor-patient HLA matching for transplantation programs, disease

association and general genetic diversity. This study aimed to quantify HLA genetic diversity amongst South African populations.

1.5.2 Objectives

1. To determine the extent of lack of HLA diversity data for South African populations in the public domain. Chapter 2 addresses this objective.
2. To document HLA diversity in previously typed individuals in public healthcare delivery systems in South Africa. Chapters 3 and 4 address this objective.
3. To use *in silico* computational methods to determine high resolution HLA alleles from NGS WGS generated from South African individuals. Chapter 5 addresses this objective.

1.6 References

1. Disotell TR. Archaic human genomics. Am J Phys Anthropol. 2012;55:24-39.
2. Stewart JR, Stringer CB. Human evolution out of Africa: the role of refugia and climate change. Science. 2012;335(6074):1317-21.
3. Relethford JH. Genetic evidence and the modern human origins debate. Heredity. 2008;100(6):555-63.
4. Ramsay M. Africa: continent of genome contrasts with implications for biomedical research and health. FEBS Lett. 2012;586(18):2813-9.
5. Alessandrini M, Asfaha S, Dodgen TM, Warnich L, Pepper MS. Cytochrome P450 pharmacogenetics in African populations. Drug Metab Rev. 2013;45(2):253-75.
6. de Wit E, Delport W, Rugamika CE, Meintjes A, Moller M, van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum Genet. 2010;128(2):145-53.
7. Belle EM, Barbujani G. Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. Am J Phys Anthropol. 2007;133:1137-46.
8. Paximadis M, Mathebula TY, Gentle NL, Vardas E, Colvin M, Gray CM, et al. Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population. Human Immunol. 2012;73:80-92.
9. Alkharsah KR, Dedicoat M, Blasczyk R, Newton R, Schulz TF. Influence of HLA alleles on shedding of Kaposi sarcoma-associated herpesvirus in saliva in an African population. J Infect Dis. 2007;195(6):809-16.
10. Tikly M, Rands A, McHugh N, Wordsworth P, Welsh K. Human leukocyte antigen class II associations with systemic sclerosis in South Africans. Tissue Antigens. 2004;63(5):487-90.
11. Yang OO, Lewis MJ, Reed EF, Gjertson DW, Kalilani-Phiri L, Mkandawire J, et al. Human leukocyte antigen class I haplotypes of human immunodeficiency virus-1-infected persons on Likoma Island, Malawi. Hum Immunol. 2011;72(10):877-80.
12. Matthews PC, Listgarten J, Carlson JM, Payne R, Huang KH, Frater J, et al. Co-operative additive effects between HLA alleles in control of HIV-1. PloS ONE. 2012;7(10):19.
13. Carr DF, Chaponda M, Jorgensen AL, Castro EC, van Oosterhout JJ, Khoo SH, et al. Association of human leukocyte antigen alleles and nevirapine

hypersensitivity in a Malawian HIV-infected population. Clin Infect Dis. 2013;56(9):1330-9.

14. Shepherd BL, Ferrand R, Munyati S, Folkard S, Boyd K, Bandason T, et al. HLA Correlates of Long-Term Survival in Vertically Infected HIV-1-Positive Adolescents in Harare, Zimbabwe. AIDS Res Hum Retroviruses. 2015;6:6.

15. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science. 2009;324(5930):1035-44.

16. Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, et al. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. PLoS Genet. 2012;8(4):26.

17. Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012;338(6105):374-9.

18. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Research. 2011;39(1):D913-D9.

19. González-Galarza Faviel F, Takeshita Louise YC, Santos Eduardo JM, Kempson F, Maia Maria Helena T, Silva Andrea Luciana Soares d, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Research. 2015;43(D1):D784-D8.

20. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43(Database issue):20.

21. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. Nucleic Acids Research. 2013 January 1, 2013;41(D1):D1222-D7.

22. Sharp B, Kleinschmidt I, Streat E, Maharaj R, Barnes K, Durrheim D, et al. Seven years of regional malaria control collaboration-Mozambique, South Africa, and Swaziland. Am J Trop Med Hyg. 2007;76:42 - 7.

23. Chapel H, Haeney M, Misbah S, Snowden N. Essentials of Clinical Immunology. 6 ed. John Wiley and Sons L, editor. West Sussex: Wiley Blackwell; 2014.

24. Lanier LL. NK cell recognition. Annu Rev Immunol. 2005;23:225-74.

25. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Med J. 2007;48(1):11-23.

26. Klein J, Sato A. The HLA system. First of two parts. N Engl J Med. 2000;343(10):702-9.

27. Zinkernagel RM, Doherty PC. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. Nature. 1974;248(5450):701-2.

28. Robinson J, Malik A, Parham P, Bodmer JG, Marsh SG. IMGT/HLA database-A sequence database for the human major histocompatibility complex. Tissue Antigens. 2000;55(3):280-7.

29. Terasaki PI, Mccleland JD. Microdroplet assay of human serum cytotoxins. Nature. 1964; 204:998-1000.

30. Dunn PPJ.2011. Human leukocyte antigen typing:techniques and technology, a critical appraisal. International Journal of Immunogenetics. 2011;38(6):463-73.

31. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16(th) IHIW : review of HLA typing by NGS. Int J Immunogenet. 2013;40(1):72-6.

32. Xiao Y, Lazaro AM, Masaberg C, Haagenson M, Vierra-Green C, Spellman S, et al. Evaluating the potential impact of mismatches outside the antigen recognition site in unrelated hematopoietic stem cell transplantation: HLA-DRB1*1454 and DRB1*140101. Tissue Antigens. 2009;73(6):595-8.

33. Bade-Doeding C, Cano P, Huyton T, Badrinath S, Eiz-Vesper B, Hiller O, et al. Mismatches outside exons 2 and 3 do not alter the peptide motif of the allele group B*44:02P. Hum Immunol. 2011;72(11):1039-44.

34. Lauterbach N, Crivello P, Wieten L, Zito L, Groeneweg M, Voorter CEM, et al. Allorecognition of HLA-DP by CD4+ T cells is affected by polymorphism in its alpha chain. Molecular Immunology. 2014;59(1):19-29.

35. Lauterbach N, Crivello P, Wieten L, Zito L, Groeneweg M, Voorter CE, et al. Allorecognition of HLA-DP by CD4+ T cells is affected by polymorphism in its alpha chain. Mol Immunol. 2014;59(1):19-29.

36. Bettens F, Schanz U, Tiercy JM. Lack of recognition of HLA class I mismatches outside alpha1/alpha2 domains by CD8+ alloreactive T lymphocytes: the HLA-B44 paradigm. Tissue Antigens. 2013;81(6):414-8.

37. Crivello P, Lauterbach N, Zito L, Sizzano F, Toffalori C, Marcon J, et al. Effects of transmembrane region variability on cell surface expression and allorecognition of HLA-DP3. Hum Immunol. 2013;74(8):970-7.
38. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, et al. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol. 2010;71(10):1033-42.
39. Gabriel C, Furst D, Fae I, Wenda S, Zollikofer C, Mytilineos J, et al. HLA typing by next-generation sequencing - getting closer to reality. Tissue Antigens. 2014;83(2):65-75.
40. Erlich H. HLA DNA typing: past, present, and future. Tissue Antigens. 2012;80(1):1-11.
41. Major E, Rigo K, Hague T, Berces A, Juhos S. HLA typing from 1000 genomes whole genome and whole exome illumina data. PloS ONE. 2013;8(11).
42. Robinson J, Soormally AR, Hayhurst JD, Marsh SGE. The IPD-IMGT/HLA Database - New developments in reporting HLA variation. Hum Immunol. 2016;77(3):233-7.
43. Parham P, Ohta T. Population biology of antigen presentation by MHC class I molecules. Science. 1996;272(5258):67-74.
44. Juhos S, Rigó K, Horváth G. On genotyping polymorphic HLA genes-ambiguities and quality measures using NGS in Next Generation Sequencing-Advances, Applications and Challenges (ed. Kulski, J.K.) 370–386 (InTech, Rijeka, Croatia, 2016.
45. Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, et al. A: Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics. 2014;15:63.
46. Park H, Hyun J, Park S, Park M, Song E. False Homozygosity Results in HLA Genotyping due to Loss of Chromosome 6 in a Patient with Acute Lymphoblastic Leukemia. The Korean Journal of Laboratory Medicine. 2011;31:302-06.
47. Walsh P, Fildes N, Reynolds R. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus VWA. Nucleic Acids Research.1996;24:2807–2812.
48. Mayor N, Robinson J, McWhinnie A, Ranade S, Eng K, Midwinter W, et al. HLA Typing for the Next Generation. PLoS ONE. 2015;10:e0127153.

49. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 2009;5(5):15.
50. Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. J Hum Genet. [Review]. 2015.
51. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics. 2014;15(325):1471-2164.
52. Ka S, Lee S, Hong J, Cho Y, Sung J, Kim H-N, et al. HLAscan: genotyping of the HLA region using next-generation sequencing data. BMC Bioinformatics. [journal article]. 2017 May 12;18(1):258.
53. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. Human Mutation. 2017;38(7):788-97.
54. Kawaguchi S, Higasa K, Yamada R, Matsuda F. Comprehensive HLA Typing from a Current Allele Database Using Next-Generation Sequencing Data. Methods Mol Biol. 2018;1802:225-33.
55. Pepper MS. Launch of the Southern African Human Genome Programme. S Afr Med J. 2011;101(5):287-8.
56. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, et al. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nat Commun. 2017;8(1):017-00663.
57. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
58. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517(7534):327-32.
59. Martin PJ. The role of donor lymphoid cells in allogeneic marrow engraftment. Bone Marrow Transplant. 1990;6(5):283-9.
60. Cornelissen JJ, Lowenberg B. Developments in T-cell depletion of allogeneic stem cell grafts. Curr Opin Hematol. 2000;7(6):348-52.

61. Moretta L, Locatelli F, Pende D, Marcenaro E, Mingari MC, Moretta A. Killer Ig-like receptor-mediated control of natural killer cell alloreactivity in haploidentical hematopoietic stem cell transplantation. Blood. 2011;117(3):764-71.

62. Bishara A, De Santis D, Witt CC, Brautbar C, Christiansen FT, Or R, et al. The beneficial role of inhibitory KIR genes of HLA class I NK epitopes in haploidentically mismatched stem cell allografts may be masked by residual donor-alloreactive T cells causing GVHD. Tissue Antigens. 2004;63(3):204-11.

63. Hsu KC, Gooley T, Malkki M, Pinto-Agnello C, Dupont B, Bignon JD, et al. KIR ligands and prediction of relapse after unrelated donor hematopoietic cell transplantation for hematologic malignancy. Biol Blood Marrow Transplant. 2006;12(8):828-36.

64. Venstrom JM, Gooley TA, Spellman S, Pring J, Malkki M, Dupont B, et al. Donor activating KIR3DS1 is associated with decreased acute GVHD in unrelated allogeneic hematopoietic stem cell transplantation. Blood. 2010;115(15):3162-5.

65. Kelher MR, Masuno T, Moore EE, Damle S, Meng X, Song Y, et al. Plasma from stored packed red blood cells and MHC class I antibodies causes acute lung injury in a 2-event in vivo rat model. Blood. 2009;113(9):2079-87.

66. Davoren A, Smith OP, Barnes CA, Lawlor E, Evans RG, Lucas GF. Case report: four donors with granulocyte-specific or HLA class I antibodies implicated in a case of transfusion-related acute lung injury (TRALI). Immunohematology. 2001;17(4):117-21.

67. Sachs UJ, Wasel W, Bayat B, Bohle RM, Hattar K, Berghofer H, et al. Mechanism of transfusion-related acute lung injury induced by HLA class II antibodies. Blood. 2011;117(2):669-77.

68. Saito S, Ota M, Komatsu Y, Ota S, Aoki S, Koike K, et al. Serologic analysis of three cases of neonatal alloimmune thrombocytopenia associated with HLA antibodies. Transfusion. 2003;43(7):908-17.

69. WHO. Global Health Report. Geneva. 2013.

70. Disotell TR. Archaic human genomics. . Am J Phys Anthropol 2012;55:24-39.

71. Ramsay M. Africa: continent of genome contrasts with implications for biomedical research and health. FEBS Lett. 2012;586:2813-9.

72. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet. 2013;14:301-23.

73. Hendel H, Caillat-Zucman S, Lebuanec H, Carrington M, O'Brien S, Andrieu JM, et al. New class I and II HLA alleles strongly associated with opposite patterns of progression to AIDS. J Immunol. 1999;162(11):6942-6.
74. Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, et al. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. Science. 1999;283(5408):1748-52.
75. Carrington M, O'Brien S. The influence of HLA genotype on AIDS. Annual Review of Medicine. 2003;54:535-51.
76. Rohowsky-Kochan C, Skurnick J, Molinaro D, Louria D. HLA antigens associated with susceptibility/resistance to HIV-1 infection. Hum Immunol. 1998;59(12):802-15.
77. Tshabalala M, Morse GD, Zijenah LS. HLA Genetic Polymorphisms: Role in HIV-1 Susceptibility, Disease Progression and Treatment Outcomes. Retrovirology: Research and Treatment. 2013;5:1-8.
78. Garamszegi LZ. Global distribution of malaria-resistant MHC-HLA alleles: the number and frequencies of alleles and malaria risk. Malar J. 2014;13(349):1475-2875.
79. Sanchez-Mazas A, Cerny V, Di D, Buhler S, Podgorna E, Chevallier E, et al. The HLA-B landscape of Africa: Signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection. Mol Ecol. 2017;26(22):6238-52.
80. Yim JJ, Selvaraj P. Genetic susceptibility in tuberculosis. Respirology. 2010;15(2):241-56.
81. Brumme ZL, John M, Carlson JM, Brumme CJ, Chan D, Brockman MA, et al. HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. PloS ONE. 2009;4(8):0006687.
82. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PloS ONE. 2011;6(2):0014643.
83. Sanchez-Mazas A, Fernandez-Vina M, Middleton D, Hollenbach JA, Buhler S, Di D, et al. Immunogenetics as a tool in anthropological studies. Immunology. 2011;133(2):143-64.
84. Sanchez-Mazas A, Thorsby E. HLA in anthropology: the enigma of Easter Island. Clin Transpl. 2013:167-73.

85. HapMapProject. The International HapMap Project. Nature. 2003;426(6968):789-96.
86. Ayele FT, Hailu E, Finan C, Aseffa A, Davey G, Newport MJ, et al. Prediction of HLA Class II Alleles Using SNPs in an African Population. PloS ONE. 2012;7(6):e40206.
87. Hayhurst JD, du Toit ED, Borrill V, Schlaphoff TEA, Brosnan N, Marsh SGE. Two novel HLA alleles, HLA-A*30:02:01:03 and HLA-C*08:113, identified in a South African bone marrow donor. Tissue Antigens. 2015;85(4):291-3.

CHAPTER 2

Human leukocyte antigen (HLA) diversity: a southern African perspective

Mqondisi Tshabalala[1], Juanita Mellet[1] and Michael S. Pepper[1*]

[1]Department of Immunology and Institute for Cellular and Molecular Medicine, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

This chapter has been prepared in the format of a manuscript, and has been accepted and published in peer reviewed journal (Journal of Immunology Research). The publication can accessed under J Immunol Res. 2015; 2015:746151. doi: 10.1155/2015/746151. I designed the study, performed the experimental work, analysis and drafted the manuscript. Juanita Mellet contributed in analysis and writing the manuscript. Prof M.S Pepper conceived the study, obtained funding for the study and provided critical review of the manuscript.

2.1 Abstract

Despite the increasingly well-documented evidence of high genetic, ethnic and linguistic diversity amongst African populations, there is limited data on human leukocyte antigen (HLA) diversity in these populations. HLA is part of the host defense mechanism mediated through antigen presentation to effector cells of the immune system. With the high disease burden in southern Africa, HLA diversity data is increasingly important in the design of population specific vaccines and the improvement of transplantation therapeutic interventions. This review highlights the paucity of HLA diversity data amongst southern African populations and defines a need for information of this kind. This information will support disease association studies, provide guidance in vaccine design and improve transplantation outcomes.

2.1 Introduction

The human leukocyte antigen (HLA) complex on chromosome 6, also known as the major histocompatibility complex (MHC) in all mammals, consists of highly polymorphic genes whose protein products present antigens to T cells as part of an immune response to infections[1,2]. HLA molecules also impact on the development and effectiveness of vaccines, and play a determining role in the outcomes of transplantation[3-10].

The World Health Organization (WHO) indicates that there is a high burden of disease in southern Africa, especially communicable diseases such as HIV/AIDS, TB and malaria[11]. Despite the increasingly well-documented high genetic diversity observed amongst human populations in southern Africa[12], there is limited information on HLA diversity[8]. Understanding HLA diversity in these populations will provide insight into HLA disease associations, and may help in vaccine development. Transplantation as a therapeutic intervention requires strict HLA allele matching between donors and recipients to reduce rejection and the incidence of graft versus host disease (GVHD). Good clinical outcomes in transplant recipients are observed in cases of high resolution HLA matching[13,14], with the number of mismatches correlating with the risk of rejection and/or GVHD[15-17]. It is currently very difficult to match donor-recipient pairs in bone marrow registries in southern Africa, partly because of the great genetic diversity in this population. A recent study identified Black and Caucasian South African population-specific alleles[18], highlighting the need to investigate HLA diversity amongst southern Africans to improve global representation in the International ImMunoGeneTics® information system IMGT/HLA database[1,2]. HLA typing methods use the IMGT/HLA database as a reference; it is thus difficult to match individuals who have alleles which are not captured in the database.

HLA typing methods have evolved from low resolution serology typing to high resolution DNA sequencing based technologies (SBT). Despite high resolution, SBT has limitations of mostly typing certain exons within the HLA loci[19]. The antigen-binding groove encoded by exons 2 and 3 (class I) and exon 2 (class II) are routinely sequenced in most laboratories, thereby giving partial sequences of about 10% of

the reported alleles[19]. Another potential source of ambiguity in SBT HLA typing is the *cis/trans* assignment of DNA bases in a heterogeneous sample[20], yielding limited resolution data and thereby making it difficult to assign HLA allele types. It is possible to sequence the entire HLA region with current methods, but at a very high cost and a need for expert analysis. There have been advances in the use of next generation sequencing (NGS) in HLA typing to improve coverage of the HLA gene loci by high throughput, while at the same time reducing ambiguity associated with SBT typing[19,21,22]. To fully appreciate the NGS HLA typing tool, there is need for a complete HLA allele database[21] highlighting the need to quantify HLA diversity in the genetically diverse southern African populations[23].

African populations have been shown to be genetically diverse[24] , and are believed to be the cradle of humankind[25,26]. In general, genetic diversity of African populations is poorly understood[27] thereby limiting our understanding of human health and susceptibility to diseases, hence the need for further analysis/evaluation to map disease association and theraupetic gene targets. Despite the general similarities in culture and shared geographical location, genetic differences exist among populations at every 1000 base pairs[28,29]. In this review, we examine available HLA diversity data in southern Africa with a view to understanding disease burden, planning registry recruitment and donor-recipient matching, and to providing insights into the evolution of the ethnic and linguistic diversity in this region. This review specifically focuses on classical HLA diversity in southern African countries (characterized by genetically, culturally and linguistically diverse Bantu ethnicities and admixed populations[30-33]) herein defined as Zambia, Malawi, Zimbabwe, Mozambique, Angola, Namibia, Botswana, South Africa, Lesotho and Swaziland.

2.2 HLA diversity

There is an ever increasing number of HLA alleles, reflecting the rate of discovery of the diversity of the gene loci[1,2]. There are currently 13412HLA alleles described by the HLA nomenclature and included in the IMGT/HLA Database (based on IMTG/HLA 3.21.0 release, 06 July 2015), with HLA-B having the highest number of alleles (3977)[34]. HLA genetic variation does not vary in an individual's lifetime, but

high diversity is observed at the population level[1,2,35-40]. High HLA allelic diversity in humans is reflected by the high number of pseudogenes, and can be explained by natural selection and co-evolution with pathogens. There is an advantage of HLA diversity related to pathogen-derived peptide presentation to effector T cells: heterozygous individuals can potentially present more antigens than homozygotes for the different HLA alleles (heterozygosity advantage)[35,41]. In non-human species, low MHC diversity has been observed in several species (Tasmanian devils, cheetah, panda) and has been associated with disease susceptibility in some Tasmania devils[42], highlighting the advantage of HLA diversity in presenting many different antigens to effector cells of the immune system.

Prugnolle *et al* suggested that up to 39% of observed HLA class I diversity was due to geographical distance (and consequently human migration history) from the source of modern humans (assumed to be Ethiopia in this study), with the unaccounted source of diversity most likely being from pathogen driven selection[43]. Generally, populations exposed to a high pathogen burden show high HLA diversity, and there is a decreasing HLA diversity away from Africa (geographically measured by landmasses away from Africa)[43]. In related studies, microsatellite data has suggested that geographic distance from East Africa (probable source of modern humans) explains about 85% of a decreasing genetic diversity within human populations from the source (reviewed in[44]). Interestingly, HLA C is less expressed on cell surfaces; hence its diversity is least likely to be driven by viral pathogens (reviewed in[43]). It is historically accepted that TB was a major selective pressure in the evolution of Western European populations[45], with malaria acting on African populations[46]. These pathogens exerted a high selective pressure mostly on genes of the immune system (particularly those involved in protective immunity).

There is growing evidence for positive selection being responsible for maintaining HLA polymorphisms, most likely due to over dominant selection (heterozygote advantage) which maintains allelic lineages for much longer periods of time than neutral polymorphisms[40,47-49]. Globally, HLA diversity seems to be highest within populations than between populations (evidenced by major differences amongst continents)[1,2,37,50]. Several studies have highlighted alternative splicing of HLA class I genes giving rise to diverse isoforms[51] which might contribute to this diversity. For

example, alternative splicing to exclude exon 5 has been reported to give rise to several isoforms of HLA-A and -B[52]. Alternative splicing in other HLA class I exons has also been reported[53] including the non-classical HLA-G gene[54].

Other mechanisms of HLA diversity generation include point mutations (substitution, deletion, insertion): gene conversion (unidirectional gene transfer) and gene cross over (bidirectional gene transfer). Gene cross over, which is a form of recombination that can be intra/inter HLA loci during meiosis, enables exchange of genetic material linked to the generation of novel alleles in offspring as described by Carrington[55]. Other recombination events include gene conversion, a bidirectional donation of DNA between two homologous chromosomes. A recent study reports novel HLA alleles resulting from (a) non-synonymous amino acid change (HLA B*41:21, HLA DQB1*02:10, HLA QA1*01:12); (b) deletion leading to frame shift (HLA A*01:123N); (c) intralocus gene conversion (HLA B*35:231, HLA B*53:31); and (d) interlocus gene conversion (HLA C*07:294)[56]. It is important to note the low frequency of interlocus generated alleles as reported by several other studies as reviewed by Adamek *et al*[56].

2.3 HLA diversity in transplantation and transfusion

The human immune system uses HLA's uniqueness in every individual to recognize self from non-self; hence the body only mounts an immune response against foreign cells/molecules under normal conditions. Transplantation as a therapeutic intervention matches donor and recipient HLA molecules to decrease the likelihood of rejection[35]. The likelihood of two individuals having identical HLA molecules on all loci is very low, except for siblings, who have a 25% chance of being HLA-identical as a result of HLA molecules being codominantly expressed and inherited as haplotypes from both parents. The degree of HLA matching is a predictor of clinical outcome.

GVHD is an immunocompetent donor T cell mediated response against the recipient's immune system which is common in unmatched donor-recipient pairs. Acute GVHD can be reduced by donor T cell depletion, but this increases the risk of rejection, malignant disease relapse and impaired immune recovery[57,58]. In addition to HLA matching, killer-cell immunoglobulin-like receptors (KIRs) have been

documented to affect the clinical outcome of allogeneic transplantation[59-62]. In severe immunocompromised individuals, allogeneic transfusion with immune competent T cell-containing blood products might lead to transfusion associated GVHD. Transfusion related lung injury (TRALI) is an anti-HLA (mostly class I[63,64]) antibody related complication which may be fatal. Anti-HLA class II antibodies induce TRALI through monocyte and subsequent neutrophil activation[63,65]. Anti-HLA class I antibodies have been reported to be a cause of neonatal alloimmune thrombocytopenia together with platelet-derived specific antigens[66]. It is critical to know the population HLA diversity in order to improve donor-recipient matching in both transplantation and transfusion therapeutic interventions. Diversity data informs decision making in transplantation and transfusion aimed at reducing rejection while at the same time improving the outcome of the intended therapeutic intervention. Recruitment of donors from minority or under-represented populations might help to improve HLA diversity in registries[36] which improves the chances of donor-recipient matching.

2.4 HLA diversity in human disease associations

The high disease burden in southern Africa[11] offers a unique opportunity to study HLA disease association[8]. Several autoimmune conditions have been directly associated with specific class I and II HLA alleles, including rheumatoid arthritis, multiple sclerosis, ankylosing spondylitis, Grave's disease and many more, as reviewed by Trowsdale and Knight[67]. Several alleles have been associated with varying rates of HIV disease progression[4,41,68-70], susceptibility, transmission and treatment outcomes (reviewed in[70]). HLA has likewise been associated with malaria[6], TB susceptibility and protection[71] in various populations. In another example, although not directly related to southern Africa, the HLA-B locus has been linked to fatal and non-fatal Sudanese Ebola strains. Thus, HLA-B*67 and -B*15 have been associated with fatal outcomes and B*07 and B*14 have been associated with non-fatal Ebola infections[72].

Haplotype analysis gives information on disease/condition associated alleles, which are assumed to be inherited as blocks due to strong linkage disequilibrium[73]. HLA

alleles can be imputed from analyzing identity by descent (IBD) patterns within the HLA region of specific populations. This approach leverages on the observation that chromosomes with high IBD within MHC most likely share the same alleles. Haplotype analysis or SNP-based HLA allele imputation is important for disease association studies, but will not replace classical HLA typing for transplantation applications where a high degree of haplomatching is required for a good clinical outcome[74]. Currently several imputation methods are available to type HLA genes *in silico* and to fine-map associations within classical HLA genes[74]. Unfortunately, limited HLA diversity data from populations such as those in southern Africa make this difficult[74].

2.5 HLA diversity in population studies

There is documented evidence of geographical distribution of human genetic variation, which helps to understand human evolution, migration and adaptation to different environments and pathogens[75]. Several efforts aimed at understanding global human genetic diversity including the Hap Map Project[76], 1000 Genomes Project[77] and recently the African Genome Variation Project[33]; however, all of these have limited information on southern African populations. Some African genetic diversity studies have focused on targeted populations like hunter gatherers[78,79] or have had very limited sample size[80], and are therefore not representative of southern Africa. The low representation of southern African genetic data in global efforts makes it difficult to use the currently available reference panels for these populations, especially in disease association studies[33]. This suggests that targeted HLA sequencing of these diverse populations is necessary to improve their representation in reference panels.

There are marked differences in HLA diversity distribution globally, with geographically separated regions showing varying degrees of diversity[37,43,44,50]. Most HLA loci show high allele numbers across populations[37,81]. HLA DPA1 has the least number of alleles (40 as of July 2015)[82] compared to other classical HLA loci (for example HLA DQB1 which has 807 alleles). This is generally due to the fact that DPB1 loci are not routinely sequenced for transplantation purposes as are other HLA

genes. The global distribution of HLA diversity provides insight into human migration patterns, and could help understand past pathogen exposures[40]. As an example, HLA studies have been used to trace the spread of modern humans from East Africa, and model for co-evolution of genes and languages in Africa[83]. Interpretation of HLA in population studies can be improved by extensive knowledge of HLA diversity in these populations.

2.6 Contemporary studies on HLA diversity in southern Africa

To highlight the paucity of HLA diversity data in southern Africa, this review used a comprehensive literature search for previously published work on HLA diversity together with the Allele Frequency Net Database (AFND) to determine the information in the public domain. The key search terms for articles were "HLA AND genetic diversity AND southern Africa". Allele frequency data from AFND was extracted for sub-Saharan African countries, from which southern African data was compiled (Supplementary Table S2). Table 2.1 summarizes allele frequency data from the AFND web search (http://www.allelefrequencies.net/)[50] used in this review. The AFND is a public global database of alleles, genotypes and haplotype frequencies of HLA and KIRs from different studies, reports and proceedings of international workshops in immunogenetics and histocompatibility. HLA data is generated by different typing methods, but is curated in the database in accordance with the updated IMGT/HLA guidelines (this review used the 3.15.0 release - 17 January 2014)[1,2,37,50]. For this review, only positive allele frequencies from all ethnic groups within sub-Saharan Africa were extracted from the database (http://www.allelefrequencies.net/)[37,50]. The number of alleles reported in Mozambicans, Black South Africans, Caucasian South Africans, Tamil South Africans, Zulu South Africans, Tswana South Africans, Zambians and Shona Zimbabweans respectively was 18, 33, 25, 16, 37, 15, 20 and 32 alleles for HLA-A, and 25, 30, 41, 23, 45, 14, 29 and 46 alleles for HLA-B. HLA-C alleles were only reported for Black South Africans (28 alleles), Caucasian South Africans (29 alleles), Tamil South Africans (21 alleles), Zambians (12 alleles) and Shona Zimbabweans (24 alleles). All HLA class II alleles in the AFND were only reported for Shona Zimbabweans and South African Vendas as summarized in Supplementary Table

S2. Tables 2.1 and 2.2 summarize the selected allele frequencies from southern African populations and the total number of classical HLA alleles reported across different global regions as defined in the AFND[37,50], respectively.

South African had the highest number of HLA data sets from the AFND compared to other southern African countries (Table 2.1A). Some southern African countries (Angola, Lesotho, Malawi, Namibia and Swaziland) have no HLA data available (Table 2.1A). As summarized in Table 1(B and C), HLA-A*30 and its derivatives (A*30:01, A*30:02) are common in black populations (Mozambicans, Black South Africans, Zulus, Tswanas, Zambians and Zimbabwean Shonas). Caucasians and Tamils had a completely different HLA A allele frequency distribution compared to the other populations. HLA-A*02:01:01 was most frequent (0.26) in South African Caucasians, as has been reported by Solberg *et al* (HLA-A*02:01) in European (27%) and white American (20%) populations[84]. This suggests that South African Caucasians have a common ancestry with the Europeans and Americans, with the A*02:01 allele and its derivatives being restricted mostly to white populations. For the HLA-B locus, B*58 (B*58:02, B*58:01) was most common in Mozambicans, Black South Africans (including Zulus and Tswanas) as highlighted in Table 2.1(B and C). All HLA-B allele frequencies were less than 0.1 in Black South Africans and Shonas. All HLA-C frequencies were less than 0.2, with C*06:02 being commonly high in Black South Africans and Tamils. Although more than ten years old (2004), the study by Cao *et al* identified A*02:02, A*34:02, A*36:01, A*74:01, B*15:03, B*42:01, B*53:01, B*57:03 and B*58:02 as unique African alleles. Recently, diverse and novel HLA alleles have been reported in sub Saharan populations, for example HLA class II as reviewed in Ayele *et al*[85] and HLA class I as described by Paximadis *et al*[18] to further support high genetic diversity in Africans, and intra African diversity. Interestingly five new class I alleles ((A*30:01:02, A*30:02:02, A*68:27, B*42:06, and B*45:07) were reported in a recent South African study[18]. Additionally, Shepherd *et al* recently reported an overrepresentation of HLA-A*02:01, -A*34:02, and -B*58:02 in HIV negative controls in Zimbabwe[86] compared to the HIV positive group, which supports the earlier notion of African specific alleles.

The AFND reports very few HLA class II alleles amongst southern African populations; only Zimbabwean Shonas and Black South Africans[18] had HLA-DP

data. The reported allele frequencies (Table 2.1B and 2.1C) for the DP locus were: most frequent DPB1*01:01:01 (0.355) in Shona Zimbabweans and DPB1*13:01 (0.148) in Black South Africans; and least frequent DPB1*01:01:02, DPB1*02:02, DPB1*62:01, DPB1*65:01 and DPB1*80:01 (0.002) in Shona Zimbabweans. No alleles were reported for the DPA1 and DQA1 loci. The DQB1 locus was reported only in Botswana, Black South Africans, Shona Zimbabweans and Venda South Africans. DQB1*06 in Black South Africans was the most frequent (0.555) with DQB1*06:15 in Shona Zimbabweans being least frequent (0.002). DRB1 alleles were reported in all the studied populations except in some South Africans (Tswana, Tamil and Zulu). The most frequent allele was DRB1*11 (0.366) in Black South Africans, while the least frequent were DRB1*16 (0.002) in Mozambicans, and DRB1*03, DRB1*04:04, DRB1*12:04, DRB1*13 and DRB1*15:01 (all at 0.002) in Shona Zimbabweans.

The number of classical HLA alleles (Table 2.2) varies greatly in each geographical region, with North Africa having the highest number of AFND reported alleles globally, and sub-Saharan Africa (including southern Africa) in the top 5. In terms of HLA class II alleles, sub-Saharan Africa falls in the bottom 5 regions (with the least number of alleles - Table 2.2) for most of the HLA loci (DQA1, DQB1, DRB1). The DP locus generally has fewer numbers of reported alleles globally (http://www.allelefrequencies.net/)[37,50]. Interestingly, more than 50% of HLA class I alleles reported for sub-Saharan Africa are in southern Africa (Table 2.2), further highlighting diversity in this region. No HLA-DPA1 alleles were reported by the AFND in southern Africa, with less than 50% of the other class II alleles reported in sub-Saharan Africa coming from southern Africa.

The number of southern African HLA studies in the AFND is relatively low, reflecting the underrepresentation of this region. The data currently available is mostly low resolution with low sample numbers, and is not a true reflection of HLA diversity in the southern African context. This highlights the need for continual submission of southern African HLA diversity data to centralized databases like the AFND. The few studies from southern Africa also highlight the knowledge gap on HLA diversity in this region in this era of high resolution typing. Several HLA disease association studies with allele frequency data have been reported in the region[7,87-90]; these

frequencies might not be a true reflection of the general population owing to the confounding effect of the diseases. Allele frequency is highly dependent on sample size, and hence might not give a clear picture of HLA diversity.

Table 2.1 Contemporary studies which provide insight into HLA diversity in southern Africa

HLA allele frequency from the studies cited was extracted from the AFND[37,50] to assess HLA diversity in southern Africa. The AFND curated allele frequency data was generated from Mozambique, South Africa, Zambia and Zimbabwe as shown in (**A**) with the most and least frequent classical HLA alleles in these populations as shown in (**B** and **C**).

A. General description of studies used in this review

Country	Year	Population	n	Typing method	Loci typed	Comments
Bots	2005		55	SSP	DRB, DQB1	55 HIV negative compared to 74 HIV positive[7]
Moza	2010	Mostly Black	202	SSOP	A, B, DRB1	91.8% Black, rest admixture. Assane *et al*[37,50,108]
RSA	2012	Black	200	SBT,SSP	A, B, C, DRB1	Blacks from different ethno linguistic groups in RSA. Paximadis *et al*[18,37,50]
RSA	2012	Caucasians	102	SBT,SSP	A, B, C, DRB1	English and Afrikaner ancestry. Paximadis *et a* [18,37,50]
RSA	2002	Tamil/Natal	51	SSOP	A, B, C	Hammond[37,50,109]
RSA	2000	Black Zulu/Natal	100	SSOP	A, B	Could not distinguish A*0301 from A*0303N, and B*0705 from B*0706[37,39,50,110]
RSA	2006	Black/Tswana	41		A,B	Coetzee *et al*[37,50,111]
RSA	2004	Black	112	SSP	DRB1, DQB1, DPB1	112 Sclerosis controls compared to cases[90]
Zam	2002	Black/Lusaka	44	SSOP	A,B, C	Alleles similar at exons 2 and 3 could not be distinguished[37,50,107,112]
Zim	2002	Shona/Harare	230	SSOP	A,B,C,DPB1, DQA1,DQB1,DRB1	Louie[37,50,113]

B. Most frequent alleles in different southern African populations[37,50]

	Loci					
Population	**A**	**B**	**C**	**DP**	**DQ**	**DRB1**
Black RSA	A*30:01 (0.101)	B*42:01 (0.089), B*58:02 (0.094)	C*06:02 (0.149)	DPB1*13:01 (0.148)[90]	DQB1*06 (0.555)[90]	DRB1*11 (0.366)[90], DRB1*13:01 (0.124)
Bots					DQB1*16 (0.509)[7]	DRB1*11 (0.364)[7]
Caucasian RSA	A*01:01:01 (0.2), A*02:01:01 (0.26)	B*07:02:01 (0.149)	C*07:01 (0.172), C*07:02:01 (0.137)			DRB1*03:01 (0.122)
Moza	A*30 (0.239)	B*15 (0.156)				DRB1*11 (0.196), DRB1*13 (0.198)
Shona Zim	A*30:02 (0.147)	B*45:01 and B*53:01 (0.093)	C*04:01 (0.148)	DPB1*01:01:01 (0.355)	DQA1*01:02 (0.343), DQB1*05:01 (0.227), DQB1*06:02 (0.247)	DRB1*11:01 (0.144), DRB1*15:03 (0.153)
Tamil RSA	A*01:01 (0.17), A*11:01 (0.18)	B*40:06 (0.143)	C*06:02 (0.177)			
Tswana RSA	A*02 (0.146), A*30 (0.159)	B*58 (0.22)				
Venda RSA					DQB1*06 (0.437)	DRB1*11 (0.184)
Zam	A*30:02 (0.233)	B*42:01 (0.148)	C*17:01 (0.156)			
Zulu RSA	A*30 (0.195)	B*15 (0.15), B*58 (0.145)				

C. Least frequent alleles in different southern African populations[37,50]

	Loci					
Population	**A**	**B**	**C**	**DP**	**DQ**	**DRB1**
Bots					DQB1*02 (0.127)[7]	DRB1*10 and DRB1*12 (0.074)[7]
Caucasian RSA	A*02:05, A*02:17, A*11:12, A*24:07, A*25:01:01 A*33:03:01 and A*69:01 (0.005)	B*07:06, B*14:01, B*15:02, B*15:03, B*15:10, B*15:13, B*15:16, B*15:24, B*27:02, B*35:05, B*40:06:01 B*41:01, B*44:04, B*44:27 B*45:01, B*49:01, B*50:01 and B*58:02 (0.005)	C*02:05, C*03:16, C*04:08, C*04:09N, C*06:11, C*07:22 C*08:01, C*14:04 and C*17:01 (0.005)			DRB1*03:02, DRB1*04:08, DRB1*12:02, DRB1*14:04 and DRB1*15:07 (0.005)
Moza	A*32 (0.002)	B*27, B*37, B*73 and B*82 (0.002)				DRB1*16 (0.002)
Shona Zim	A*02:17, A*32:02, A*34:01 A*80:01, A*66:02, A*66:03 and A*74 (0.002)	B*07:12, B*13:04, B*14:04, B*15:17, B*15:18, B*35:02, B*39:10, B*40:01, , B*40:16, B*50:02 and B*73:01 (0.002)	C*03:04:01, C*07:08 C*12:04:02 and C*15:05 (0.02)	DPB1*01: 01:02, DPB1*02: 02, DPB1*62: 01, DPB1*65: 01 and DPB1*80: 01 (0.002)	DQA1*05:02 (0.004), DQB1*06:08 and DQB1*06:15 (0.002)	DRB1*03, DRB1*04:04, DRB1*12:04, DRB1*13 and DRB1*15:01 (0.002)
Tamil RSA	A*02:01, A*02:03 A*03:02 A*24:07,	B*15:25, B*27:05, B*44:07 B*50:01 and B*56:01 (0.01)	C*02:02:01, C*12:03, C*15:02			

	A*30:01 and A*32:01 (0.001)		and C*16:01 (0.01)			
Tswana RSA	A*01, A*31, A*32, A*36 and A*80 (0.012),	B*35, B*40, B*50 and B*53 (0.012)				
Venda RSA					DQB1*04 (0.094)	DRB1*10:01 (0.004)
Zam	A*02:06, A*02:14 A*26:01, A*33:01, A*34:02, A*43:01 and A*66:01 (0.012)	B*07:05, B*13:02, B*15:18, B*18:03 B*41:01, B*44:05, B*47:01 B*49:01 and B*57:01 (0.011)	C*03:03 and C*07:04 (0.022)			
Zulu RSA	A*31, A*31:01:02, A*33 and A*33:03 (0.005)	B*15:01, B*15:16, B*41:01, B*41:02, B*67, B*67:01 B*82 and B*82:01 (0.005)				

n=sample size, Bots=Botswana, Moza=Mozambique, RSA=Republic of South Africa, Zam=Zambia, Zim=Zimbabwe, SSP=sequence specific primers, SBT=sequence based typing, SSOP=sequence specific oligonucleotide primers, (number) is allele frequency in the population stated. Blanks indicate no alleles reported in the population or ethnicity not defined or typing method not specified

Table 2.2 Number of classical HLA alleles reported in each geographical region

Sub-Saharan Africa (including southern Africa) generally has a high number of class I alleles (ranked in the top 5 regions) with a low number of class II alleles (ranked in the bottom 5 regions). More than half of the reported class I alleles in sub-Saharan Africa come from southern Africa, with less than half of the reported class II alleles in the sub-Saharan region coming from southern Africa, data from AFND[37,50].

	HLA loci							
Region	**A**	**B**	**C**	**DPA1**	**DPB1**	**DQA1**	**DPB1**	**DRB1**
Australia	49	95	33	*	20	12	17	40
Europe	714	1121	387	16	137	47	89	602
N. Africa	982	1559	600	*	32	30	89	269
N. America	721	1166	390	7	74	29	93	574
N.E. Asia	262	477	131	12	78	47	57	318
S.Central America	121	288	59	12	78	28	60	549
S./S.E. Asia	407	731	227	10	99	21	64	280
SubSahara Africa	154	313	94	12	87	23	48	220
W. Asia	215	366	167	*	29	21	57	138
Ocenia	163	256	85	16	84	10	48	93
Southern Africa	131	291	54	*	21[a]	8[a]	20	58

N. Africa=North Africa, N. America=North America, N.E. Asia =North East Asia, S.Central America=South and Central America, S/S.E. Asia=South and South East Asia, W. Asia=West Asia * No loci specific alleles were reported in this region in the AFND. [a]Alleles only reported in Zimbabwean black Shona population in the AFND

2.7 Concluding remarks

There is limited data on HLA diversity in southern Africa, with most having been generated from disease association studies and which is therefore not a true reflection of the general population. It is often difficult to assign causality of a specific HLA allele to an infection/condition, because of linkage disequilibrium and other factors such as selection pressure, which are dependent on the condition/infection and the other arms of the immune system which are HLA independent[91]. As evidenced by the HIV example, several HLA B alleles have been associated with control of viremia[4,92,93] yet some individuals with these protective alleles develop AIDS (fail to control the virus)[94]. Recently Chen *et al* showed that HLA B*27 restricted CD8 T cells had variable viral replication inhibition capabilities in HIV controllers *versus* progressors due to a modulation by specific T cell receptor clonotypes[5]. There are few high resolution HLA datasets from southern African populations[1,2,37,50] despite growing advancement in NGS HLA typing.

HLA diversity data forms the cornerstone of population-specific vaccine development, and taking into consideration the high disease burden in southern Africa, information of this nature is particularly important in this region[11]. This review highlights the paucity of information on HLA genotypic data and documents the extent of HLA diversity data from the southern African perspective based on the limited data available. This underpins an urgent need for HLA data from the general populations in this region and for studies which elucidate the extent of this diversity. There is a need to build an HLA diversity resource for southern Africa (or Africa as a whole) such as for example the HLA-net (a European network)[95] which focuses on HLA diversity and its applications in histocompatibility, transplantation, epidemiology and population genetics. This network has developed analysis pipelines and guidelines for HLA diversity data for mostly European populations[95,96]. It is thus possible to build such a resource for the genetically diverse and disease burdened African continent to be used as a guideline for future studies including donor recruitment strategies[36], population studies[40,83,96] and disease association studies[6,8,71,72]. Furthermore, advancement in HLA typing methods such as NGS will help to finely investigate HLA diversity, as previous strategies have targeted a few

exons per locus thereby missing some of medically important variants outside the typed regions.

An understanding of HLA diversity will provide insight into allele frequency dependent selection fitness which varies between populations. This might help understand the high disease burden (especially with regard to HIV), and form the basis of vaccine development for the many infectious diseases as well as in the planning of vaccine clinical trials in the region. The paucity of HLA data from this region is a major hurdle in vaccine design[7]. Brumme *et al* highlight for example the need to elucidate HLA-restricted CTL responses in HIV vaccine design[97]. HLA class II antigens presented to $CD4^+$ T cells induce B cells leading to an antigen specific humoral immune response[98]. HLA class II alleles have been associated with humoral immune response inducing vaccines for malaria[99], active anticancer immunotherapy[100] and HIV[101]. The combined use of HLA class II T helper (Th) epitopes with CD8+ CTL epitopes theoretically generates a high efficacy vaccine as reviewed by Minzhen et al[100]. HLA diversity data might be useful in predicting the relative population coverage of a specific vaccine, add knowledge on epitope targets for vaccines[102], mechanisms of immune evasion[103,104], and evaluation of drug efficacy[105]. Posteraro *et al* reviewed the significance of HLA diversity in efficacy of vaccination, highlighting the need to further understand the link between genetic variation and immune responses[106].

It is generally easier to match donor-recipient pairs from populations with known HLA genotypes than in areas with information gaps[3], highlighting the need to understand population HLA diversity in order to improve on donor-recipient matching. It is generally difficult to find a donor HLA match for patients of African descent owing to the paucity of Africans in global registries together with the occurrence of African specific alleles and or haplotypes, and the high genetic diversity in these populations[107].

It is thus important to fully understand HLA diversity in the southern African context, to establish HLA-disease associations, to use this data for the informed design of population-specific vaccines against the many diseases, and to improve on donor-recipient matching.

2.8 Supplementary Data

Table S2: A Microsoft Excel spreadsheet listing all classical HLA alleles, their frequencies as reported by the AFND and a limited number of disease association studies in southern African populations has been made available online as supplementary Material (S2) http://dx.doi.org/10.1155/2015/746151. Additionally, as supplementary data to this thesis, supplementary data is available in Addendum 1.

2.9 References

1. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. Nucleic Acids Research. 2013 January 1, 2013;41(D1):D1222-D7.
2. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43(Database issue):20.
3. Beatty PG, Boucher KM, Mori M, Milford EL. Probability of finding HLA-mismatched related or unrelated marrow or cord blood donors. Human Immunology. 2000;61(8):834-40.
4. Carrington M, O'Brien S. The influence of HLA genotype on AIDS. Annual Review of Medicine. 2003;54:535-51.
5. Chen H, Ndhlovu ZM, Liu D, Porter LC, Fang JW, Darko S, et al. TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. Nat Immunol. [10.1038/ni.2342]. 2012;13(7):691-700.
6. Garamszegi LZ. Global distribution of malaria-resistant MHC-HLA alleles: the number and frequencies of alleles and malaria risk. Malar J. 2014;13(349):1475-2875.
7. Ndung'u T, Gaseitsiwe S, Sepako E, Doualla-Bell F, Peter T, Kim S, et al. Major histocompatibility complex class II (HLA-DRB and -DQB) allele frequencies in Botswana: association with human immunodeficiency virus type 1 infection. Clin Diagn Lab Immunol. 2005;12(9):1020-8.
8. Ramsay M. Africa: continent of genome contrasts with implications for biomedical research and health. FEBS Lett. 2012;586:2813-9.
9. Brander C, Frahm N, Walker BD. The challenges of host and viral diversity in HIV vaccine design (impedes of vaccine development owing to incomplete HLA information). Curr Opin Immunol. 2006;18:430-7.
10. Ovsyannikova I, Poland G. Vaccinomics: Current Findings, Challenges and Novel Approaches for Vaccine Development. AAPS J. 2011 2011/09/01;13(3):438-44.
11. WHO. Global Health Report. Geneva2013.
12. Disotell TR. Archaic human genomics. . Am J Phys Anthropol 2012;55:24-39.

13. Kunze-Schumacher H, Blasczyk R, Bade-Doeding C. Soluble HLA Technology as a Strategy to Evaluate the Impact of HLA Mismatches: J Immunol Res. 2014;2014:246171. Epub 2014 Sep 1.; 2014.

14. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood. 2007;110(13):4576-83.

15. Petersdorf EW. Optimal HLA matching in hematopoietic cell transplantation. Curr Opin Immunol. 2008;20(5):588-93.

16. Furst D, Muller C, Vucinic V, Bunjes D, Herr W, Gramatzki M, et al. High-resolution HLA matching in hematopoietic stem cell transplantation: a retrospective collaborative analysis. Blood. 2013;122(18):3220-9.

17. Pidala J, Wang T, Haagenson M, Spellman SR, Askar M, Battiwalla M, et al. Amino acid substitution at peptide-binding pockets of HLA class I molecules increases risk of severe acute GVHD and mortality. Blood. 2013;122(22):3651-8.

18. Paximadis M, Mathebula TY, Gentle NL, Vardas E, Colvin M, Gray CM, et al. Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population. Human Immunol. 2012;73:80-92.

19. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16(th) IHIW : review of HLA typing by NGS. Int J Immunogenet. 2013;40(1):72-6.

20. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, et al. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol. 2010;71(10):1033-42.

21. Gabriel C, Furst D, Fae I, Wenda S, Zollikofer C, Mytilineos J, et al. HLA typing by next-generation sequencing - getting closer to reality. Tissue Antigens. 2014;83(2):65-75.

22. Erlich H. HLA DNA typing: past, present, and future. Tissue Antigens. 2012;80(1):1-11.

23. Parham P, Ohta T. Population biology of antigen presentation by MHC class I molecules. Science. 1996;272(5258):67-74.

24. Disotell TR. Archaic human genomics. Am J Phys Anthropol. 2012;55:24-39.

25. Stewart JR, Stringer CB. Human evolution out of Africa: the role of refugia and climate change. Science. 2012;335(6074):1317-21.

26. Relethford JH. Genetic evidence and the modern human origins debate. Heredity. 2008;100(6):555-63.

27. Ramsay M. Africa: continent of genome contrasts with implications for biomedical research and health. FEBS Lett. 2012;586(18):2813-9.
28. www.broadinstitute.org.
29. Belle EM, Barbujani G. Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. Am J Phys Anthropol. 2007;133:1137-46.
30. Alessandrini M, Asfaha S, Dodgen TM, Warnich L, Pepper MS. Cytochrome P450 pharmacogenetics in African populations. Drug Metab Rev. 2013;45(2):253-75.
31. de Wit E, Delport W, Rugamika CE, Meintjes A, Moller M, van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum Genet. 2010;128(2):145-53.
32. http:en.wikipedia.org/wiki/Bantu_peoples.
33. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517(7534):327-32.
34. http://www.ebi.ac.uk/ipd/imgt/hla/stats.html.
35. Chapel H, Haeney M, Misbah S, Snowden N. Essentials of Clinical Immunology. 6 ed. John Wiley and Sons L, editor. West Sussex: Wiley Blackwell; 2014.
36. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Med J. 2007;48(1):11-23.
37. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Research. 2011;39(1):D913-D9.
38. Klein J, Sato A. The HLA system. First of two parts. N Engl J Med. 2000;343(10):702-9.
39. Middleton D, Williams F, Meenagh A, Daar AS, Gorodezky C, Hammond M, et al. Analysis of the distribution of HLA-A alleles in populations from five continents. Hum Immunol. 2000;61(10):1048-52.
40. Sanchez-Mazas A, Fernandez-Vina M, Middleton D, Hollenbach JA, Buhler S, Di D, et al. Immunogenetics as a tool in anthropological studies. Immunology. 2011;133(2):143-64.

41. Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, et al. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. Science. 1999;283(5408):1748-52.

42. Siddle HV, Kreiss A, Eldridge MDB, Noonan E, Clarke CJ, Pyecroft S, et al. Transmission of a fatal clonal tumor by biting occurs due to depleted MHC diversity in a threatened carnivorous marsupial. Proceedings of the National Academy of Sciences. 2007 October 9, 2007;104(41):16221-6.

43. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-Driven Selection and Worldwide HLA Class I Diversity. Current Biology. 2005;15(11):1022-7.

44. Handley LJL, Manica A, Goudet J, Balloux F. Going the distance: human population genetics in a clinal world. Trends in Genetics. 2007;23(9):432-9.

45. Cooke GS, Hill AV. Genetics of susceptibility to human infectious disease. Nat Rev Genet. 2001;2(12):967-77.

46. Miller LH. Impact of malaria on genetic polymorphism and genetic diseases in Africans and African Americans. Proc Natl Acad Sci U S A. 1994;91(7):2415-9.

47. Hughes AL, Nei M. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci U S A. 1989;86(3):958-62.

48. Yeager M, Hughes AL. Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. Immunol Rev. 1999;167:45-58.

49. Hughes AL, Ota T, Nei M. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol Biol Evol. 1990;7(6):515-24.

50. González-Galarza Faviel F, Takeshita Louise YC, Santos Eduardo JM, Kempson F, Maia Maria Helena T, Silva Andrea Luciana Soares d, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Research. 2015;43(D1):D784-D8.

51. Belicha-Villanueva A, Blickwedehl J, McEvoy S, Golding M, Gollnick S, Bangia N. What is the role of alternate splicing in antigen presentation by major histocompatibility complex class I molecules? Immunol Res. 2010 2010/03/01;46(1-3):32-44.

52. Krangel MS. Secretion of HLA-A and -B antigens via an alternative RNA splicing pathway. The Journal of Experimental Medicine. 1986 May 1, 1986;163(5):1173-90.
53. Norgaard L, Fugger L, Madsen HO, Svejgaard A. Identification of 4 different alternatively spliced HLA-A transcripts. Tissue Antigens. 1999;54(4):370-8.
54. Paul P, Adrian Cabestre F, Ibrahim EC, Lefebvre S, Khalil-Daher I, Vazeux G, et al. Identification of HLA-G7 as a new splice variant of the HLA-G mRNA and expression of soluble HLA-G5, -G6, and -G7 transcripts in human transfected cells. Human Immunology. 2000;61(11):1138-49.
55. Carrington M. Recombination within the human MHC. Immunol Rev. 1999;167:245-56.
56. Adamek M, Klages C, Bauer M, Kudlek E, Drechsler A, Leuser B, et al. Seven novel HLA alleles reflect different mechanisms involved in the evolution of HLA diversity: description of the new alleles and review of the literature. Hum Immunol. 2015;76(1):30-5.
57. Martin PJ. The role of donor lymphoid cells in allogeneic marrow engraftment. Bone Marrow Transplant. 1990;6(5):283-9.
58. Cornelissen JJ, Lowenberg B. Developments in T-cell depletion of allogeneic stem cell grafts. Curr Opin Hematol. 2000;7(6):348-52.
59. Moretta L, Locatelli F, Pende D, Marcenaro E, Mingari MC, Moretta A. Killer Ig-like receptor-mediated control of natural killer cell alloreactivity in haploidentical hematopoietic stem cell transplantation. Blood. 2011;117(3):764-71.
60. Bishara A, De Santis D, Witt CC, Brautbar C, Christiansen FT, Or R, et al. The beneficial role of inhibitory KIR genes of HLA class I NK epitopes in haploidentically mismatched stem cell allografts may be masked by residual donor-alloreactive T cells causing GVHD. Tissue Antigens. 2004;63(3):204-11.
61. Hsu KC, Gooley T, Malkki M, Pinto-Agnello C, Dupont B, Bignon JD, et al. KIR ligands and prediction of relapse after unrelated donor hematopoietic cell transplantation for hematologic malignancy. Biol Blood Marrow Transplant. 2006;12(8):828-36.
62. Venstrom JM, Gooley TA, Spellman S, Pring J, Malkki M, Dupont B, et al. Donor activating KIR3DS1 is associated with decreased acute GVHD in unrelated allogeneic hematopoietic stem cell transplantation. Blood. 2010;115(15):3162-5.

63. Kelher MR, Masuno T, Moore EE, Damle S, Meng X, Song Y, et al. Plasma from stored packed red blood cells and MHC class I antibodies causes acute lung injury in a 2-event in vivo rat model. Blood. 2009;113(9):2079-87.

64. Davoren A, Smith OP, Barnes CA, Lawlor E, Evans RG, Lucas GF. Case report: four donors with granulocyte-specific or HLA class I antibodies implicated in a case of transfusion-related acute lung injury (TRALI). Immunohematology. 2001;17(4):117-21.

65. Sachs UJ, Wasel W, Bayat B, Bohle RM, Hattar K, Berghofer H, et al. Mechanism of transfusion-related acute lung injury induced by HLA class II antibodies. Blood. 2011;117(2):669-77.

66. Saito S, Ota M, Komatsu Y, Ota S, Aoki S, Koike K, et al. Serologic analysis of three cases of neonatal alloimmune thrombocytopenia associated with HLA antibodies. Transfusion. 2003;43(7):908-17.

67. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet. 2013;14:301-23.

68. Hendel H, Caillat-Zucman S, Lebuanec H, Carrington M, O'Brien S, Andrieu JM, et al. New class I and II HLA alleles strongly associated with opposite patterns of progression to AIDS. J Immunol. 1999;162(11):6942-6.

69. Rohowsky-Kochan C, Skurnick J, Molinaro D, Louria D. HLA antigens associated with susceptibility/resistance to HIV-1 infection. Hum Immunol. 1998;59(12):802-15.

70. Tshabalala M, Morse GD, Zijenah LS. HLA Genetic Polymorphisms: Role in HIV-1 Susceptibility, Disease Progression and Treatment Outcomes. Retrovirology: Research and Treatment. 2013;5:1-8.

71. Yim JJ, Selvaraj P. Genetic susceptibility in tuberculosis. Respirology. 2010;15(2):241-56.

72. Sanchez A, Wagoner KE, Rollin PE. Sequence-based human leukocyte antigen-B typing of patients infected with Ebola virus in Uganda in 2000: identification of alleles associated with fatal and nonfatal disease outcomes. J Infect Dis. 2007;196(2):S329-36.

73. Kauppi L, Sajantila A, Jeffreys AJ. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. Human Molecular Genetics. 2003 January 1, 2003;12(1):33-40.

74. de Bakker PIW, Raychaudhuri S. Interrogating the major histocompatibility complex with high-throughput genomics. Human Molecular Genetics. 2012 October 15, 2012;21(R1):R29-R36.

75. Botigue LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, et al. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc Natl Acad Sci U S A. 2013;110(29):11791-6.

76. HapMapProject. The International HapMap Project. Nature. 2003;426(6968):789-96.

77. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

78. Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, et al. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. PLoS Genet. 2012;8(4):26.

79. Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012;338(6105):374-9.

80. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science. 2009;324(5930):1035-44.

81. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PloS ONE. 2011;6(2):0014643.

82. http://hla.alleles.org/nomenclature/stats.html.

83. Sanchez-Mazas A, Thorsby E. HLA in anthropology: the enigma of Easter Island. Clin Transpl. 2013:167-73.

84. Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum Immunol. 2008;69(7):443-64.

85. Ayele FT, Hailu E, Finan C, Aseffa A, Davey G, Newport MJ, et al. Prediction of HLA Class II Alleles Using SNPs in an African Population. PloS ONE. 2012;7(6):e40206.

86. Shepherd BL, Ferrand R, Munyati S, Folkard S, Boyd K, Bandason T, et al. HLA Correlates of Long-Term Survival in Vertically Infected HIV-1-Positive Adolescents in Harare, Zimbabwe. AIDS Res Hum Retroviruses. 2015;6:6.

87. Alkharsah KR, Dedicoat M, Blasczyk R, Newton R, Schulz TF. Influence of HLA alleles on shedding of Kaposi sarcoma-associated herpesvirus in saliva in an African population. J Infect Dis. 2007;195(6):809-16.

88. Carr DF, Chaponda M, Jorgensen AL, Castro EC, van Oosterhout JJ, Khoo SH, et al. Association of human leukocyte antigen alleles and nevirapine hypersensitivity in a Malawian HIV-infected population. Clin Infect Dis. 2013;56(9):1330-9.

89. Yang OO, Lewis MJ, Reed EF, Gjertson DW, Kalilani-Phiri L, Mkandawire J, et al. Human leukocyte antigen class I haplotypes of human immunodeficiency virus-1-infected persons on Likoma Island, Malawi. Hum Immunol. 2011;72(10):877-80.

90. Tikly M, Rands A, McHugh N, Wordsworth P, Welsh K. Human leukocyte antigen class II associations with systemic sclerosis in South Africans. Tissue Antigens. 2004;63(5):487-90.

91. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet. 2009;10(5):318-29.

92. Kaslow RA, Carrington M, Apple R, Park L, Munoz A, Saah AJ, et al. Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. Nat Med. 1996 Apr;2(4):405-11.

93. Migueles SA, Sabbaghian MS, Shupert WL, Bettinotti MP, Marincola FM, Martino L, et al. HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. Proc Natl Acad Sci U S A. 2000 Mar 14;97(6):2709-14.

94. Pereyra F, Addo MM, Kaufmann DE, Liu Y, Miura T, Rathod A, et al. Genetic and Immunologic Heterogeneity among Persons Who Control HIV Infection in the Absence of Therapy. Journal of Infectious Diseases. 2008 February 15, 2008;197(4):563-71.

95. Nunes JM, Buhler S, Roessli D, Sanchez-Mazas A. The HLA-net GENE[RATE] pipeline for effective HLA data analysis and its application to 145 population samples from Europe and neighbouring areas. Tissue Antigens. 2014;83(5):307-23.

96. Sanchez-Mazas A, Vidan-Jeras B, Nunes JM, Fischer G, Little AM, Bekmane U, et al. Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA-NET methodological recommendations. International Journal of Immunogenetics. 2012;39(6):459-76.

97. Brumme ZL, Chopera DR, Brockman MA. Modulation of HIV reservoirs by host HLA: bridging the gap between vaccine and cure. Curr Opin Virol. 2012;2(5):599-605.

98. Delves PJ, Roitt IM. The Immune system-second of two parts. New England Journal of Medicine. 2000;343(2):108-17.

99. Stephens HA, Browns AE, Chandanayingyong D, Webster HK, Sirikong M, Longta P, et al. The presence of the HLA class II allele DPB1*0501 in ethnic Thais correlates with an enhanced vaccine-induced antibody response to a malaria sporozoite antigen. Eur J Immunol. 1995;25(11):3142-7.

100. Xu M, Kallinteris NL, von Hofe E. CD4+ T-cell activation for immunotherapy of malignancies using Ii-Key/MHC class II epitope hybrid vaccines. Vaccine. 2012;30(18):2805-10.

101. Paris R, Bejrachandra S, Thongcharoen P, Nitayaphan S, Pitisuttithum P, Sambor A, et al. HLA class II restriction of HIV-1 clade-specific neutralizing antibody responses in ethnic Thai recipients of the RV144 prime-boost vaccine combination of ALVAC-HIV and AIDSVAX® B/E. Vaccine. 2012;30(5):832-6.

102. Zhao L, Zhang M, Cong H. Advances in the study of HLA-restricted epitope vaccines. Hum Vaccin Immunother. 2013;9(12):2566-77.

103. Yagita Y, Kuse N, Kuroki K, Gatanaga H, Carlson JM, Chikata T, et al. Distinct HIV-1 escape patterns selected by cytotoxic T cells with identical epitope specificity. J Virol. 2013;87(4):2253-63.

104. Carlson JM, Le AQ, Shahid A, Brumme ZL. HIV-1 adaptation to HLA: a window into virus-host immune interactions. Trends Microbiol. 2015;23(4):212-24.

105. Paul S, Kolla RV, Sidney J, Weiskopf D, Fleri W, Kim Y, et al. Evaluating the immunogenicity of protein drugs by applying in vitro MHC binding data and the immune epitope database and analysis resource. Clin Dev Immunol. 2013;467852(10):8.

106. Posteraro B, Pastorino R, Di Giannantonio P, Ianuale C, Amore R, Ricciardi W, et al. The link between genetic variation and variability in vaccine responses: systematic review and meta-analyses. Vaccine. 2014;32(15):1661-9.
107. Cao K, Moormann AM, Lyke KE, Masaberg C, Sumba OP, Doumbo OK, et al. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens. 2004;63(4):293-325.
108. Assane AA, Fabricio-Silva GM, Cardoso-Oliveira J, Mabunda NE, Sousa AM, Jani IV, et al. Human leukocyte antigen-A, -B, and -DRB1 allele and haplotype frequencies in the Mozambican population: a blood donor-based population study. Hum Immunol. 2010;71(10):1027-32.
109. Hammond MG, Anley D, editors. Tamil from Natal Province, South Africa. Proceedings of the 13th International Histocompatibility Workshop; 2006; Seattle: International Histocompatibility Working Group Press.
110. Hammond MG, Middleton D, Anley D, editors. Zulu from Natal Province, South Africa. Proceedings of the 13th International Histocompatibility Workshop and Conference; 2006; Seattle, WA: IHWG Press.
111. Coetzee V, Barrett L, Greeff JM, Henzi SP, Perrett DI, Wadee AA. Common HLA Alleles Associated with Health, but Not with Facial Attractiveness. PloS ONE. 2007;2(7):e640.
112. Cao K, Masaberg C, Yu J, Mann DL, Fernández-Viña MA, editors. Zambians from Lusaka, Zambia. Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference; 2006; Seattle, WA: IHWG Press, 2007.
113. Louie L, Mather K, Meyer D, Hollenbach J, Jackman R, Schultz K, et al., editors. Shona from Harare, Zimbabwe. Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference; 2006; Seattle, WA: IHWG Press, 2007.

CHAPTER 3

Human Leukocyte Antigen -A, -B, -C, -DRB1 and -DQB1 allele and haplotype frequencies in a subset of 237 donors in the South African Bone Marrow Registry

Mqondisi Tshabalala[1], Charlotte Ingram[2], Terry Schlaphoff[2,4], Veronica Borrill[2,4], Alan Christoffels[3] and Michael S. Pepper[1*]

[1]Institute for Cellular and Molecular Medicine, Department of Immunology, and SAMRC Extramural Unit for Stem Cell Research and Therapy, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

[2]South African Bone Marrow Registry (SABMR), P.O. Box 13353 Mowbray, Cape Town, 7705, South Africa

[3]SA MRC Bioinformatics Unit, South African National Bioinformatics Institute, University of Western Cape, Robert Sibukwe Road, Bellville, Cape Town 7535, South Africa

[4]Laboratory for Tissue Immunology, National Health Laboratory Service, C21 Groote Schuur Hospital, Anzio Street, Observatory, Cape Town, 7925 Cape Town, South Africa

This chapter has been prepared in the format of a manuscript, and has been accepted and published in peer reviewed journal (Journal of Immunology Research). The publication can accessed under J Immunol Res. 2018 Apr 23; 2018:2031571. doi: 10.1155/2018/2031571. I designed the study, performed experimental work, data analysis and drafted the manuscript. Charlotte Ingram, Terry Schlaphoff and Veronica Borrill recruited the study participants, provided HLA data contributed in writing the manuscript. Prof Alan Christoffels contributed in data analysis and writing the manuscript. Prof M.S Pepper conceived the study, obtained funding for the study and provided critical review of the manuscript.

3.1 Abstract

Human leukocyte antigen (HLA) -A, -B, -C, -DRB1 and -DQB1 allele and haplotype frequencies were studied in a subset of 237 volunteer bone marrow donors registered at the South African Bone Marrow Registry (SABMR). Hapl-o-Mat software was used to compute allele and haplotype frequencies from individuals typed at various resolutions, with some alleles in multiple allele code (MAC) format. Four hundred and thirty eight HLA -A, 235 HLA -B, 234 HLA -DRB1, 41 HLA -DQB1 and 29 HLA -C alleles are reported. The most frequent alleles were A*02:02g (0.096), B*07:02g (0.082), C*07:02g (0.180), DQB1*06:02 (0.157) and DRB1*15:01 (0.072). The most common haplotype was A*03:01g~B*07:02g~C*07:02g~DQB1*06:02~DRB1*15:01 (0.067), which has also been reported in other populations. Deviations from Hardy-Weinberg equilibrium were observed in A, B and DRB1 loci, with C~DQB1 being the only locus pair in linkage disequilibrium. This study describes allele and haplotype frequencies from a subset of donors registered at SABMR, the only active bone marrow donor registry in Africa. Although sample size was small, our results form a key resource for future population studies, disease association studies and donor recruitment strategies.

Keywords: HLA alleles, HLA haplotypes, South African Bone Marrow Registry

3.2 Introduction

The ~4Mb human leukocyte antigen (HLA) complex on chromosome 6 in humans is amongst the most polymorphic gene regions in the genome[1]. Seventeen thousand eight hundred and seventy-four (17 874) HLA alleles have been described in the IMTG/HLA database to date[2]. HLA gene products drive antigen presentation to T cells, and form the basis of host defense mechanisms against pathogens[3]. HLA also plays a role in vaccine development, and has a determining role in transplantation outcome[4-11]. In hematopoietic stem cell transplantation (HSCT), good clinical outcomes are associated with high resolution HLA matching[12,13], with the number of mismatches correlating with the risk of rejection and/or graft versus host disease (GVHD)[14-16].

Bone Marrow Donors Worldwide (BMDW) is a centralized databank of HLA phenotypes and other relevant data of unrelated stem cell donors which aims to support HSCT programmes[17]. The South African Bone Marrow Registry (SABMR), a nonprofit initiative based in Cape Town, was started in 1991 with the objective of providing HLA matched unrelated donors for South African patients and the world at large. The registry, listed in the BMDW, has more than 73 000 HLA typed volunteer donors from South Africa[18]. Unrelated donor registries globally, including the SABMR, increase chances of HLA matches for many patients in need of transplantation. Despite the high donor numbers globally, it is still difficult to find HLA matches for patients of black African origin, partly because of (a) the great genetic diversity in these populations[19] and (b) limited information on HLA diversity[9]. Most transplants facilitated by the SABMR are from foreign donors, mainly due to the limited number of donors in the registry, particularly those of black African and Asiatic/Indian origin[20]. There is thus a need to improve recruitment from these under represented populations into the SABMR, which, since 1997, has been the only registry on the African continent supporting an HLA matched unrelated donor stem cell transplantation programme[20,21].

Donor registries continuously try to improve their recruitment strategies through increasing donor numbers[22], recruiting young males[23], minority recruitment[24-26], recruiting donors with rare HLA phenotypes[27] or alternatively, using currently

available HLA allele and haplotype frequencies[25,28]. Although there is limited HLA diversity data for southern Africans (reviewed in[29]), Africans are considered to be genetically diverse[19] as has been determined using multiple markers[30-32], including HLA[33]. Most HLA families that exist globally are found in African populations[34], further confirming genetic diversity in these populations.

In this study, we describe HLA allele and haplotype frequency data from 237 donors registered with the SABMR, which serves as the source of unrelated marrow donors in South Africa. Frequencies of HLA- A, -B, -C, -DRB1 and -DQB1 alleles and haplotypes were analysed with the aim of developing a resource for disease association, anthropology and evolutionary studies. Furthermore, these data will support models for population specific vaccine development[35], and will improve donor recruitment strategies in South African populations

3.3 Methods

3.3.1 Study population, data access and ethics

Two hundred and thirty seven (237) SABMR registered consenting volunteer bone marrow donors HLA typed at varying resolutions were included in this study. This subset was accessed following an extensive re-consenting procedure of donors in the SAMBR. The self-reported ethnic grouping of the study population was Asian, Black, Chinese, Coloured, White and some unknown. High resolution typing has recently been adopted by SABMR, with most donors having low resolution typing (two digit)[20,21] which did not meet the current study criteria. For ethical compliance, the current study had to re consent donors to participate in the study. As a result only 237 of the potential 400 participants provided consent. Ethical clearance for this study was granted by the University of Pretoria, Faculty of Health Sciences Research Ethics Committee (220/2015) and the SABMR Board. Participants' data accessed included HLA -A, -B, -C, -DRB1 and -DQB1 loci molecular typing and self-reported ethnicity. Some typings in this data set were represented by multiple allele codes (MAC, formerly NMDP allele codes) as described in https://hml.nmdp.org/MacUI/.

3.3.2 HLA allele and haplotype frequency analysis

Allele and haplotype (two, three, four and five loci) frequencies were estimated by resolving phase and allelic ambiguities using the expectation-maximization (EM) algorithm[36,37] in Hapl-o-Mat open source software[38]. This software allows for allele verification using the IMTG/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/)[2,3] and recognizes ambiguities including MACs. Deviations from Hardy Weinberg equilibrium (HWE) were assessed at locus level using a chi-squared test[39]. Global linkage disequilibrium (LD) and HWE were implemented in Arlequin v3.5.2[40]. MAC coded alleles were dropped to two digit level resolution for HWE and LD analysis.

3.4 Results

3.4.1 Demographics and allele diversity

Self-reported ethnicity was not considered for analysis in this study owing to redundancy and simplicity of this classification as previously discussed[41,42]. One hundred and thirty-one (131) Black, 69 Caucasian, 19 Mixed-ancestry (Coloured), 15 Asian, 2 unknown and 1 Chinese individuals were included in this study. Nine hundred and seventy-seven (977) different possible alleles are reported in this study (Table S1). There were 438 HLA -A, 235 HLA -B, 29 HLA -C, 234 HLA -DRB1and 41 HLA -DQB1 alleles (Table S3.1), with the HLA-C locus having the lowest allelic diversity.

3.4.2 Hardy-Weinberg equilibrium and global LD analysis

In this donor subset, HLA-A, -B and -DRB1 genotypes deviated from the expected HWE proportions ($p<0.05$), with HLA-C and -DQB1 having insignificant ($p>0.05$) differences between expected and observed heterozygosity (Table 3.1). No significant global LD was detected between A~B, A~C, B~C, A~DRB1, B~DRB1, C~DRB1, A~DQB1, B~DQB1, DRB1~DQB1 locus pairs (Table 3.2). In addition, the C~DQB1 locus pair showed significant LD ($p<0.001$), as summarized in Table 3.2.

3.4.3 HLA allele frequency

The full list of alleles including those derived from MACs, and their frequencies, are listed in Table S3.1. The top 20 most frequent alleles across the five loci are summarized in Table 3.3 with the top three alleles per locus being A*02:01g (0.096), A*03:01g (0.093), A*01:01g (0.057); B*07:02g (0.082), B*08:01g (0.049), B*58:02 (0.048); C*07:02g (0.180), C*07:01g (0.104), C*04:01g (0.091); DRB1*15:01 (0.072), DRB1*15:03 (0.065), DRB1*07:01 (0.057) and DQB1*06:02 (0.157), DQB1*03:01 (0.139), DQB1*05:01 (0.118).

3.4.4 HLA haplotype frequency

All two, three, four and five (extended) haplotype frequencies are detailed in Supplementary Table 2 (Table S3.2), with the 20 most frequent haplotypes summarized in Tables 3.4 and 3.5 (extended haplotypes). The most common computed two, three and four loci haplotypes were B*07:02g~C*07:02g (0.145); C*07:02g~DRB1*15:01~DQB1*06:02 (0.107) and B*07:02g~C*07:02g~DRB1*15:01~DQB1*06:02 (0.108) respectively. We report a possible 7498 two locus, 6446 three locus and 773 four locus haplotypes in the SABMR subset of donors (Table S2). A*33:95~B*07:231N (1.08725E-06), A*03:01g~C*07:02g~DQB1*03:02 (1.03519E-06) and A*11:01g~C*01:02g~DRB1*01:01~DQB1*05:01 (2.8507E-06) were less frequent two, three and four locus haplotypes respectively (Table S2). The twenty most frequent extended haplotypes (five loci) are summarized in Table 3.5, with A*03:01g~B*07:02g~C*07:02g~DRB1*15:01~DQB1*06:02 being the most frequent (0.067).

Table 3.1 Hardy-Weinberg Equilibrium (HWE) parameters for the 237 donors studied

Locus	Obs Het	Exp Het	SD	Steps done	P value
HLA -A	1.00000	0.96196	0.00000	1001000	<0.001*
HLA -B	0.99554	0.97382	0.00001	1001000	0.00074*
HLA -C	1.00000	0.93582	0.00020	1001000	0.07316
HLA -DRB1	0.98958	0.95618	0.00000	1001000	<0.001*
HLA -DQB1	1.00000	0.91336	0.00027	1001000	0.15049

SD standard deviation; * statistically significant (*p<0.005*)

Table 3.2 Pair-wise global LD estimates across the five loci

haplotype	Chi-square test value	Degrees of freedom	P value
A~B	1672.062	3696	1.000
A~C	845.290	1488	1.000
B~C	1220.641	2387	1.000
A~DRB1	1288.195	2256	1.000
B~DRB1	1713.476	3619	1.000
C~DRB1	847.773	1457	1.000
A~DQB1	596.485	816	1.000
B~DQB1	777.193	1309	1.000
C~DQB1	732.281	527	<0.001*
DRB1~DQB1	802.780	799	0.456

* Statistically significant (*p<0.005*)

Table 3.3 The twenty most frequent HLA -A, -B, -C, -DRB1 and -DQB1 alleles from the 237 donor subset (Full list in Table S1)

A	frequency	B	frequency	C	frequency	DRB1	frequency	DQB1	frequency
A*02:01g	0.096	B*07:02g	0.082	C*07:02g	0.180	DRB1*15:01	0.072	DQB1*06:02	0.157
A*03:01g	0.093	B*08:01g	0.049	C*07:01g	0.104	DRB1*15:03	0.065	DQB1*03:01	0.139
A*01:01g	0.057	B*58:02	0.048	C*04:01g	0.091	DRB1*07:01	0.057	DQB1*05:01	0.118
A*24:02g	0.051	B*42:01	0.039	C*06:02g	0.074	DRB1*13:01	0.053	DQB1*02:01	0.090
A*30:02g	0.050	B*44:03	0.033	C*08:02g	0.057	DRB1*11:01	0.053	DQB1*03:02	0.083
A*68:02g	0.048	B*15:10	0.032	C*02:02g	0.051	DRB1*03:01	0.046	DQB1*06:03	0.068
A*11:01g	0.044	B*15:01g	0.031	C*15:02g	0.045	DRB1*04:01	0.038	DQB1*04:02	0.066
A*30:01g	0.043	B*15:03g	0.031	C*05:01g	0.045	DRB1*03:02	0.034	DQB1*02:02	0.063
A*29:02g	0.035	B*35:01g	0.031	C*03:04g	0.045	DRB1*13:02	0.033	DQB1*05:03	0.049
A*23:01g	0.034	B*14:02	0.028	C*12:03g	0.040	DRB1*01:02	0.029	DQB1*03:03	0.045
A*68:01g	0.025	B*58:01g	0.028	C*03:03g	0.034	DRB1*01:01	0.029	DQB1*06:01	0.042
A*43:01	0.024	B*18:01g	0.026	C*01:02g	0.034	DRB1*15:02	0.026	DQB1*03:19	0.021
A*66:01g	0.023	B*51:01g	0.025	C*17:01g	0.028	DRB1*11:02	0.021	DQB1*06:04	0.021
A*33:03g	0.023	B*15:16	0.021	C*12:02g	0.028	DRB1*13:03	0.020	DQB1*06:09	0.021
A*34:02	0.022	B*13:02g	0.021	C*16:01g	0.023	DRB1*11:04	0.018	DQB1*04:04	0.003
A*74:01g	0.020	B*58:60	0.019	C*14:02g	0.023	DRB1*12:01	0.016	DQB1*03:30	0.003
A*31:01g	0.020	B*53:01g	0.018	C*18:01g	0.017	DRB1*12:02	0.015	DQB1*06:40	0.003
A*24:07	0.017	B*45:01g	0.018	C*08:04	0.017	DRB1*08:04	0.014	DQB1*06:11	0.003
A*02:05g	0.016	B*81:01g	0.018	C*07:04g	0.017	DRB1*14:04	0.013	DQB1*06:218	0.000
A*33:01g	0.015	B*27:05g	0.017	C*03:02g	0.017	DRB1*03:102	0.013	DQB1*06:185	0.000

"g" groups are expressed and null alleles with identical amino acid sequences across class I exons 2 and 3 and class II exon 2[43]

Table 3.4 The twenty most frequent two, three and four locus haplotype frequencies in the 237 donor subset (Full list in Table S2)

Two loci	freq	Three loci	freq	Four loci	freq
B*07:02g~C*07:02g	0.145	C*07:02g~DRB1*15:01~DQB1*06:02	0.107	B*07:02g~C*07:02~DRB1*15:01g~DQB1*06:02	0.108
DRB1*15:01~DQB1*06:02	0.125	B*07:02g~DRB1*15:01~DQB1*06:02	0.106	B*08:01g~C*07:01g~DRB1*03:01~DQB1*02:01	0.067
C*07:02g~DQB1*06:02	0.105	B*07:02g~C*07:02g~DQB1*06:02	0.101	A*03:01g~B*07:02g~C*07:02g~DQB1*06:02	0.063
B*07:02g~DQB1*06:02	0.099	B*07:02g~C*07:02g~DRB1*15:01	0.084	A*03:01g~B*07:02g~DRB1*15:01~DQB1*06:02	0.061
DRB1*03:01~DQB1*02:01	0.096	A*03:01g~B*07:02g~C*07:02g	0.081	A*03:01g~C*07:02g~DRB1*15:01~DQB1*06:02	0.057
C*07:02g~DRB1*15:01	0.091	B*08:01g~DRB1*03:01~DQB1*02:01	0.076	A*03:01g~B*07:02g~C*07:02g~DRB1*15:01	0.051
A*03:01g~C*07:02g	0.079	C*07:01g~DRB1*03:01~DQB1*02:01	0.066	A*01:01g~C*07:01g~DRB1*03:01~DQB1*02:01	0.049
B*08:01g~DQB1*02:01	0.071	B*08:01g~C*07:01g~DQB1*02:01	0.063	A*01:01g~B*08:01g~C*07:01g~DQB1*02:01	0.047
DRB1*13:01~DQB1*06:03	0.071	A*03:01g~C*07:02g~DQB1*06:02	0.062	A*01:01g~B*08:01g~DRB1*03:01~DQB1*02:01	0.045
A*03:01g~DQB1*06:02	0.061	A*03:01g~B*07:02g~DQB1*06:02	0.057	A*01:01g~B*08:01g~C*07:01g~DRB1*03:01	0.037
B*08:01g~C*07:01g	0.058	B*08:01g~C*07:01g~DRB1*03:01	0.052	B*15:01g~C*03:03g~DRB1*13:01~DQB1*06:03	0.029
C*07:01g~DQB1*02:01	0.053	A*03:01g~DRB1*15:01~DQB1*06:02	0.051	B*44:02g~C*05:01g~DRB1*01:01~DQB1*05:01	0.025
DRB1*01:01~DQB1*05:01	0.051	A*01:01g~B*08:01g~C*07:01g	0.047	A*11:01g~C*01:02g~DRB1*15:01~DQB1*06:02	0.025
DRB1*11:01~DQB1*03:01	0.051	A*01:01g~C*07:01g~DQB1*02:01	0.046	A*03:01g~C*07:02g~DRB1*01:01~DQ	0.025

				B1*05:01	
C*07:01g~DRB1*03:01	0.049	A*03:01g~C*07:02g~DRB1*15:01	0.044	A*03:01g~B*07:02g~C*07:02g~DQB1*03:01	0.023
C*04:01g~DQB1*05:01	0.045	A*01:01g~B*08:01g~DQB1*02:01	0.043	A*11:01g~B*51:01g~DRB1*15:01~DQB1*06:02	0.023
DRB1*07:01~DQB1*02:02	0.044	A*01:01g~DRB1*03:01~DQB1*02:01	0.042	A*02:01g~B*07:02g~DRB1*15:01~DQB1*06:02	0.023
B*07:02g~DRB1*15:01	0.043	A*01:01g~C*07:01g~DRB1*03:01	0.037	B*42:01~C*17:01g~DRB1*03:02~DQB1*04:02	0.021
A*01:01g~DQB1*02:01	0.042	A*11:01g~DRB1*13:01~DQB1*06:03	0.037	B*57:01g~C*06:02g~DRB1*07:01~DQB1*03:03	0.021
B*14:02~C*08:02g	0.041	A*24:02g~B*07:02g~C*07:02g	0.035	A*01:01g~C*06:02g~DRB1*07:01~DQB1*03:03	0.020

freq" frequency; "g" groups are expressed and null alleles with identical amino acid sequences across class I exons 2 and 3 and class II exon 2[43]

Table 3.5 The twenty most frequent extended (five loci) haplotype frequencies from the 237 donor subset in the SABMR (full list in Table S2)

A~B~C~DQB1~DRB1 haplotype	frequency
A*03:01g~B*07:02g~C*07:02g~DRB1*15:01~DQB1*06:02	0.067
A*01:01g~B*08:01g~C*07:01g~DRB1*03:01~DQB1*02:01	0.050
A*01:01g~B*57:01g~C*06:02g~DRB1*07:01~DQB1*03:03	0.021
A*03:01g~B*07:02g~C*07:02g~DRB1*01:01~DQB1*05:01	0.017
A*11:01g~B*15:01g~C*03:03g~DRB1*13:01~DQB1*06:03	0.017
A*24:02g~B*07:02g~C*07:02g~DRB1*15:01~DQB1*06:02	0.017
A*02:11g~B*40:06~C*15:02g~DRB1*15:01~DQB1*06:01	0.017
A*33:01g~B*14:02~C*08:02g~DRB1*13:01~DQB1*06:03	0.017
A*68:02g~B*14:01~C*08:02g~DRB1*07:01~DQB1*02:02	0.017
A*11:01g~B*51:01g~C*01:02g~DRB1*04:01~DQB1*03:02	0.017
A*31:01g~B*27:05g~C*02:02g~DRB1*15:01~DQB1*06:02	0.017
A*03:01g~B*07:02g~C*07:02g~DRB1*11:01~DQB1*03:01	0.017
A*68:02g~B*14:02~C*08:02g~DRB1*13:03~DQB1*03:01	0.017
A*69:01~B*15:17~C*07:01g~DRB1*11:01~DQB1*03:01	0.017
A*02:01g~B*07:02g~C*07:02g~DRB1*15:01~DQB1*06:02	0.017
A*30:01g~B*42:01~C*17:01g~DRB1*03:02~DQB1*04:02	0.013
A*24:02g~B*15:32~C*12:03g~DRB1*12:02~DQB1*03:01	0.008
A*23:01g~B*49:01g~C*07:01g~DRB1*15:02~DQB1*05:03	0.008
A*25:01g~B*08:01g~C*07:01g~DRB1*03:01~DQB1*02:01	0.008
A*26:01g~B*58:01g~C*05:01g~DRB1*15:03~DQB1*06:02	0.008

"g" groups are expressed and null alleles with identical amino acid sequences across class I exons 2 and 3 and class II exon 2[43]

3.5 Discussion

Although this study had a limited sample size of 237, we provide an in-depth analysis of HLA diversity in a subset of donors in the SABMR. Mixed resolution HLA typing data with multiple allele codes (https://hml.nmdp.org/MacUI) were analyzed using a robust Hapl-o-Mat[38] package to compute allele and haplotype frequencies through the EM algorithm. In addition, the package supports typing ambiguities in NMDP codes (MAC), G group and GL string formats. Since Hapl-o-Mat does not compute LD and HWE, we reduced all MAC encoded typing in our data set to two digit resolution to estimate these parameters in Arlequin v3.5.2[40]. Although there was the possibility of underestimation due to loss of some allele information, global LD and HWE deviation is important in genetic studies.

Strong LD of C~DQB1 locus pairs ($p<0.001$ in Table 3.2) in our study suggests limited chances of recombination between alleles from these loci in our population, hence a greater chance of being inherited together. LD patterns of HLA or other genes may be used to infer evolutionary relatedness of populations[44]. Generally, individuals with haplotypes in LD are more likely to find haplomatches and strong LD is indicative of evolutionary relatedness of those alleles/loci. Carvallo and colleagues[45] report HLA -A, -B and -DRB1 in HWE ($p>0.05$), which contrasts to the significant deviation ($p<0.05$) observed in the current study (Table 3.1). Sample size and mixed typing resolution in the current study may have affected HWE proportions. When there is no deviation from HWE, HLA data may be used to infer human peopling history in anthropological studies[46]. Furthermore, there is evidence of large HWE deviations influencing EM algorithm based allele and haplotype frequency estimations[47]. It is thus important to note the sample size and mixed typing resolution limitations of the current study in interpreting HWE and LD analysis.

Taking into account the nature of the HLA data in the current study, we report 977 possible alleles (Table S3.1). HLA -C had the lowest number (29) of alleles compared to HLA -A (438 alleles) which had the highest. There are generally more reported HLA-B alleles in the HLA database[2,3]. We note though that previously, most registries routinely typed HLA -A, -B, -DRB1 for new donors with few being typed for HLA -C and HLA -DQB1[48]. This might explain the observed allele numbers in our

study. There is an ever increasing number of alleles in the database (currently 17 874 in the IMTG/HLA database release 3.31)[2,3], with South Africa contributing some unique alleles[49,50].

HLA -A*02:01g with a frequency of 9.6% in the current study has been reported in North West England Caucasians at a higher frequency of 28.9%[51]. This English study also reported B*07:02g, C*07:02g and DRB1*15:01 at frequencies of 15.3%, 15.6% and 15.9% respectively[51] compared to 8.2%, 18.0% and 7.2% in the current study. It is important to note that the fifth most common allele in our study, namely A*30:02g (5% frequency in Table 3.3 and Table S3.1), is identical (exon 2 and 3 amino acid sequence) to a novel A*30:02:01:03 allele previously reported in a SABMR donor[49]. HLA -DQB1*06:02 (15.7%) has been observed at higher frequencies in previous studies in West Africans (30.8%), Shona Zimbabweans (24.7%) and is lower in Kenyans (14.6%), Colombians (15.0%) and people from Papua New Guinea (15.0%)[26]. HLA -DRB1*15:01 (7.2%) in the current study (Table 3.3) has been reported previously in South African populations at varying frequencies: 11.2% in Caucasians and 2.4% in Black Africans[26]. Additionally, DRB1*15:01 had a 3.8% frequency in Inuit women[52], 11.65% in Chinese[53] and more than 50% in North Africans, Asians, people from Oceania and Europeans[54].

The main thrust of our study has been the ability to estimate with high confidence, haplotype frequencies from mixed resolution typings including MAC (https://hml.nmdp.org/MacUI) encoded alleles[38]. No record of the most frequent two, three and four loci haplotypes reported in this study (Table 3.4 and Table S3.2) is found in the allele frequency database[2,3,55]. The most frequent (6.7%) extended haplotype A*03:01g~B*07:02g~C*07:02g~DRB1*15:01~DQB1*06:02 has previously been reported amongst Chinese populations at varying frequencies (0.93-5.20 %)[53] compared to our 6.7 %. There is no record of this haplotype in African populations in the AFND allele frequency database[56]. A lower frequency (3.31%) of this haplotype has also been reported in a German registry as described by Sauter and colleagues[57].

Haplotype frequencies from a specific population may be useful for resolving typing ambiguities using statistical approaches in typing prospective individuals from the

same population[58]. It is important though to note that sample size affects these computations, with a tendency towards haplotype overestimation in small sample sized studies[35]. Other confounders include typing ambiguity as previously described[59]. Additionally, multi-locus haplotype frequency estimation better informs disease association studies than allele frequency[47]. A complete list of donor registry HLA haplotype frequencies better informs donor-patient matching tools like Easymatch®[60], NMDP HapLogic[61,62] and Optimatch[63] especially for patients of African origin who might benefit from donors in the SABMR. These tools use haplotype frequencies to compute the likelihood of a donor-patient match, and also anticipate the most likely mismatches. Haplotype frequency may be used to estimate the probability of finding a recipient match, or may give an indication of the likelihood of mismatches from initial registry searches[35]. Additionally, haplotypes are better indicators of HLA match estimation compared to allele frequency alone[35]. Variations in allele frequency distribution in populations in general provide insight into peopling history[64,65]. HLA genetic makeup of populations provides insight into history including selective pressures by pathogens[33], migration, admixture and changes in population size[54,66-68].

Allele and haplotype frequencies from this study highlight the need for continued analysis by the SABMR for a better understanding of HLA diversity in the region. There is limited HLA diversity data for South African populations (reviewed in[29]), despite the evident value in transplantation, donor recruitment, disease association and population studies. In addition, some registries specifically aim to improve recruitment from ethnic minorities[25] to increase the HLA diversity, and hence the probability of finding an appropriate donor for a given patient. In this context, knowledge of the distribution of alleles and haplotypes in many different population groups, as determined by high-resolution typing, may allow for modification of recruitment strategies.

3.6 Conclusions

Although results reported here are from a small subset of SABMR registered donors, allele and haplotype frequencies generated by Hapl-o-Mat tool[38] could be a useful resource for future anthropological and population genetics studies in South Africans. Furthermore, these findings may better inform donor recruitment strategies for the SABMR. The small sample size limitation of this study also highlights the need for larger studies in order to better understand HLA diversity in South African populations. It would also be interesting to analyze the whole donor registry and compare its HLA diversity data to other registries globally.

3.7 Supplementary Information

Supplementary Table 3.1 (Table S3.1): HLA -A, -B, -C, -DRB1 and -DQB1 allele frequencies in 237 volunteer bone marrow donors registered in the South African Bone Marrow Registry. The 237 individuals described herein are a subset of all SABMR registered donors. Accessible through J Immunol Res. 2018 Apr 23; 2018:2031571. doi: 10.1155/2018/2031571, additionally available as supplementary data to this thesis in Addendum 1..

Supplementary Table 3.2 (Table S3.2): Two, three, four and five loci Haplo-o-Mat 38 estimated haplotype frequencies in 237 volunteer bone marrow donors registered in the South African Bone Marrow Registry. The 237 individuals described herein are a subset of all SABMR registered donors. Accessible through J Immunol Res. 2018 Apr 23; 2018:2031571. doi: 10.1155/2018/2031571, additionally available as supplementary data to this thesis in Addendum 1.

3.8 References

1. Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, et al. Deep whole-genome sequencing of 100 southeast Asian Malays. Am J Hum Genet. 2013;92(1):52-66.
2. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43(Database issue):20.
3. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. Nucleic Acids Research. 2013 January 1, 2013;41(D1):D1222-D7.
4. Beatty PG, Boucher KM, Mori M, Milford EL. Probability of finding HLA-mismatched related or unrelated marrow or cord blood donors. Human Immunology. 2000;61(8):834-40.
5. Carrington M, O'Brien S. The influence of HLA genotype on AIDS. Annual Review of Medicine. 2003;54:535-51.
6. Chen H, Ndhlovu ZM, Liu D, Porter LC, Fang JW, Darko S, et al. TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. Nat Immunol. 2012;13(7):691-700.
7. Garamszegi LZ. Global distribution of malaria-resistant MHC-HLA alleles: the number and frequencies of alleles and malaria risk. Malar J. 2014;13(349):1475-2875.
8. Ndung'u T, Gaseitsiwe S, Sepako E, Doualla-Bell F, Peter T, Kim S, et al. Major histocompatibility complex class II (HLA-DRB and -DQB) allele frequencies in Botswana: association with human immunodeficiency virus type 1 infection. Clin Diagn Lab Immunol. 2005;12(9):1020-8.
9. Ramsay M. Africa: continent of genome contrasts with implications for biomedical research and health. FEBS Lett. 2012;586:2813-9.
10. Brander C, Frahm N, Walker BD. The challenges of host and viral diversity in HIV vaccine design (impedes of vaccine development owing to incomplete HLA information). Curr Opin Immunol. 2006;18:430-7.
11. Ovsyannikova I, Poland G. Vaccinomics: Current Findings, Challenges and Novel Approaches for Vaccine Development. AAPS J. 2011 2011/09/01;13(3):438-44.

12. Kunze-Schumacher H, Blasczyk R, Bade-Doeding C. Soluble HLA Technology as a Strategy to Evaluate the Impact of HLA Mismatches: J Immunol Res. 2014;2014:246171. Epub 2014 Sep 1.; 2014.

13. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood. 2007;110(13):4576-83.

14. Petersdorf EW. Optimal HLA matching in hematopoietic cell transplantation. Curr Opin Immunol. 2008;20(5):588-93.

15. Furst D, Muller C, Vucinic V, Bunjes D, Herr W, Gramatzki M, et al. High-resolution HLA matching in hematopoietic stem cell transplantation: a retrospective collaborative analysis. Blood. 2013;122(18):3220-9.

16. Pidala J, Wang T, Haagenson M, Spellman SR, Askar M, Battiwalla M, et al. Amino acid substitution at peptide-binding pockets of HLA class I molecules increases risk of severe acute GVHD and mortality. Blood. 2013;122(22):3651-8.

17. Worldwide BMD. https://www.bmdw.org/.

18. Worldwide BMD. https://www.bmdw.org/numberofdonors/.

19. Disotell TR. Archaic human genomics. . Am J Phys Anthropol 2012;55:24-39.

20. du Toit E, Schlaphoff T, Borrill V. The South African Bone Marrow Registry–role in providing unrelated donors for allogeneic stem cell transplantation. Continuing Medical Education. 2012;30(8):293-94.

21. du Toit ED, Borrill V, Schlaphoff TEA. The South African bone marrow registry (SABMR) in 2004. Transfusion and apheresis science : official journal of the World Apheresis Association : official journal of the European Society for Haemapheresis. 2005;32(1):25-6.

22. Kollman C, Abella E, Baitty RL, Beatty PG, Chakraborty R, Christiansen CL, et al. Assessment of optimal size and composition of the U.S. National Registry of hematopoietic stem cell donors. Transplantation. 2004;78(1):89-95.

23. Schmidt AH, Biesinger L, Baier D, Harf P, Rutt C. Aging of registered stem cell donors: implications for donor recruitment. Bone Marrow Transplant. 2008;41(7):605-12.

24. Laver JH, Hulsey TC, Jones JP, Gautreaux M, Barredo JC, Abboud MR. Assessment of barriers to bone marrow donation by unrelated African-American potential donors. Biol Blood Marrow Transplant. 2001;7(1):45-8.

25. Johansen KA, Schneider JF, McCaffree MA, Woods GL. Efforts of the United States' National Marrow Donor Program and Registry to improve utilization and representation of minority donors. Transfus Med. 2008;18(4):250-9.

26. Schmidt AH, Solloch UV, Baier D, Yazici B, Ozcan M, Stahr A, et al. Criteria for initiation and evaluation of minority donor programs and application to the example of donors of Turkish descent in Germany. Bone Marrow Transplant. 2009;44(7):405-12.

27. Schmidt AH, Stahr A, Baier D, Schumacher S, Ehninger G, Rutt C. Selective recruitment of stem cell donors with rare human leukocyte antigen phenotypes. Bone Marrow Transplant. 2007;40(9):823-30.

28. Pingel J, Solloch UV, Hofmann JA, Lange V, Ehninger G, Schmidt AH. High-resolution HLA haplotype frequencies of stem cell donors in Germany with foreign parentage: how can they be used to improve unrelated donor searches? Hum Immunol. 2013;74(3):330-40.

29. Tshabalala M, Mellet J, Pepper MS. Human Leukocyte Antigen Diversity: A Southern African Perspective. J Immunol Res. 2015;746151(10):12.

30. Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet. 1995;57(1):133-49.

31. Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, et al. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet. 2000;66(3):979-88.

32. Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, et al. Nuclear DNA diversity in worldwide distributed human populations. Gene. 1997;205(1-2):161-71.

33. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-Driven Selection and Worldwide HLA Class I Diversity. Current Biology. 2005;15(11):1022-7.

34. Cao K, Moormann AM, Lyke KE, Masaberg C, Sumba OP, Doumbo OK, et al. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens. 2004;63(4):293-325.

35. Gourraud PA, Pappas DJ, Baouz A, Balere ML, Garnier F, Marry E. High-resolution HLA-A, HLA-B, and HLA-DRB1 haplotype frequencies from the French Bone Marrow Donor Registry. Hum Immunol. 2015;76(5):381-4.

36. Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet. 1995;56(3):799-810.
37. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 1977;39(1):1-38.
38. Schäfer C, Schmidt AH, Sauter J. Hapl-o-Mat: open-source software for HLA haplotype frequency estimation from ambiguous and heterogeneous data. BMC Bioinformatics. [journal article]. 2017;18(1):284.
39. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics. 1992;48(2):361-72.
40. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 2010;10(3):564-7.
41. Barbujani G, Ghirotto S, Tassi F. Nine things to remember about human genome diversity. Tissue Antigens. 2013;82(3):155-64.
42. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Human Genomics. 2015;9(1):1.
43. Schmidt AH, Baier D, Solloch UV, Stahr A, Cereb N, Wassmuth R, et al. Estimation of high-resolution HLA-A, -B, -C, -DRB1 allele and haplotype frequencies based on 8862 German stem cell donors and implications for strategic donor registry planning. Hum Immunol. 2009;70(11):895-902.
44. Martinez-Laso J, Sartakova M, Allende L, Konenkov V, Moscoso J, Silvera-Redondo C, et al. HLA molecular markers in Tuvinians: a population with both Oriental and Caucasoid characteristics. Ann Hum Genet. 2001;65(Pt 3):245-61.
45. Carvalho MG, Tsuneto LT, Moita Neto JM, Sousa LC, Sales Filho HL, Macedo MB, et al. HLA-A, HLA-B and HLA-DRB1 haplotype frequencies in Piaui's volunteer bone marrow donors enrolled at the Brazilian registry. Hum Immunol. 2013;74(12):1598-602.
46. Romon I, Montes C, Ligeiro D, Trindade H, Sanchez-Mazas A, Nunes JM, et al. Mapping the HLA diversity of the Iberian Peninsula. Hum Immunol. 2016;77(10):832-40.
47. Single RM, Meyer D, Hollenbach JA, Nelson MP, Noble JA, Erlich HA, et al. Haplotype frequency estimation in patient populations: the effect of departures from

Hardy-Weinberg proportions and collapsing over a locus in the HLA region. Genet Epidemiol. 2002;22(2):186-95.

48. Hurley CK, Maiers M, Marsh SG, Oudshoorn M. Overview of registries, HLA typing and diversity, and search algorithms. Tissue Antigens. 2007;1:3-5.

49. Hayhurst JD, du Toit ED, Borrill V, Schlaphoff TEA, Brosnan N, Marsh SGE. Two novel HLA alleles, HLA-A*30:02:01:03 and HLA-C*08:113, identified in a South African bone marrow donor. Tissue Antigens. 2015;85(4):291-3.

50. Paximadis M, Mathebula TY, Gentle NL, Vardas E, Colvin M, Gray CM, et al. Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population. Human Immunol. 2012;73:80-92.

51. Alfirevic A, Gonzalez-Galarza F, Bell C, Martinsson K, Platt V, Bretland G, et al. In silicoanalysis of HLA associations with drug-induced liver injury: use of a HLA-genotyped DNA archive from healthy volunteers. Genome Medicine. [journal article]. 2012 June 25;4(6):51.

52. Metcalfe S, Roger M, Faucher M-C, Coutlée F, Franco EL, Brassard P. The frequency of HLA alleles in a population of Inuit women of northern Quebec. International Journal of Circumpolar Health. 2013;72:10.3402/ijch.v72i0.21350.

53. Zhou X-Y, Zhu F-M, Li J-P, Mao W, Zhang D-M, Liu M-L, et al. High-Resolution Analyses of Human Leukocyte Antigens Allele and Haplotype Frequencies Based on 169,995 Volunteers from the China Bone Marrow Donor Registry Program. PloS ONE. 2015;10(9):e0139485.

54. Sanchez-Mazas A, Fernandez-Vina M, Middleton D, Hollenbach JA, Buhler S, Di D, et al. Immunogenetics as a tool in anthropological studies. Immunology. 2011;133(2):143-64.

55. http://www.who.int/hiv/pub/arv/summary-recommendations.pdf?ua=1.

56. González-Galarza Faviel F, Takeshita Louise YC, Santos Eduardo JM, Kempson F, Maia Maria Helena T, Silva Andrea Luciana Soares d, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Research. 2015;43(D1):D784-D8.

57. Sauter J, Solloch UV, Giani AS, Hofmann JA, Schmidt AH. Simulation shows that HLA-matched stem cell donors can remain unidentified in donor searches. Scientific Reports. [Article]. 2016;6:21149.

58. Gourraud PA, Lamiraux P, El-Kadhi N, Raffoux C, Cambon-Thomsen A. Inferred HLA haplotype information for donors from hematopoietic stem cells donor registries. Hum Immunol. 2005;66(5):563-70.

59. Castelli EC, Mendes-Junior CT, Veiga-Castelli LC, Pereira NF, Petzl-Erler ML, Donadi EA. Evaluation of computational methods for the reconstruction of HLA haplotypes. Tissue Antigens. 2010;76(6):459-66.

60. Dubois V, Detrait M, Sobh M, Morisset S, Labussiere H, Giannoli C, et al. Using EasyMatch(R) to anticipate the identification of an HLA identical unrelated donor: A validated efficient time and cost saving method. Hum Immunol. 2016;77(11):1008-15.

61. Dehn J, Setterholm M, Buck K, Kempenich J, Beduhn B, Gragert L, et al. HapLogic: A Predictive Human Leukocyte Antigen-Matching Algorithm to Enhance Rapid Identification of the Optimal Unrelated Hematopoietic Stem Cell Sources for Transplantation. Biol Blood Marrow Transplant. 2016;22(11):2038-46.

62. Gragert L, Eapen M, Williams E, Freeman J, Spellman S, Baitty R, et al. HLA Match Likelihoods for Hematopoietic Stem-Cell Grafts in the U.S. Registry. New England Journal of Medicine. 2014;371(4):339-48.

63. Bochtler W, Beth M, Eberhard H-P, Müller CR. OptiMatch (R) - a universally configurable HLA matching framework. 22nd European Immunogenetics and Histocompatibility Conference; Berne, Switzerland: Tissue Antigens; 2008. p. 265-398.

64. Arnaiz-Villena A, Reguera R, Parga-Lozano C, Abd-El-Fatah-Khalil S, Monleon L, Barbolla L, et al. HLA Genes in Afro-American Colombians (San Basilio de Palenque): The First Free Africans in America. The Open Immunology Journal. 2009; 2:59-66.

65. Arnaiz-Villena A, Palacio-Gruber J, Muniz E, Campos C, Alonso-Rubio J, Gomez-Casado E, et al. Genetic HLA Study of Kurds in Iraq, Iran and Tbilisi (Caucasus, Georgia): Relatedness and Medical Implications. PloS ONE. 2017;12(1).

66. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PloS ONE. 2011;6(2):0014643.

67. Parham P, Ohta T. Population biology of antigen presentation by MHC class I molecules. Science. 1996;272(5258):67-74.

68. Kijak GH, Walsh AM, Koehler RN, Moqueet N, Eller LA, Eller M, et al. HLA class I allele and haplotype diversity in Ugandans supports the presence of a major east African genetic cluster. Tissue Antigens. 2009;73(3):262-9.

CHAPTER 4

Mixed resolution HLA~A, ~B, ~C, ~DRB1, ~DQA1, ~DQB1 and ~DPB1 diversity in South African populations

Mqondisi Tshabalala[1], Kuben Vather[2], Derrick Nelson[2], Fathima Mohamed[2], Alan Christoffels[3] and Michael S. Pepper[1*]

[1]Institute for Cellular and Molecular Medicine, Department of Immunology, and SAMRC Extramural Unit for Stem Cell Research and Therapy, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

[2]South African National Blood Services, Constantia Kloof Extension 22, Weltevreden Park 1715

[3]SAMRC Bioinformatics Unit, South African National Bioinformatics Institute, University of Western Cape, Robert Sibukwe Road, Bellville, Cape Town 7535, South Africa

This chapter has been prepared in the format of a manuscript, and is under review by the BMC Medical Genetics (manuscript number MGTC-D-19-00228). I designed the study, performed experimental work, data analysis and drafted the manuscript. Vather, Derrick Nelson and Fathima Mohamed recruited the study participants, provided HLA data and contributed in manuscript writing. Prof Alan Christoffels contributed in data analysis and writing the manuscript. Prof M.S Pepper conceived the study, obtained funding for the study and provided critical review of the manuscript.

4.1 Abstract

Background/Aim: Lack of HLA data in southern African populations hampers disease association studies and our understanding of genetic diversity in these populations. We aimed to determine HLA diversity in South African populations using 3007 high resolution HLA ~A, HLA ~B, HLA ~C, HLA ~DRB1, HLA ~DQA1and HLA ~DQB1 and 51 891 low resolution previously typed individuals.

Materials and Methods: We determined allele and haplotype frequencies, deviations from Hardy-Weinberg equilibrium (HWE), linkage disequilibrium and neutrality test. South African HLA class I data was additionally compared to other global populations using non-metrical multidimensional scaling (NMDS), genetic distances and principal component analysis

Results: All loci strongly ($p<0.0001$) deviated from HWE, coupled with excessive heterozygosity in most loci. Two of the three most frequent alleles HLA ~DQA1*05:02 (0.370) and HLA ~C*17:01 (0.281) were previously reported in South African populations at lower frequencies. NMDS showed genetic distinctness of South African populations. Phylogenetic and principal component analysis clustered our current dataset with previous South African studies. Additionally, South Africans seem to be related to other sub Saharan populations using HLA class I allele frequencies.

Conclusion: We uniquely provide a large sample size HLA data from South Africans, which might be a useful resource to support anthropological studies, disease association studies, population based vaccine development and door recruitment programs. We additionally provide simulated high resolution HLA class I data to augment the mixed resolution typing results generated from this study.

Key words:
HLA, Mixed resolution HLA typing, South Africa

4.2 Introduction

The human leukocyte antigen (HLA) gene region is considered to be one of the most polymorphic regions in the human genome[1,2]. Currently, there are 18 955 reported alleles in the IMGT/HLA database (3.33 release of July 2018)[3]. HLA genes encode proteins involved in antigen presentation[4], and play a key determining role in transplantation clinical outcomes[5-12]. Despite the growing documented evidence of genetic diversity of Africans[13-17], there remains an information gap on HLA diversity in these populations (reviewed in Chapter 2[18]). This lack of HLA data hampers disease association studies (reviewed in[19]), population specific vaccine development[20] and donor recruitment programs into registries [21]. Additionally, there is high disease burden in these populations[22]; hence understanding HLA diversity will further support efforts to eliminate these health challenges.

In addition to its key role in the human immune system, HLA has been used to understand human genetic diversity, population genetics and anthropology. HLA has been widely used to understand genetic relatedness of different populations as well as demographic events in those populations[23]. The HLA genetic makeup of populations provides insight into their histories including selective pressures by pathogens[16] migration, admixture and changes in population size[24-27]. The availability of population HLA data is thus critical, in understanding peopling history and general evolution of the human immune system[28,29]

The South African population comprises 55.6 million people (2011 census)[30] who are burdened by disease and harbor one of the oldest modern human lineages, Homo naledi[31]. Additionally, new HLA alleles have been reported in South African populations[32,33] supporting the idea of high genetic diversity in these populations[34,35]. In Chapter 3, allele and haplotype frequencies from the South African Bone Marrow Registry (SABMR) are described in an effort to understand HLA diversity in South Africans[36]. The current study is aimed at improving our understanding of HLA diversity in South Africans using retrospectively typed individuals in the National Health laboratory Services (NHLS) and the South African National Blood Transfusion Services (SANBS). We additionally sought to compare HLA data from South Africans with other global populations using population genetics approaches.

4.3 Methods

4.3.1 Study population, HLA data access and ethics

Approval for this study was granted by Research Ethics Committee of the University of Pretoria Faculty of Health Sciences (approval no. 220/2015), the SANBS Human Research Ethics Committee (SANBS HREC) and NHLS Academic Affairs and Research. We analysed a combined total (SANBS and NHLS) of 3007 high resolution (four digit typing HLA ~A, HLA ~B, HLA ~C, HLA ~DRB1, HLA ~DQA1and HLA ~DQB1) and 51 891 low resolution (two digit HLA ~A, HLA ~B, HLA ~C, HLA ~DRB1, HLA ~DQA1, HLA ~DQB1 and HLA ~DPB1) results. The mixed resolution typing data (a mixture of 2 and 4 digit typing resolution) set has resulted from the retrospective nature of the study, with typing methods evolving from low resolution serology typing to higher resolution DNA based methods in SANBS and NHLS. All available HLA data from SANBS (up to 20 November 2016) plus NHLS data (05 June 2003 to 12 April 2016) was accessed. The NHLS offers national diagnostic pathology services (http://www.nhls.ac.za/) whilst SANBS aims to supply safe blood and blood products (https://sanbs.org.za/). Only HLA data was accessed, with no additional data accessed due to ethical considerations. Participants' personal identifiers were not accessed to maintain confidentiality following the Helsinki ethical guidelines[37]. All the accessed HLA data was checked for allele validity, and all pre-2010 nomenclature designations converted using current nomenclature conversion tables and conversion tools provided by IMGT/HLA (https://www.ebi.ac.uk). HLA data missingness in our dataset was defined by the lack of typing methods to call two alleles at a given locus, resulting in one allele for that individual at that particular locus. Unfortunately, a distinction between homozygous typing and data missingness could not be established due to the retrospective nature of the study.

4.3.2 Statistical analysis

Low (2 digit) and high (4 digit) resolution data were separately analysed to estimate LD, HWE proportions, homozygosity test of neutrality, allele and haplotype frequencies. Low and high resolution typing allele frequencies were determined by

direct counting, and haplotype frequencies estimated by resolving phase and allelic ambiguities using the expectation-maximization (EM) algorithm[38,39] both implemented in PyPop ver 0.7.0[40]. Excoffier et al[38] allows estimation of random haplotypes based on sample allele frequencies. For pair wise linkage disequilibrium (LD), we used Hedrick's D′[41] and Cramer's V Statistic (W_n)[42], all implemented in PyPop ver 0.7.0[40]. HLA genotypes were converted to Arlequin v3.5.2[43] input files using CREATEv1.37 software[44] to assess deviations from Hardy-Weinberg equilibrium (HWE) {modified hidden Markov chain[45] with 100 000 dememorization steps}. Slatkin's implementation of Ewens-Watterson homozygosity test of neutrality[46,47] was done in PyPop ver 0.7.0[40].

4.3.3 Population comparison

To better understand the HLA diversity in our dataset, we compared our findings to other global populations. Our current data was compared with multiple population datasets from selected world regions by non-metrical multidimensional scaling analysis (NMDS) in gene[RATE] tools[48]. Due to the HLA mixed resolution typing nature and data missingness in our dataset, we performed HLA class I completion of our data set to get high resolution (four digit typing) using the PhyloD tool as previously described[49]. The PhyloD HLA completion tool uses statistical *in silico* methods to probabilistically predict four digit HLA -A, -B and -C[49]. We further compared our class I HLA allele frequency data with PhyloD generated allele frequency data[49], and 28 other publicly available HLA ~A, ~B and ~C allele frequency (four digit resolution) sub Saharan Africa data from the allele frequency database (AFND)[50] including previous South African studies[36,51-53]. Specifically, our HLA data (RSA) was compared with the following AFND defined populations (population codes we used for phylogenetic analysis): Burkina Faso Fulani (BFF)[54] Burkina Faso Mossi (BFM)[54], Burkina Faso Rimaibe (BFR)[54], Cameroon Baka Pygmy (CBP)[55], Cameroon Bakola Pygmy (CBkP)[56], Cameroon Bamileke (CaB)[55], Cameroon Beti (CBt)[55], Cameroon Sawa (CSw)[55], Central African Republic Mbenzele Pygmy (CARMP)[56], Ghana Ga-Adangbe (GGA)[57], Kenya (KEN)[58], Kenya Luo (KENL)[59], Kenya Nandi (KENN)[59], Kenya, Nyanza Province, Luo tribe (KENNy)[60], PhyloD generated data (PSA)[49], Rwanda (RWA)[61], Senegal Niokholo

Mandenka (SenMAND)[62], South Africa Black (SoAB)[33], South Africa Caucasians (SoAC)[33], South Africa Natal Tamil (SANT)[63], South Africa Natal Zulu (SANZ)[64], South Africa Worcester (WOR)[51], South African Bone Marrow Registry (SAB) described in Chapter 3[36], South African Indian population (SAI)[52], South African Mixed ancestry (RMX)[53], Uganda Kampala (UgaKam)[59], Uganda Kampala pop 2 (UgaKam2)[27], Zambia Lusaka (ZaL)[59] and Zimbabwe Harare Shona (ZiHS)[65]. HLA class I allele frequencies from the above 30 populations were used to compute pair wise population differentiation (F_{ST}) and Neis' genetic distances[66] in POPTREE software[67,68]. An unrooted tree was constructed based on Neighbour-Joining (NJ) method[69] implemented in POPTREE software[67,68] using Nei's genetic distances. Furthermore, the pair wise F_{ST} matrix was used for principal component analysis (PCA) in ClustVis (a web tool for visualizing clustering of multivariate data using PCA and heatmap)[70].

4.4 Results

4.4.1 HWE proportions and neutrality test

All loci (both low resolution and high resolution typing) showed a strong significant deviation from the expected HWE proportions ($p<0.0001$) as detailed in Table 4.1. Generally, more genotypes were observed in low resolution compared to high resolution typing which was characterized by data missingness (Table 4.1). Extremely excessive heterozygosity ($p<0.0001$) in high resolution HLA ~A and excessive heterozygosity ($p<0.05$) in low resolution HLA ~B, ~C, ~DQA1 and ~DPB1 was observed. Excessive homozygosity ($p>0.05$) was observed in high resolution HLA ~B, ~C, ~DRB1, ~DQA1 and ~DQB1 and low resolution HLA ~A, ~DRB1 and ~DQB1 (Table 4.2).

4.4.2 Allele frequencies

The full list of alleles is detailed in Supplementary Table 1 (Table S4.1) which includes both low and high resolution typing frequencies. The top 20 most frequent

alleles across the different loci typed at low or high resolution are summarized in Table 4.3. HLA ~ DQB1*06 (0.428), ~DPB1*52 (0.427) and ~DPB1*53 (0.407) were the three most common allele groups (low resolution typing). High resolution typed HLA ~ DQA1*05:02 (0.370), ~DQA1*04:02 (0.303) and ~C*17:01 (0.281) were the three most common alleles in our dataset (Table 3). We additionally include PhyloD generated[49] HLA ~A, ~B and ~C estimated genotypes (with probabilities) and allele frequencies in supplementary Table 4.2 (Table S4.2) for population comparison and as a future resource for other researchers.

4.4.3 Haplotype frequencies and LD

For low resolution typing (two digit), all two, three, four, five and six haplotype frequencies are detailed in Supplementary Table 4.3 (Table S4.3), with the 20 most frequent haplotypes summarized in Table 4.4. DQB1*03~DPB1*53 (0.297), B*44~C*07~DPB1*53 (0.333), B*44~C*07~DQB1*03~DPB1*53 (0.333), B*44~C*07~DRB1*04~DQB1*03~DPB1*53 (0.333) and A*02~B*58~C*07~DRB1*11~DQA1*05~DQB1*03 (0.018) were the most common computed two, three, four, five and six loci haplotypes. PyPop ver 0.7.0[40] could not estimate some haplotype frequencies due to an excessive number of rows, or no data left after filtering (Table S4.3). No seven loci haplotypes were estimated for low resolution typing (Table S4.3). The most common estimated high resolution two, three and four loci haplotypes were A*02:05~C*14:02 (0.500), A*30:02~B*45:01~DRB1*15:03 (1.00) and A*30:02~B*45:01~DRB1*15:03~DQB1*05:01 (0.500) respectively as summarised in Table 4.5 and Supplementary Table S4.4. PyPop ver 0.7.0[40] could not estimate any five and six loci haplotypes at high resolution (Table S4.4) due to lack of data after filtering. In all low and high resolution typing results, all pair wise linkage disequilibrium (LD) measured by Hedrick's D′[41] and Cramer's V Statistic (W_n)[42] were strongly significant ($p<0.0001$) and significant ($p<0.05$) except for insignificant low resolution A:DPB1, C:DPB1 and high resolution C:DQB1 loci pairs (Table 4.6).

4.4.4 Population comparison

NMDS analyses implemented in gene[RATE] tools[48] suggest high genetic diversity of high resolution HLA ~DRB1 and low resolution HLA ~A and ~DRB1 (Figure 4.1). Global populations show less diversity in high resolution HLA ~A loci, with only two clusters (our data set and other populations) shown by NMDS (Figure 4.1). Additionally, our data set distinctly clustered away from other global populations (Supplementary Figures 4.1 and 4.2~Figure S4.1 and Figure S4.2 respectively). NMDS analysis suggests high genetic diversity in high resolution HLA ~B, ~DQA1, ~DRB1, ~DQB1 (Figure S4.1) and low resolution HLA ~A, ~B, ~C, ~DRB1, ~DQA1 and ~DQB1 (Figure S4.2) with low diversity in low resolution HLA ~C loci (Figure S4.2). Global NMSD comparison for HLA ~DPB1 loci was not available in gene[rate] tools (both at low and high resolution)[48]. The NJ generated tree (Figure 4.2) shows a close relation of the current data (RSA) with other previously described South African studies ~SoAC[33], SoAB[33] and SANT[63], but not with SANZ[64], SAB[36], SAI[52], RMX[53] and WOR[51]. Interestingly, although our probability simulated data PSA did not cluster with the data it was generated from (RSA), it was closely related to a previous South African study SAB[36] (Figure 4.2). Pair wise F_{ST} based principal component analysis showed 69.6% and 11.1% total population variability explained by PCA1 and PCA 2 respectively (Figure 4.3). PCA (Figure 4.3) suggests Central African Republic Mbenzele Pygmy (CARMP) are completely different from other sub Saharan populations. Additional outliers include Cameroon Baka Pygmy (CBP) and Cameroon Sawa (CSw). Our data (RSA) seem to cluster together with Cameroon Bakola Pygmy (CBkP) and South Africa Natal Tamil (SANT). Probability simulated data (PSA) clusters with the other remaining populations, with Ghana Ga-Adangbe (GGA), Senegal Niokholo Mandenka (SenMAND and Zambia Lusaka (ZaL) forming a small separate cluster (Figure 4.3).

Table 4.1 HWE parameters for low and high resolution typing

Exact Test using Markov chain for all loci with 100000 dememorization steps

	Locus	#Genotypes	Obs Het	Exp Het	p-HWE
High resolution	HLA ~A	111	0.07207	0.96714	<0.0001*
	HLA ~B	345	0.27536	0.95592	<0.0001*
	HLA ~C	128	0.03906	0.86489	<0.0001*
	HLA ~DRB1	1927	0.10223	0.94003	0.0015**
	HLA ~DQA1	104	0.12500	0.71363	<0.0001*
	HLA ~DQB1	325	0.55077	0.93905	<0.0001*
Low resolution	HLA ~A	23048	0.92030	0.90148	<0.0001*
	HLA ~B	25434	0.97067	0.93540	<0.0001*
	HLA ~C	3510	0.74074	0.86568	<0.0001*
	HLA ~DRB1	13605	0.66645	0.88341	<0.0001*
	HLA ~DQA1	221	0.31674	0.76767	<0.0001*
	HLA ~DQB1	8057	0.25977	0.72241	<0.0001*
	HLA ~DPB1	198	1.00000	0.62638	<0.0001*

#Genotypes (number of genotypes), **Obs Het** (observed heterozygosity), **Exp Het** (expected Heterozygosity), **p-HWE** *(p* value for HWE deviation), **significant (*highly significant) at p<0.01 (p<0.0001) difference between observed and expected heterozygosity

Table 4.2 Slatkin's implementation of Ewens-Watterson homozygosity test of neutrality

Observed homozygosity (homozygosity F statistic ~ a sum of squared allele frequencies) compared to expected homozygosity (simulated under neutrality/equilibrium expectations for the same sample taking into account unique alleles) [46,47].

	Locus	Observed F	Expected F	Variance in F	Fnd	Fp
High resolution	HLA ~A	0.0362	0.0657	0.0003	-1.7622	<0.0001**
	HLA ~B	0.0461	0.0367	0.0001	1.2062	0.8965
	HLA ~C	0.1385	0.1496	0.0026	-0.2165	0.5070
	HLA ~DRB1	0.0602	0.0446	0.0001	1.3792	0.9163
	HLA ~DQA1	0.2898	0.4738	0.0262	-1.1368	0.0960
	HLA ~DQB1	0.0626	0.1091	0.0013	-1.3042	0.0133
Low resolution	HLA ~A	0.0985	0.3228	0.0182	-1.6614	<0.0001**
	HLA ~B	0.0646	-0.3230	0.0179	2.8974	0.9999.
	HLA ~C	0.1344	0.3735	0.0227	-1.5871	0.0007
	HLA ~DRB1	0.1166	0.4355	0.0292	-1.8656	<0.0001**
	HLA ~DQA1	0.2341	0.5145	0.0310	-1.5917	0.0071*
	HLA ~DQB1	0.2776	0.6947	0.0428	-2.0154	0.0044*
	HLA ~DPB1	0.3752	0.7331	0.0367	-1.8675	0.0186*

Observed F: observed homozygosity F statistic, **Expected F:** expected homozygosity F statistic, **Fp:** p value F statistic **Fnd:** Normalised deviate of F statistic **highly statistically significant at p<0.0001 *significant at p<0.05

Table 4.3 Top 20 HLA alleles by locus and typing resolution (Full list in S4.1)

Low resolution (two digit)			High resolution (four digit)		
loci	**freq**	**count**	**loci**	**freq**	**count**
DQB1*06	0.428	6887	DQA1*05:02	0.370	77
DPB1*52	0.427	169	DQA1*04:02	0.303	63
DPB1*53	0.407	161	C*17:01	0.281	72
DQA1*01	0.342	151	DQA1*02:01	0.240	50
C*07	0.282	1979	C*16:01	0.141	36
DQA1*05	0.267	118	DRB1*15:03	0.135	521
DQB1*03	0.215	3457	B*15:10	0.132	91
A*02	0.206	9501	DQB1*03:19	0.129	84
DRB1*15	0.197	5357	DRB1*15:01	0.122	471
DQB1*05	0.179	2889	C*03:04	0.121	31
DRB1*13	0.171	4644	C*16:02	0.109	28
DPB1*51	0.167	66	A*43:01	0.093	21
DQA1*04	0.143	63	DRB1*13:01	0.090	347
C*06	0.136	956	B*42:01	0.087	60
B*07	0.125	6339	B*15:03	0.086	59
DRB1*11	0.124	3369	C*14:02	0.082	21
DQA1*02	0.118	52	DQB1*02:01	0.080	52
A*03	0.116	5363	DQB1*05:01	0.078	51
A*01	0.113	5207	DQB1*03:01	0.077	50
DQB1*02	0.110	1767	DQB1*06:02	0.074	48

Allele frequency (freq) number of individuals with allele (count)

Table 4.4 Top twenty most frequent low resolution two, three, four, five and six loci haplotype frequencies (Full list in Table S4.3)

Two loci	freq	Three loci	freq	Four loci	freq	Five loci	freq	Six loci	freq
DQB1*03~DPB1*53	0.297	B*44~C*07~DPB1*53	0.333	B*44~C*07~DQB1*03~DPB1*53	0.333	B*44~C*07~DRB1*04~DQB1*03~DPB1*53	0.333	A*02~B*58~C*07~DRB1*11~DQA1*05~DQB1*03	0.018
DQB1*02~DPB1*52	0.277	DRB1*04~DQB1*03~DPB1*53	0.265	A*01~B*44~C*07~DPB1*52	0.167	A*01~B*44~C*07~DRB1*04~DPB1*52	0.167	A*30~B*42~C*17~DRB1*03~DQA1*04~DQB1*04	0.016
C*04~DPB1*52	0.250	C*04~DRB1*04~DPB1*52	0.250	A*02~B*44~C*05~DPB1*51	0.167	A*02~B*44~C*05~DRB1*04~DPB1*51	0.167	A*01~B*08~C*07~DRB1*03~DQA1*05~DQB1*02	0.014
C*07~DPB1*53	0.250	C*07~DRB1*04~DPB1*53	0.250	A*03~B*07~C*07~DPB1*53	0.167	A*03~B*07~C*07~DRB1*15~DPB1*53	0.167	A*30~B*42~C*17~DRB1*12~DQA1*01~DQB1*05	0.012
DRB1*04~DPB1*53	0.245	C*04~DQB1*03~DPB1*52	0.250	A*24~B*08~C*05~DPB1*53	0.167	A*24~B*08~C*05~DRB1*03~DPB1*53	0.167	A*01~B*15~C*03~DRB1*03~DQA1*05~DQB1*03	0.009
DQA1*01~DQB1*06	0.188	C*07~DQB1*03~DPB1*53	0.250	A*68~B*15~C*04~DPB1*53	0.167	A*68~B*15~C*03~DRB1*03~DPB1*53	0.167	A*02~B*08~C*07~DRB1*04~DQA1*03~DQB1*03	0.009
DRB1*07~DPB1*53	0.174	DRB1*03~DQB1*02~DPB1*52	0.236	A*74~B*35~C*03~DPB1*52	0.167	A*74~B*35~C*04~DRB1*04~DPB1*52	0.167	A*30~B*42~C*17~DRB1*15~DQA1*04~DQB1*04	0.009
DRB1*15~DQB1*06	0.159	A*02~DQB1*03~DPB1*53	0.191	B*07~C*05~DQB1*06~DPB1*51	0.167	B*07~C*05~DRB1*15~DQB1*06~DPB1*51	0.167	A*68~B*15~C*03~DRB1*15~DQA1*05~DQB1*06	0.009
A*02~DPB1*53	0.158	B*07~C*05~DPB1*51	0.167	B*08~C*05~DQB1*02~DPB1*52	0.167	B*08~C*05~DRB1*03~DQB1*02~DPB1*52	0.167	A*01~B*44~C*07~DRB1*07~DQA1*02~DQB1*02	0.007
DQB1*06~DPB1*51	0.153	B*08~C*05~DPB1*52	0.167	B*15~C*03~DQB1*03~DPB1*52	0.167	B*15~C*04~DRB1*04~DQB1*02~DPB1*53	0.167	A*23~B*07~C*07~DRB1*15~DQA1*01~DQB1*06	0.007
DRB1*13~DQB1*06	0.146	B*15~C*04~DPB1*52	0.167	B*35~C*04~DQB1*02~DPB1*53	0.167	B*35~C*03~DRB1*03~DQB1*03~DPB1*52	0.167	A*24~B*58~C*06~DRB1*15~DQA1*02~DQB1*02	0.007
DRB1*03~DPB1*52	0.130	B*35~C*03~DPB1*53	0.167	A*01~B*08~DRB1*03~DPB1*52	0.111	B*42~C*17~DRB1*03~DQA1*04~DQB1*04	0.027	A*29~B*07~C*07~DRB1*01~DQA1*05~DQB1*05	0.007
DQA1*05~DQB1*03	0.127	A*01~C*07~DPB1*53	0.125	A*03~B*07~DRB1*15~DPB1*51	0.111	A*02~C*07~DRB1*11~DQA1*05~DQB1*03	0.026	A*30~B*18~C*07~DRB1*11~DQA1*01~DQB1*06	0.007
C*03~DPB1*53	0.125	A*02~C*04~DPB1*53	0.125	B*42~C*17~DQA1*04~DQB1*04	0.041	B*08~C*07~DRB1*03~DQA1*05~DQB1*02	0.023	A*30~B*42~C*17~DRB1*11~DQA1*01~DQB1*03	0.007

C*05~DPB1*51	0.125	A*02~C*07~DPB1*51	0.125	A*02~DRB1*11~DQA1*05~DQB1*03	0.040	A*01~B*08~C*07~DRB1*03~DQB1*02	0.019	A*30~B*58~C*06~DRB1*12~DQA1*01~DQB1*05	0.007
C*05~DPB1*52	0.125	A*03~C*05~DPB1*53	0.125	B*08~C*07~DRB1*03~DQA1*05	0.039	B*58~C*07~DRB1*11~DQA1*05~DQB1*03	0.018	A*33~B*50~C*07~DRB1*15~DQA1*02~DQB1*06	0.007
C*12~DPB1*53	0.125	A*24~C*05~DPB1*52	0.125	C*07~DRB1*11~DQA1*05~DQB1*03	0.039	A*02~B*58~C*07~DRB1*11~DQA1*05	0.018	A*68~B*58~C*06~DRB1*13~DQA1*03~DQB1*03	0.007
A*01~DPB1*52	0.122	A*24~C*12~DPB1*52	0.125	A*01~B*41~DRB1*07~DPB1*52	0.037	A*02~B*58~DRB1*11~DQA1*05~DQB1*03	0.018	A*30~B*58~C*06~DRB1*12~DQA1*01~DQB1*05	0.007
DQA1*01~DQB1*05	0.115	A*68~C*03~DPB1*52	0.125	A*02~B*27~DRB1*04~DPB1*53	0.037	A*03~B*07~C*07~DRB1*15~DQB1*06	0.016	A*33~B*50~C*07~DRB1*15~DQA1*02~DQB1*02	0.007
DRB1*15~DQA1*01	0.112	A*74~C*04~DPB1*53	0.125	A*02~B*44~DRB1*04~DPB1*53	0.037	A*30~B*42~C*17~DRB1*03~DQA1*04	0.016	A*68~B*14~C*08~DRB1*13~DQA1*05~DQB1*03	0.007

“freq” frequency

Table 4.5 The twenty most frequent high resolution two, three, four, five and six loci haplotype frequencies (Full list in Table S4.4)

No data was available after filtering to compute five and six loci haplotype frequencies in Pypop[40]. Only 13 four loci haplotypes were identified.

Two loci	freq	Three loci	freq	Four loci	freq
A*02:05~C*14:02	0.500	A*30:02~B*45:01~DRB1*15:03	1.00	A*30:02~B*45:01~DRB1*15:03~DQB1*05:01	0.500
A*29:02~C*17:01	0.500	DRB1*11:02~DQA1*05:02~DQB1*03:19	1.00	A*30:02~B*45:01~DRB1*15:03~DQB1*06:02	0.500
C*17:01~DQA1*04:02	0.579	C*17:01~DRB1*11:02~DQB1*03:19	0.667	B*42:01~C*17:01~DRB1*15:03~DQA1*04:02	0.571
B*42:01~DQA1*04:02	0.556	B*42:01~C*17:01~DQA1*04:02	0.657	B*42:02~C*17:01~DRB1*11:02~DQB1*03:19	0.333
A*23:01~DQA1*02:01	0.500	A*23:01~DQA1*02:01~DQB1*02:01	0.500	B*15:10~C*17:01~DRB1*11:02~DQB1*03:19	0.167
A*80:01~DQA1*02:01	0.500	A*80:01~DQA1*02:01~DQB1*02:01	0.500	B*52:02~C*03:04~DRB1*11:02~DQB1*03:19	0.167
B*42:01~C*17:01	0.406	B*42:01~DRB1*15:03~DQA1*04:02	0.444	B*41:02~C*17:01~DRB1*11:02~DQB1*03:19	0.167
C*17:01~DQB1*03:19	0.313	C*17:01~DRB1*15:03~DQA1*04:02	0.444	B*41:02~C*17:01~DRB1*15:03~DQB1*03:19	0.167
C*17:01~DQB1*04:01	0.313	A*30:02~B*45:01~DQB1*05:01	0.400	B*42:01~C*17:01~DRB1*11:02~DQA1*04:02	0.143
A*30:02~DRB1*15:03	0.267	A*30:02~B*45:01~DQB1*06:02	0.400	B*42:01~C*17:01~DRB1*03:02~DQA1*04:02	0.071
DQA1*02:01~DQB1*02:01	0.222	C*17:01~DQA1*04:02~DQB1*04:01	0.313	B*57:03~C*17:01~DRB1*03:02~DQA1*04:02	0.071

C*03:04~DQA1*05:02	0.218	C*17:01~DQA1*05:02~DQB1*04:01	0.312	B*15:10~C*17:01~DRB1*15:03~DQA1*05:02	0.071
A*30:02~B*45:01	0.211	A*30:02~DRB1*15:03~DQB1*05:01	0.250	B*42:02~C*03:04~DRB1*15:03~DQA1*05:02	0.071
A*30:02~DQB1*06:02	0.208	A*30:02~DRB1*15:03~DQB1*06:02	0.250	B*42:01~C*17:01~DQA1*04:02~DQB1*04:01	0.345
DRB1*15:03~DQA1*02:01	0.207	A*68:01~DRB1*03:01~DQB1*02:01	0.250	B*15:10~C*03:04~DQA1*05:02~DQB1*04:01	0.220
B*42:02~C*17:01	0.188	A*68:01~DRB1*11:01~DQB1*03:01	0.250	B*42:01~C*17:01~DQA1*05:02~DQB1*04:01	0.155
C*03:04~DQB1*04:01	0.188	B*42:01~C*17:01~DQB1*04:01	0.250	B*15:10~C*03:04~DQA1*04:02~DQB1*04:01	0.155
C*17:01~DRB1*15:03	0.178	B*42:01~DRB1*11:02~DQA1*04:02	0.222	B*15:10~C*17:01~DQA1*05:02~DQB1*04:01	0.125
A*30:02~DQB1*05:01	0.167	B*42:01~C*17:01~DRB1*15:03	0.211		
A*68:01~DQB1*02:01	0.167	B*42:02~DRB1*11:02~DQB1*03:19	0.200		

“freq” frequency

Table 4.6 Pair wise linkage disequilibrium (LD)

	High resolution (four digit)			Low resolution (two digit)		
Locus pair	**D'**	**Wn**	**p-value**	**D'**	**Wn**	**p-value**
A:B	0.0310	0.9501	<0.0001**	#	#	#
A:C	1.0000	1.000	<0.0001**	0.2017	0.1465	<0.0001**
A:DRB1	1.0000	1.000	<0.0001**	#	#	#
A:DQA1	0.0000	0.9721	<0.0001**	0.2778	0.3049	0.0010*
A:DQB1	0.9583	0.7958	<0.0001**	0.06878	0.0839	<0.0001**
A:DPB1	+	+	+	0.6416	0.6240	0.0290^{NS}
B:C	0.9842	0.8967	<0.0001**	0.5119	0.4418	<0.0001**
B:DRB1	0.8110	0.7693	<0.0001**	0.2179	0.1880	<0.0001**
B:DQA1	0.7458	0.6177	0.0050*	0.3420	0.3556	<0.0001**
B:DQB1	0.9328	0.8895	<0.0001**	0.1422	0.1851	<0.0001**
B:DPB1	+	+	+	0.7630	0.7801	<0.0001**
C:DRB1	0.7771	0.6520	<0.0001**	0.2213	0.1573	<0.0001**
C:DQA1	0.5335	0.5335	0.0070*	0.2993	0.2978	<0.0001**
C:DQB1	0.4583	0.7253	0.1061^{NS}	0.1636	0.2334	<0.0001**
C:DPB1	+	+	+	0.9250	0.8165	0.0671^{NS}
DRB1:DQA1	0.5978	0.6758	0.0130*	0.4850	0.4793	<0.0001**
DRB1:DQB1	0.8669	0.7042	<0.0001**	0.5676	0.5173	<0.0001**
DRB1:DPB1	+	+	+	0.9432	0.9679	<0.0001**
DQA1:DQB1	0.6288	0.6693	<0.0001**	0.5302	0.4788	<0.0001**
DQA1:DPB1	+	+	+	#	#	#
DQB1:DPB1	+	+	+	0.7082	0.7347	<0.0001**

D': Hedrick's statistic[41] **Wn:** Cramer's V statistic[42] for global LD, **highly statistically significant at p<0.0001) *significant at p<0.05, ^{NS}not significant p>0.05 +No high resolution HLA –DPB1 data.. #No data after filtering in Pypop.

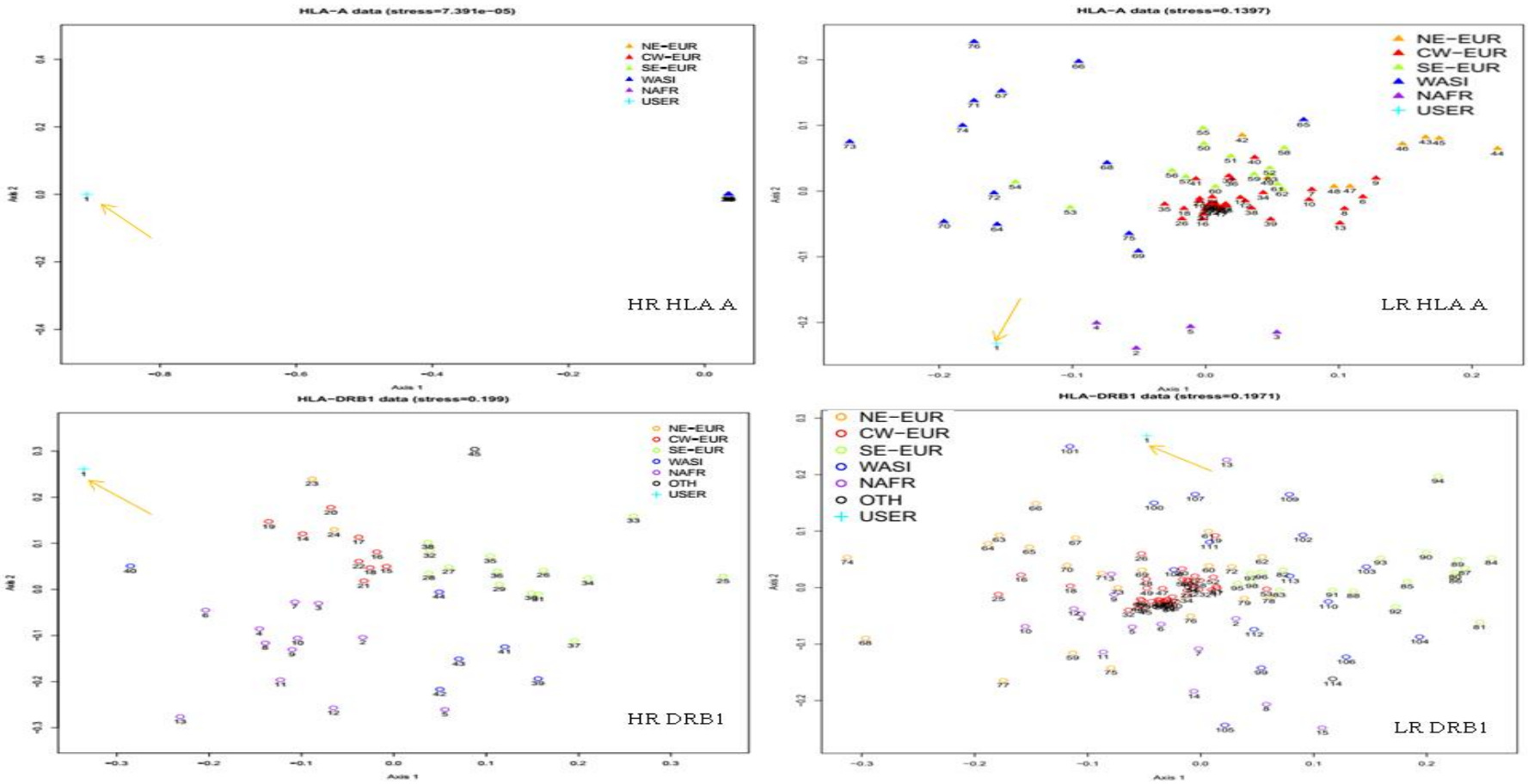

Figure 4.1 South African HLA A and DRB1 non metric multidimensional scaling analysis using gene[rate] tools48. Full list in Figures S1 and S2

The distances between each population correlate to the HLA profile dissimilarity in those populations, for example in HR HLA A, South Africans are distinctly different from the other global populations (clumped together in the far right of the HR HLA A graph). The orientation of axes in NMDS plots is arbitrary and can be rotated to any direction. South African data = orange arrows. HR HLA A (High resolution HLA ~A), LR HLA A (Low resolution HLA ~A), HR DRB1 (high resolution HLA ~DRB1), LR DRB1 (low resolution HLA ~DRB1). NMSD for all loci and description of populations compared are detailed in Supplementary Figures 1 and 2 (Figure S1and Figure S2). NE-EUR (Northeast Europe), CW-EUR (Central and West Europe), SE-EUR (Southeast Europe), WASI (Western Asia), NAFR (Northern Africa), OTH (other European populations of recent origin), USER (South African).

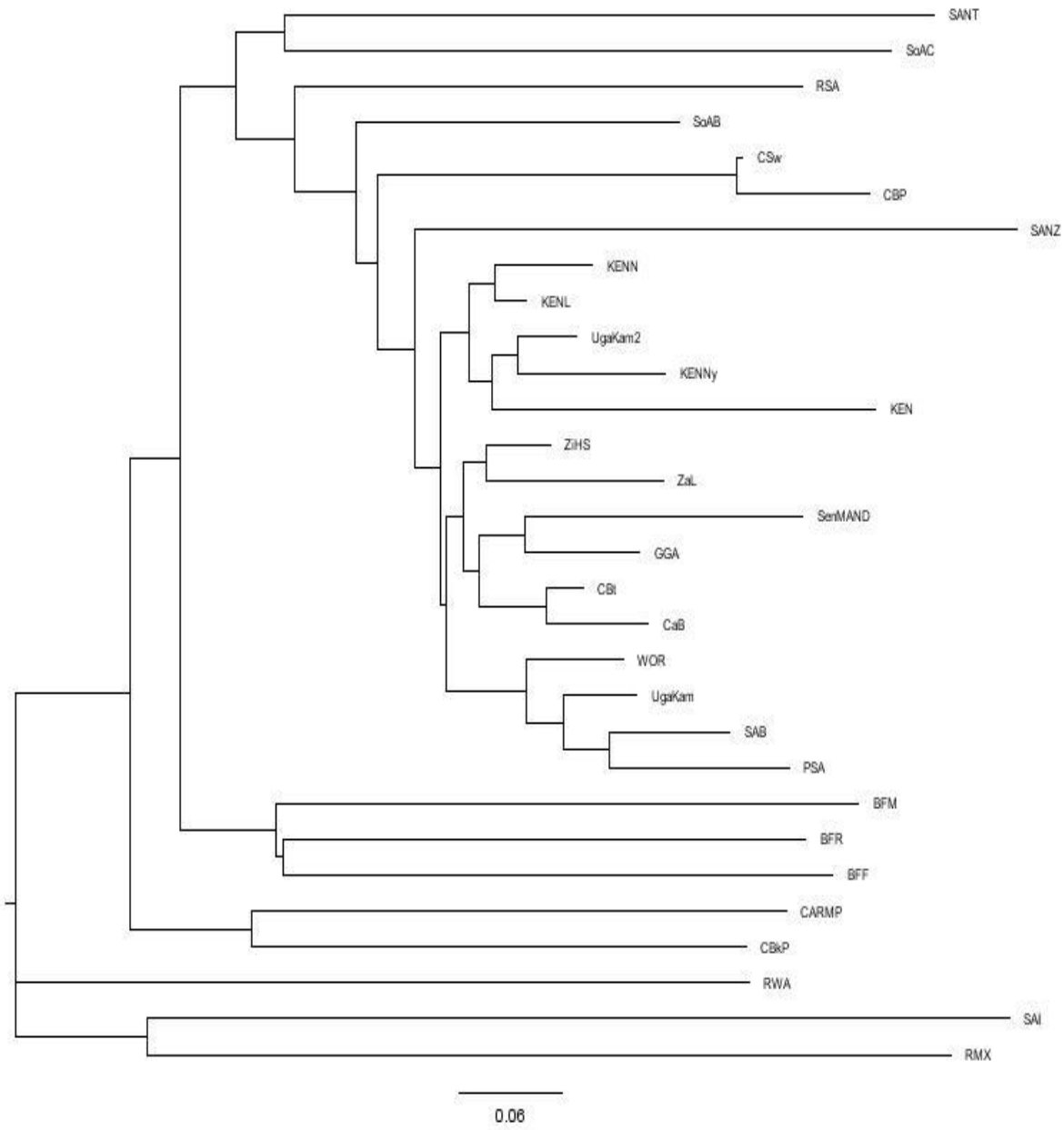

Figure 4.2 Neighbor-Joining tree based on Neis's genetic distance for HLA ~A, ~B and ~C calculated from sub Saharan populations

High resolution (4 digit typing) HLA ~A, ~B and ~C allele frequencies from the following populations were used to determine phylogenetic relatedness. Populations include: Burkina Faso Fulani (BFF)[54] Burkina Faso Mossi (BFM)[54], Burkina Faso Rimaibe (BFR)[54], Cameroon Baka Pygmy (CBP)[55], Cameroon Bakola Pygmy (CBkP)[56], Cameroon Bamileke (CaB)[55], Cameroon Beti (CBt)[55], Cameroon Sawa (CSw)[55], Central African Republic Mbenzele Pygmy (CARMP)[56], Ghana Ga-Adangbe (GGA)[57], Kenya (KEN)[58], Kenya Luo (KENL)[59], Kenya Nandi (KENN)[59], Kenya

Nyanza Province, Luo tribe (KENNy)[60], PhyloD generated data (PSA)[49], RSA (current study), Rwanda (RWA)[61], Senegal Niokholo Mandenka (SenMAND)[62], South Africa Black (SoAB)[33], South Africa Caucasians (SoAC)[33], South Africa Natal Tamil (SANT)[63], South Africa Natal Zulu (SANZ)[64], South Africa Worcester (WOR)[51], South African Bone Marrow Registry (SAB)[36], South African Indian population (SAI)[52], South African Mixed ancestry (RMX)[53], Uganda Kampala (UgaKam)[59], Uganda Kampala pop 2 (UgaKam2)[27], Zambia Lusaka (ZaL)[59] and Zimbabwe Harare Shona (ZiHS)[65]. Current NHLS and SANBS data (RSA) showed phylogenetic relatedness to some previous South African studies SoAC[33], SoAB[33] and SANT[63], but not with SANZ[64], SAB[36], SAI[52], RMX[53] and WOR[51] using the Neis' genetic distance[66].

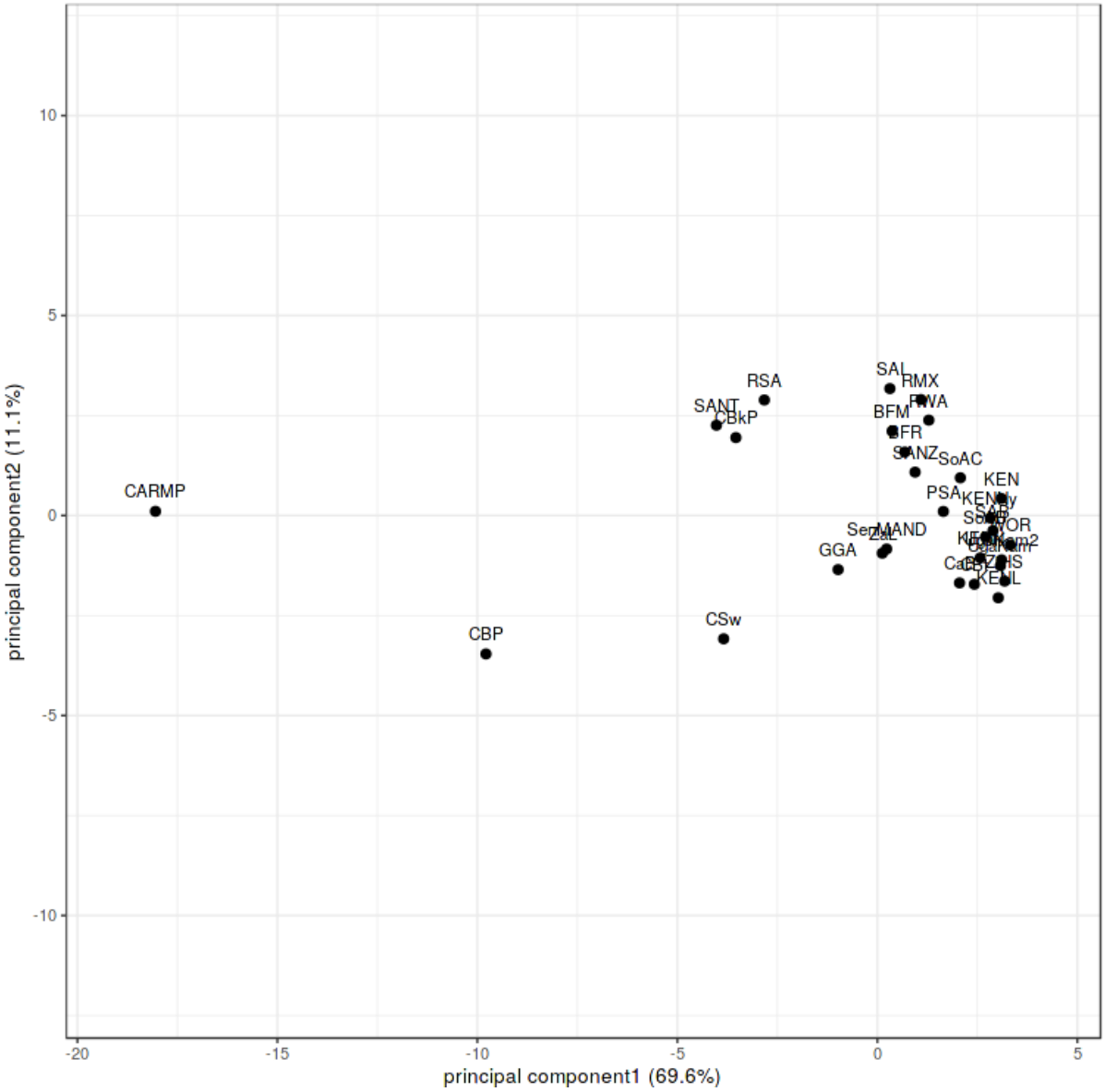

Figure 4.3 F_{ST} based principal component analysis of HLA ~A, ~B and ~C calculated from sub Saharan populations

Burkina Faso Fulani (BFF)[54] Burkina Faso Mossi (BFM)[54], Burkina Faso Rimaibe (BFR)[54], Cameroon Baka Pygmy (CBP)[55], Cameroon Bakola Pygmy (CBkP)[56], Cameroon Bamileke (CaB)[55], Cameroon Beti (CBt)[55], Cameroon Sawa (CSw)[55], Central African Republic Mbenzele Pygmy (CARMP)[56], Ghana Ga-Adangbe (GGA)[57], Kenya (KEN)[58], Kenya Luo (KENL)[59], Kenya Nandi (KENN)[59], Kenya Nyanza Province, Luo tribe (KENNy)[60], PhyloD generated data (PSA)[49], RSA (current study), Rwanda (RWA)[61], Senegal Niokholo Mandenka (SenMAND)[62], South Africa Black (SoAB)[33], South Africa Caucasians(SoAC) [33], South Africa Natal Tamil (SANT)[63], South Africa Natal Zulu (SANZ)[64], South Africa Worcester (WOR)[51], South

African Bone Marrow Registry (SAB)[36], South African Indian population (SAI)[52], South African Mixed ancestry (RMX)[53], Uganda Kampala (UgaKam)[59], Uganda Kampala pop 2 (UgaKam2)[27], Zambia Lusaka (ZaL)[59] and Zimbabwe Harare Shona (ZiHS)[65].

4.5 Discussion

Despite the retrospective nature of this study combined with data missingness, we provide detailed insight into HLA diversity in South African populations using 3007 high (four digit) and 51 891 low (two digit) resolution typing results. We attempted to address data missingness by using our dataset to simulate high resolution (four digit) class I data[49]. HLA ~A, ~B and ~C low resolution 2 digit and 4 digit typing results were combined to simulate a high resolution (4 digit) data set. The combined dataset (2 and 4 digit resolution) had some missing alleles for some participants (data missingness). High resolution HLA class I was simulated from this dataset to address data missingness and the mixed resolution typing nature of the accessed SANBS and NHLS HLA data. Additionally, the current data set was compared to other global populations accessed through the AFND. We note the limitation of not having high resolution data from nations neighboring South Africa for comparison, as previously reviewed[18]. As a result we conveniently selected high resolution class I (four digit) allele frequencies from sub Saharan populations from AFND[50] to compare with our South African data set. Additionally, data generated from this study is accessible, and may be a useful future resource for population and anthropology studies for South African populations.

Ewens-Watterson neutrality test[72] detected excessive heterozygosity ($p<0.0001$) in HLA ~A (high resolution), and HLA ~A and ~DRB1 (low resolution) which is suggestive of balancing selection in these loci (Table 4.2). Balancing selection is well documented to maintain HLA diversity amongst populations[71]. The excessive heterozygosity in South African HLA data described in this study support this previously described source of HLA diversity. Generally, although the Ewens-Watterson neutrality test[72] used to detect neutrality was designed for non recombining data, the test has been evaluated to be insensitive to recombination[73].

This test may be confidently used to detect selection in HLA genes, which are known to have a high recombination rate. Deviations from Ewens-Watterson neutrality due to recombination is expected to decrease haplotype homozygosity[74,75] but not influence balancing selection driven allele diversity. The exact mechanism of how balancing selection promotes HLA diversity is poorly understood[71]. HWE approximation may give insights into HLA genotyping quality and sampling errors. Genotyping errors or failure to detect some alleles (blank allele) increases homozygosity, which may result in significant deviation from HWE[76]. The high data missingness in the current study might explain the highly significant deviations from HWE proportions at both typing resolutions (Table 4.1). Highly significant deviations from HWE might also highlight the presence of family members. Unfortunately we did not access demographic information of the study participants.

We describe allele and haplotype frequencies in the South African population from mixed resolution HLA typing data. All three most frequent alleles (high resolution) were previously reported in different AFND populations at varying frequencies[50]. Interestingly, HLA ~DQA1*05:02 (0.370) was previously reported at lower (0.013) frequency in a South African population ~WOR[51] and in Harare Zimbabwean Shonas (0.004)[50]. Additionally, our third most common allele, HLA ~C*17:01 (0.281), was previously reported at lower frequencies in other South African studies, specifically in South Africa Worcester~WOR[51] (0.053), black South Africans~SoAB[33] (0.111), Caucasian South Africans~SoAC[33] (0.005) and in South African Bone Marrow Registry~SAB[36] (0.028). The second most common allele HLA ~DQA1*04:02 (0.303) has not been previously reported in other South African studies.

Our top three haplotypes were not reported in any population in the allele frequency data base ~AFND[50], which does not necessarily mean the haplotypes have not been reported in any global population. Publicly available HLA data is key in supporting research; hence the need to deposit HLA data into centralised publicly accessible resources. Haplotype frequencies from limited sample size are inherently affected by genetic drift, with the occurrence of some alleles due to chance. The high sample size in the current study might have addressed this problem. We acknowledge though the limitation of mixed resolution typing and data missingness. Other reported confounders to haplotype estimation include typing ambiguity[77] and sample size[20].

Additionally, the highly significant HWE deviations (as seen in this study) have been reported to influence allele and haplotype estimations[78]. There was a strong global LD between loci pairs in our study except for C:DQB1 (p = 0.1061) high resolution, and A:DPB1 (p=0.0290), C:DPB1 (p=0.0671) low resolution (Table 4.6). Haplotype diversity coupled with highly significant LD might generally give insights into purifying selection[79] in HLA region. Global LD considers all possible allele combinations from two loci studied[80], in our case Hedrick's D′[41] weights alleles in each haplotype and Cramer's V Statistic (W_n)[42] is a multi allelic correlation measure between pairs of loci. Haplotype frequency is influenced by LD, sample size, completeness of HLA data and allele frequency[81], especially if gamete phase is unknown (reviewed in[76]).

Although HLA-net gene[RATE) tools are mostly European populations (Northern Africa, Northeast Europe, Southeast Europe, Western Asia, Central and Western Europe)[48], the tool allows for population comparison in HLA diversity through NMDS. Our data was distinctly different from other mostly European population, further supporting high genetic diversity in Africans in general[13-17]. Additionally, our NMDS analysis suggests high genetic diversity in some HLA loci than others, (high resolution HLA ~B, ~DQA1, ~DRB1, ~DQB1 and low resolution HLA ~A, ~B, ~C, ~DRB1, ~DQA1 and ~DQB1 with low diversity in low resolution HLA ~C loci. Generally, in NMDS plots, closely related populations cluster together compared to those that are not related. Tight clusters separated from the rest suggest sub population structure in the dataset. We additionally compared our data with some global populations downloaded from AFND[50] and simulated PhyloD generated data~PSA[49]. Bioinformatics tools have been key in simulating high resolution typing from low/intermediate typing to further understand HLA diversity[49,82]. We acknowledge that the reference for this statistical simulation method[49] might not be ideal for African populations since it is based on African Americans (Table S4.2).

Data from the current study (RSA) was related to other South African data sets (South African studies ~SoAC[33], SoAB[33] and SANT[63], but not with SANZ[64], SAB[36], SAI[52], RMX[53] and WOR[51]) using the Neis' genetic distance[66] and NJ method[69] unrooted tree (Figure 4.2). We expected these populations to cluster together considering they are from the same population. Other South African studies

including South Africa Natal Zulu ~SANZ[64], South African Bone Marrow Registry ~SAB[36], South African Indian ~SAI[52], South African Mixed ancestry ~RMX[53] and South Africa Worcester ~WOR[51] were more related to other sub Saharan populations than our current study (RSA). This might be suggestive of high HLA diversity in South African populations, and their genetic relatedness to other African populations. Despite the use of "African-American" reference in simulating PhyloD generated data~PSA[49], it showed close relation with a previous South African study ~SAB[36] (Figure 4.2). This might give confidence in the simulated data as a future resource for South Africans. Generally, if dendograms generated from HLA data do not show the expected relatedness of populations (geographically, ethnically, anthropologically and linguistically related), it suggests diversification of the studied loci amongst those populations[76]. Genetic distance computation assumes genetic drift drives population differentiation, but there is strong evidence of balancing selection driving differentiation in HLA loci[83-86]. Caution should thus be taken when interpreting HLA genetic distance analysis between populations. Additionally, Neis' genetic distance[66] assumes new alleles arise from neutral mutation rates across all loci. The complex HLA region seems not to follow these assumptions. Other genetic distance measures, Cavalli-Sforza[87] and Reynold's[88] assume no mutation, differences between populations is attributed to genetic drift alone. It seems Neis' genetic distance[66] is favored for HLA data considering the high mutation rates in this gene region.

PCA (Figure 4.3) confirms the genetic relatedness of South Africans (current RSA study) to other sub Saharan populations. Central African Republic Mbenzele Pygmy ~CARMP showed a complete separation from other populations as shown by 69.6% variability in PCA 1 (Figure 4.3) suggesting a unique HLA class I genetic makeup amongst different populations. Additionally, From PCA, there is some degree of confidence in our simulated PhyloD generated data~PSA[49] despite the use of an African –American reference, as it clustered with some South African HLA data and other sub Saharan populations (Figure 4.3).

Generally, HLA allele frequencies provide insight into population history and not necessarily information on selection[89]. HLA data has been widely used to understand genetic relatedness of different populations, and demographic events in those populations[23]. The large sample size of the current study might shed light on some

demographic events in South Africa and how these relate to other sub Saharan populations. Population allele frequencies may be used in disease association studies and provide insight into genetic relatedness[90-92]. They may additionally be used to track population evolutionary processes including migration, selection and admixture[93].

4.6 Conclusions

Despite data missingness, mixed resolution typing and the retrospective nature of the current study, we provide an insight into HLA diversity in South Africans. Our data and simulated PhyloD generated data~PSA[49] may be a useful resource in the future to support disease association and population genetics studies. This attempt to elucidate HLA diversity in South Africans is part of our efforts to fully understand HLA diversity in Africans, and to build a resource for future studies. Key limitations include lack of ethnic data and disease state of participants; these contribute to HLA diversity. Although and individual's inherited HLA genotype does not change due to disease state, continuous exposure to many pathogens in a population result in increased HLA diversity over an evolutionary time[16]. Generally, HLA genetic makeup of populations provides insight into their population history including selective pressures by pathogens[16], migration, admixture and changes in population size[24-27]. Population comparison suggests genetic differences in our population relative to other global populations. It would be interesting to compare more high resolution data from other populations geographically close to South Africa. Unfortunately HLA data from these populations is limited (reviewed in Chapter 2[18]); hence we only managed to include data from Zambia Lusaka (ZaL)[59] and Zimbabwe Harare Shona (ZiHS)[65].

4.7 Data Availability

Previously reported [HLA allele frequencies] data used to support this study are available http://www.allelefrequencies.net/hla6006a.asp using HLA A, B and C search options and sub Sahara region options. These prior studies and other

additional datasets are cited at relevant places within the text as references[27,33,36,50-65]. Data for non-metrical multidimensional scaling (NMDS) analysis is available at https://hala-net.eu/tools/regional-analysis/ and cited in text as gene[RATE] tools[48]. Additionally, HLA allele and haplotype frequencies generated by this study, to support the findings of this study are included within the supplementary information file(s).

4.8 Supplementary Information

Supplementary Tables and Figures are available in Addendum 1 and Addendum 2 respectively. Additionally, these files were submitted together with this manuscript (under review in the BMC Medical Genetics manuscript number MGTC-D-19-00228)

Supplementary Table 4.1 (Table S4.1): Low resolution (two digit) HLA ~A, ~B, ~C, ~DRB1, ~DQA1, ~DQB1 and ~DPB1 allele frequencies in 51 891 typing results and high resolution (four digit) HLA ~A, ~B, ~C, ~DRB1, ~DQA1 and ~DQB1 allele frequencies in 3007 typing result.

Supplementary Table 4.2 (Table S4.2): High resolution (four digit) HLA ~A, ~B, ~C genotypes and allele frequencies from PhyloD generated data~PSA [49]. The data was simulated from our dataset which had a lot of missing data and low resolution typing (two digit)**.**

Supplementary Table 4.3 (Table S4.3): Low resolution (two digit) estimated haplotypes and their frequencies from 51 891 typing results

Supplementary Table 4.4 (Table S4.4): High resolution (four digit) estimated haplotypes and their frequencies from 3007 typing result.

Supplementary Figure 4.1 (Figure S4.1): High resolution (four digit) NMDS global comparison of South African HLA ~A, ~B, ~C, ~DRB1, ~DQA1, and ~DQB1 non metric multidimensional scaling analysis using gene[rate] tools[48].

Supplementary Figure 4.2 (Figure S4.2): Low resolution (two digit) NMDS global comparison South African HLA ~A, ~B, ~C, ~DRB1, ~DQA1, and ~DQB1 non metric multidimensional scaling analysis using gene[rate] tools[48].

4.9 References

1. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L, et al. The DNA sequence and analysis of human chromosome 6. Nature. [Article]. 2003;425:805-11.
2. Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, et al. Deep whole-genome sequencing of 100 southeast Asian Malays. Am J Hum Genet. 2013;92(1):52-66.
3. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43(Database issue):20.
4. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. Nucleic Acids Research. 2013 January 1, 2013;41(D1):D1222-D7.
5. Beatty PG, Boucher KM, Mori M, Milford EL. Probability of finding HLA-mismatched related or unrelated marrow or cord blood donors. Human Immunology. 2000;61(8):834-40.
6. Carrington M, O'Brien S. The influence of HLA genotype on AIDS. Annual Review of Medicine. 2003;54:535-51.
7. Chen H, Ndhlovu ZM, Liu D, Porter LC, Fang JW, Darko S, et al. TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. Nat Immunol. 2012;13(7):691-700.
8. Garamszegi LZ. Global distribution of malaria-resistant MHC-HLA alleles: the number and frequencies of alleles and malaria risk. Malar J. 2014;13(349):1475-2875.
9. Ndung'u T, Gaseitsiwe S, Sepako E, Doualla-Bell F, Peter T, Kim S, et al. Major histocompatibility complex class II (HLA-DRB and -DQB) allele frequencies in Botswana: association with human immunodeficiency virus type 1 infection. Clin Diagn Lab Immunol. 2005;12(9):1020-8.
10. Ramsay M. Africa: continent of genome contrasts with implications for biomedical research and health. FEBS Lett. 2012;586:2813-9.
11. Brander C, Frahm N, Walker BD. The challenges of host and viral diversity in HIV vaccine design (impedes of vaccine development owing to incomplete HLA information). Curr Opin Immunol. 2006;18:430-7.

12. Ovsyannikova I, Poland G. Vaccinomics: Current Findings, Challenges and Novel Approaches for Vaccine Development. AAPS J. 2011 2011/09/01;13(3):438-44.

13. Chen CH, Matthews TJ, McDanal CB, Bolognesi DP, Greenberg ML. A molecular clasp in the human immunodeficiency virus (HIV) type 1 TM protein determines the anti-HIV activity of gp41 derivatives: implication for viral fusion. J Virol. 1995;69(6):3771-7.

14. Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, et al. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet. 2000;66(3):979-88.

15. Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, et al. Nuclear DNA diversity in worldwide distributed human populations. Gene. 1997;205(1-2):161-71.

16. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-Driven Selection and Worldwide HLA Class I Diversity. Current Biology. 2005;15(11):1022-7.

17. Disotell TR. Archaic human genomics. . Am J Phys Anthropol 2012;55:24-39.

18. Tshabalala M, Mellet J, Pepper MS. Human Leukocyte Antigen Diversity: A Southern African Perspective. J Immunol Res. 2015;746151(10):12.

19. Dyer P, McGilvray R, Robertson V, Turner D. Status report from 'double agent HLA': health and disease. Mol Immunol. 2013;55(1):2-7.

20. Gourraud PA, Pappas DJ, Baouz A, Balere ML, Garnier F, Marry E. High-resolution HLA-A, HLA-B, and HLA-DRB1 haplotype frequencies from the French Bone Marrow Donor Registry. Hum Immunol. 2015;76(5):381-4.

21. Edinur HA, Manaf SM, Che Mat NF. Genetic barriers in transplantation medicine. World Journal of Transplantation. 2016;6(3):532-41.

22. WHO. Global Health Report. Geneva2013.

23. Sanchez-Mazas A, Meyer D. The relevance of HLA sequencing in population genetics studies. J Immunol Res. 2014;971818(10):15.

24. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PloS ONE. 2011;6(2):0014643.

25. Sanchez-Mazas A, Fernandez-Vina M, Middleton D, Hollenbach JA, Buhler S, Di D, et al. Immunogenetics as a tool in anthropological studies. Immunology. 2011;133(2):143-64.

26. Parham P, Ohta T. Population biology of antigen presentation by MHC class I molecules. Science. 1996;272(5258):67-74.

27. Kijak GH, Walsh AM, Koehler RN, Moqueet N, Eller LA, Eller M, et al. HLA class I allele and haplotype diversity in Ugandans supports the presence of a major east African genetic cluster. Tissue Antigens. 2009;73(3):262-9.

28. Burrell AS, Disotell TR. Panmixia postponed: ancestry-related assortative mating in contemporary human populations. Genome Biology. 2009;10(11):245-.

29. Meyer D, VR CA, Bitarello BD, DY CB, Nunes K. A genomic perspective on HLA evolution. Immunogenetics. 2018;70(1):5-27.

30. StatisticsSA. www.statssa.gov.za.

31. Berger LR, Hawks J, de Ruiter DJ, Churchill SE, Schmid P, Delezene LK, et al. Homo naledi, a new species of the genus Homo from the Dinaledi Chamber, South Africa. eLife. 2015;4:e09560.

32. Hayhurst JD, du Toit ED, Borrill V, Schlaphoff TEA, Brosnan N, Marsh SGE. Two novel HLA alleles, HLA-A*30:02:01:03 and HLA-C*08:113, identified in a South African bone marrow donor. Tissue Antigens. 2015;85(4):291-3.

33. Paximadis M, Mathebula TY, Gentle NL, Vardas E, Colvin M, Gray CM, et al. Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population. Human Immunol. 2012;73:80-92.

34. May A, Hazelhurst S, Li Y, Norris SA, Govind N, Tikly M, et al. Genetic diversity in black South Africans from Soweto. BMC Genomics. 2013;14(644):1471-2164.

35. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, et al. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nat Commun. 2017;8(1):017-00663.

36. Tshabalala M, Ingram C, Schlaphoff T, Borrill V, Christoffels A, Pepper MS. Human Leukocyte Antigen-A, B, C, DRB1, and DQB1 Allele and Haplotype Frequencies in a Subset of 237 Donors in the South African Bone Marrow Registry. J Immunol Res. 2018;23(2031571).

37. Association WM. World medical association declaration of helsinki: Ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191-4.
38. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. 1995;12(5):921-7.
39. Eberhard HP, Madbouly AS, Gourraud PA, Balere ML, Feldmann U, Gragert L, et al. Comparative validation of computer programs for haplotype frequency estimation from donor registry data. Tissue Antigens. 2013;82(2):93-105.
40. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update – a software pipeline for large-scale multilocus population genomics. Tissue Antigens. 2007;69:192-7.
41. Hedrick PW. Gametic disequilibrium measures: proceed with caution. Genetics. 1987;117(2):331-41.
42. Cramér H. Mathematical Methods of Statistics. Press PU, editor. Princeton: Princeton University Press; 1946.
43. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 2010;10(3):564-7.
44. Coombs JA, Letcher BH, Nislow KH. CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. Mol Ecol Resour. 2008;8(3):578-80.
45. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics. 1992;48(2):361-72.
46. Slatkin M. An exact test for neutrality based on the Ewens sampling distribution. Genetical Research. 1994;64.
47. Slatkin M. A correction to the exact test based on the Ewens sampling distribution. Genetical Research. 1996;68.
48. Nunes JM. Using UNIFORMAT and GENE[RATE] to Analyze Data with Ambiguities in Population Genetics. Evolutionary Bioinformatics. 2016(5411):19-26.
49. Listgarten J, Brumme Z, Kadie C, Xiaojiang G, Walker B, Carrington M, et al. Statistical resolution of ambiguous HLA typing data. PLoS Comput Biol. 2008;4(2):1000016.
50. González-Galarza Faviel F, Takeshita Louise YC, Santos Eduardo JM, Kempson F, Maia Maria Helena T, Silva Andrea Luciana Soares d, et al. Allele

frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Research. 2015;43(D1):D784-D8.

51. Grifoni A, Sidney J, Carpenter C, Phillips E, Mallal S, Scriba TJ, et al. Sequence-based HLA-A, B, C, DP, DQ, and DR typing of 159 individuals from the Worcester region of the Western Cape province of South Africa. Hum Immunol. 2018;79(3):143-4.

52. Loubser S, Paximadis M, Tiemessen CT. Human leukocyte antigen class I (A, B and C) allele and haplotype variation in a South African Indian population. Hum Immunol. 2017;78(7-8):468-70.

53. Loubser S, Paximadis M, Tiemessen CT. Human leukocyte antigen class I (A, B and C) allele and haplotype variation in a South African Mixed ancestry population. Hum Immunol. 2017;78(5-6):399-400.

54. Modiano D, Luoni G, Petrarca V, Sodiomon Sirima B, De Luca M, Simpore J, et al. HLA class I in three West African ethnic groups: genetic distances from sub-Saharan and Caucasoid populations. Tissue Antigens. 2001;57(2):128-37.

55. Torimiro JN, Carr JK, Wolfe ND, Karacki P, Martin MP, Gao X, et al. HLA class I diversity among rural rainforest inhabitants in Cameroon: identification of A*2612-B*4407 haplotype. Tissue Antigens. 2006;67(1):30-7.

56. Bruges Armas J, Destro-Bisol G, Lopez-Vazquez A, Couto AR, Spedini G, Gonzalez S, et al. HLA class I variation in the West African Pygmies and their genetic relationship with other African populations. Tissue Antigens. 2003;62(3):233-42.

57. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Guethlein LA, Hilton HG, Pando MJ, et al. Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. PLoS Genet. 2013;9(10):31.

58. Luo M, Embree J, Ramdahin S, Ndinya-Achola J, Njenga S, Bwayo JB, et al. HLA-A and HLA-B in Kenya, Africa: Allele frequencies and identification of HLA-B*1567 and HLA-B*4426. Tissue Antigens. 2002;59(5):370-80.

59. Cao K, Moormann AM, Lyke KE, Masaberg C, Sumba OP, Doumbo OK, et al. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens. 2004;63(4):293-325.

60. Arlehamn CSL, Copin R, Leary S, Mack SJ, Phillips E, Mallal S, et al. Sequence-based HLA-A, B, C, DP, DQ, and DR typing of 100 Luo infants from the Boro area of Nyanza Province, Kenya. Human Immunology. 2017;78(4):325-6.

61. Tang J, Naik E, Costello C, Karita E, Rivers C, Allen S, et al. Characteristics of HLA class I and class II polymorphisms in Rwandan women. Exp Clin Immunogenet. 2000;17(4):185-98.

62. Sanchez-Mazas A, Steiner QG, Grundschober C, Tiercy JM. The molecular determination of HLA-Cw alleles in the Mandenka (West Africa) reveals a close genetic relationship between Africans and Europeans. Tissue Antigens. 2000;56(4):303-12.

63. Hammond MG, Anley D, editors. Tamil from Natal Province, South Africa. Proceedings of the 13th International Histocompatibility Workshop; 2006; Seattle: International Histocompatibility Working Group Press.

64. Hammond MG, Middleton D, Anley D, editors. Zulu from Natal Province, South Africa. Proceedings of the 13th International Histocompatibility Workshop and Conference; 2006; Seattle, WA: IHWG Press.

65. Louie L, Mather K, Meyer D, Hollenbach J, Jackman R, Schultz K, et al., editors. Shona from Harare, Zimbabwe. Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference; 2006; Seattle, WA: IHWG Press, 2007.

66. Nei M. Genetic Distance between Populations. The American Naturalist. 1972;106(949):283-92.

67. Takezaki N, Nei M, Tamura K. POPTREE2: Software for Constructing Population Trees from Allele Frequency Data and Computing Other Population Statistics with Windows Interface. Molecular Biology and Evolution. 2010;27(4):747-52.

68. Takezaki N, Nei M, Tamura K. POPTREEW: Web Version of POPTREE for Constructing Population Trees from Allele Frequency Data and Computing Some Other Quantities. Molecular Biology and Evolution. 2014;31(6):1622-4.

69. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406-25.

70. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Research. 2015;43(W1):W566-W70.

71. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nature Genetics. 2008;40:340.

72. Watterson GA. Homozygosity test of neutrality. Genetics. 1978;88.

73. Zeng K, Mano S, Shi S, Wu CI. Comparisons of site- and haplotype-frequency methods for detecting positive selection. Mol Biol Evol. 2007;24(7):1562-74.

74. Wright SI, Foxe JP, DeRose-Wilson L, Kawabe A, Looseley M, Gaut BS, et al. Testing for effects of recombination rate on nucleotide diversity in natural populations of Arabidopsis lyrata. Genetics. 2006;174(3):1421-30.

75. Sanchez-Mazas A, Lemaitre JF, Currat M. Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. Philos Trans R Soc Lond B Biol Sci. 2012;367(1590):830-9.

76. Mack SJ, Gourraud P-A, Single RM, Thomson G, Hollenbach JA. Analytical Methods for Immunogenetic Population Data. Methods in Molecular Biology (Clifton, NJ). 2012;882:215-44.

77. Castelli EC, Mendes-Junior CT, Veiga-Castelli LC, Pereira NF, Petzl-Erler ML, Donadi EA. Evaluation of computational methods for the reconstruction of HLA haplotypes. Tissue Antigens. 2010;76(6):459-66.

78. Single RM, Meyer D, Hollenbach JA, Nelson MP, Noble JA, Erlich HA, et al. Haplotype frequency estimation in patient populations: the effect of departures from Hardy-Weinberg proportions and collapsing over a locus in the HLA region. Genet Epidemiol. 2002;22(2):186-95.

79. Alter I, Gragert L, Fingerson S, Maiers M, Louzoun Y. HLA class I haplotype diversity is consistent with selection for frequent existing haplotypes. PLOS Computational Biology. 2017;13(8):e1005693.

80. Klitz W, Stephens JC, Grote M, Carrington M. Discordant patterns of linkage disequilibrium of the peptide-transporter loci within the HLA class II region. Am J Hum Genet. 1995;57(6):1436-44.

81. Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics. 1964;49(1):49-67.

82. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. Human Immunology. 2013;74(10):1313-20.

83. Hedrick PW, Thomson G. Evidence for Balancing Selection at Hla. Genetics. 1983;104(3):449-56.

84. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335(6186):167-70.

85. Meyer D, Thomson G. How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet. 2001;65(Pt 1):1-26.

86. Lawlor DA, Ward FE, Ennis PD, Jackson AP, Parham P. HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. Nature. 1988;335(6187):268-71.

87. Cavalli-Sforza LL, Edwards AW. Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet. 1967;19(3 Pt 1):233-57.

88. Reynolds J, Weir BS, Cockerham CC. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics. 1983;105(3):767-79.

89. Blagitko N, O'HUigin C, Figueroa F, Horai S, Sonoda S, Tajima K, et al. Polymorphism of the HLA-DRB1 locus in Colombian, Ecuadorian, and Chilean Amerinds. Hum Immunol. 1997;54(1):74-81.

90. Sanchez-Mazas A, Vidan-Jeras B, Nunes JM, Fischer G, Little AM, Bekmane U, et al. Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA-NET methodological recommendations. International Journal of Immunogenetics. 2012;39(6):459-76.

91. Mack SJ, Tu B, Lazaro A, Yang R, Lancaster AK, Cao K, et al. HLA-A, -B, -C, and -DRB1 allele and haplotype frequencies distinguish Eastern European Americans from the general European American population. Tissue Antigens. 2009;73(1):17-32.

92. Romphruk AV, Romphruk A, Kongmaroeng C, Klumkrathok K, Paupairoj C, Leelayuwat C. HLA class I and II alleles and haplotypes in ethnic Northeast Thais. Tissue Antigens. 2010;75(6):701-11.

93. Fernandez Vina MA, Hollenbach JA, Lyke KE, Sztein MB, Maiers M, Klitz W, et al. Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. Philos Trans R Soc Lond B Biol Sci. 2012;367(1590):820-9.

CHAPTER 5

In silico HLA typing of 24 whole genome sequences generated by the Southern African Human Genome Programme (SAHGP)

5.1 Abstract

Background: Despite the importance of human leukocyte antigen (HLA) typing results in research and clinical applications, HLA typing is still generally inaccessible in most resource limited settings. There is however an increasing number of next generation sequencing studies generating sequence data that may be used to determine HLA alleles *in silico.* This chapter describes determination of HLA alleles from 24 whole genomes from South African individuals using *in silico* methods to augment the paucity of HLA diversity data in these populations.

Methods: Ethical approval was granted by University of Pretoria and the Southern African Human Genome Program (SAHGP) ethics committees. Whole genome sequence data was used to determine HLA alleles by HLAscan and HLA-HD imputation tools.

Results: The two *in silico* HLA imputation methods predicted high resolution (up to 8 digits) HLA alleles from the 24 South African genomes. Classical, non-classical and non-HLA alleles were predicted by the two methods using the whole genome sequences. There was generally high concordance between the two methods in predicting classical class I alleles compared to classical class II alleles.

Conclusions/Significance: This chapter demonstrates the feasibility of using whole genome sequence data in understanding HLA diversity, especially in populations with limited HLA typing data. With the increasing availability of human genomic data at the population level through improvements in NGS and reduction of sequencing costs, HLA imputation might augment HLA typing. Results from this study benchmark the use of sequencing data to support HLA disease association studies, population genetics and better inform donor recruitment strategies into registries epidemiology

5.2 Introduction

Precise HLA typing at high resolution has an impact on clinical outcomes in transplantation[1,2] highlighting the critical need for accurate high resolution HLA typing methods. The polymorphic nature of the HLA gene region makes high resolution HLA typing challenging. It is often difficult to accurately determine an individual's HLA genotype at high resolution. The HLA gene region is considered to be one of the most polymorphic regions in the human genome[3,4], with 20 088 alleles described in the IMGT HLA database version 3.34.0 of October 2018 (https://www.ebi.ac.uk/ipd/imgt/hla/stats.html)[5]. Additionally, high linkage disequilibrium (LD) is a distinctive feature of the HLA region[6,7], adding to the challenge of HLA typing. Generally, classical HLA typing is commonly performed by sequencing exons 2–4 of Class I genes (HLA ~A, ~B and ~C) and exons 2 and/or 3 of Class II genes (HLA ~DRB1 and ~DQB1)[1].. But next generation sequencing (NGS) has revolutionized HLA typing with whole class I genes being sequenced and more exons being sequenced for class II alleles[8]. Despite these improvements, NGS HLA typing remains relatively expensive and generally inaccessible to most developing countries' public health systems, e.g. South Africa. As a result, few individuals (in relation to population size) are HLA typed at high resolution for clinical applications. This contributes to the limited availability of high resolution HLA data from these populations (reviewed in this thesis Chapter 2[9]). Additionally, short and long read sequences generated by NGS HLA typing have challenges including read coverage of target HLA gene/gene region, chromosome phasing and reduced ability to identify novel alleles.

Despite the key function of HLA in host immunity and association with several diseases[10], HLA typing is not routinely done in many settings due to high costs and expertise needed. With the current global push towards precision medicine, it becomes critical to have HLA genotypes at high resolution for better diagnosis and management. At least four digit typing (amino acid level) is clinically relevant to reduce graft versus host disease (GVHD), and reduce the chance of graft rejection[11]. There are several HLA typing methods, from serology, polymerase chain reaction sequence specific primer (PCR-SSP); polymerase chain reaction sequence specific oligonucleotide (PCR-SSO) Sanger sequence based typing and NGS HLA

typing[12,13]. HLA imputation provides a low cost broadly available HLA typing method owing to advances in NGS and availability of large numbers of whole genome sequence (WGS), whole exome sequence (WES) and single nucleotide polymorphisms (SNP) data sets across many populations. Single nucleotide polymorphisms, WGS and WES data sets may be used to accurately determine high resolution HLA alleles of the sequenced individuals[14-21]. Even discovery of novel alleles using *in silico* methods is possible[18] through in *silico* HLA typing (HLA imputation). Several large sequencing projects like 1000 Genomes[22-25], African Genome Variome Project[26], H3 Africa (https://h3africa.org/) and the Southern African Human Genome Programme (SAHGP) datasets are valuable resources for HLA imputation to better understand HLA diversity in African populations.

The SAHGP is a South African government funded initiative aimed at understanding genetic diversity of southern Africans, and was officially launched in January 2011[27]. The pilot study describes genetic diversity in 24 South African male individuals (8 South African colored and 16 black South Africans from the eastern Bantu speaking lineage) using WGS. The study highlights high genetic diversity amongst the 24 whole genomes. Additionally, the study showed genetic variability amongst the eastern Bantu speakers suggesting more extensive genetic diversity than previously thought[28]. Generally, African populations are considered genetically diverse[29-33] with a high disease burden[34], and they are believed to be the cradle of modern humans[35,36]. The South African ethnolingistic diversity comprises the following groups: 79.6% eastern Bantu speakers, 8/9% Coloured (mixed race), 8.9% whites, 2.5% Indian and 0.1% unclassified (http://www.statssa.gov.za/). The SAHGP pilot project generated a bioresource of unbiased deep sequencing data from the South African genomes. The study data analysis was done by South Africans supported by government funding as an initiative to build capacity, and demonstrates political will in understanding human genetic diversity[27,28].

This study aimed at determining HLA alleles from 24 whole genome sequences generated in the SAHGP[27,28] using *in silico* methods as a pilot in using HLA imputation to understanding HLA diversity in South Africans.

5.3 Materials and Methods

5.3.1 Ethics and data Access

Ethical approval and access to the data was granted through the University of Pretoria Faculty of Health Sciences Ethics committee (ref: 220/2015) and the SAHGP data access committee (ref: SAHGP004) with all participants in the SAHGP study having signed written informed consent to participate in the main study[28]. The European Genome-phenome archive (EGA) client tool was used to download sequence data of the 24 individuals (accession number EGAD00001003791) in BAM file format[37] from the EGA (https://ega-archive.org/datasets/EGAD00001003791). Briefly, the tool offers a secure download of the data (password protected and encrypted data is downloaded after ethical approval). The commands used to download sequence data from EGA are summarised in Appendix 7 and detailed in (https://www.ebi.ac.uk/ega/about/your_EGA_account/download_streaming_client).

5.3.2 Description of data and file pre processing

The SAHGP data was sequenced at about 50X coverage (≥ 30X) on Illumina HiSeq2000 (~100bp paired end reads, ~314bp insert size)[28]. Sequence reads were aligned to NCBI37 (hg19) human reference genome using Isaac Alignment tool[38]. Quality of alignments was determined by Samtools ver 1.1-26[37]. For the current study, reads covering chromosome 6 (chr6:28866528-33775446) were extracted using Samtools ver 1.1-26[37]. The chr6:28866528-33775446 covers and overlaps the HLA region; hence all HLA sequence reads were extracted. SamToFastq tool in picard-2.17.11 tools (https://github.com/broadinstitute/picard) was used to convert SAM files to paired end fastq files[39]. The extracted chromosome 6 (chr6:28866528-33775446) fastq files[39] were used as input for HLA imputation. In Appendix 8 is a customized python script used to automate chromosome 6 (chr6:28866528-33775446) read extraction and conversion from BAM file format[37] to paired end fastq file formats[39].

5.3.3 HLA imputation using HLA scan and HLA-HD tools

Two alignment based HLA imputation tools were independently used to determine HLA alleles of the 24 whole genomes generated by the SAHGP[27,28]. HLA scan[40] and HLA typing from High-quality Dictionary (HLA-HD)[41,42] alignment based tools were used for HLA imputation. HLAscan[40] and HLA-HD[41,42] tools were downloaded onto a local University of Pretoria Unix server together with dependencies outlined by the developers. The environment variables for these imputation tools were set to run in the folders with the SAHGP BAM file[37] and paired end fastq file[39] file formats.

Figure 5.1 summarises the step by step imputation using these two methods to obtain high resolution HLA typing results. For both methods, the IMGT HLA database version 3.34.0 of October 2018 (https://www.ebi.ac.uk/ipd/imgt/hla/stats.html)[5] was used as a reference. Briefly, HLA scan[40] is an alignment-based program that determines HLA alleles taking into account sequence read coverage to reduce false allele calling. The software performs alignment of reads to HLA sequences from the international ImMunoGeneTics project/human leukocyte antigen (IMGT/HLA) database (https://www.ebi.ac.uk/ipd/imgt/hla/)[5,43]. The distribution of the HLA region aligned reads is used to calculate a score function and to determine correctly phased alleles by progressively removing false-positive alleles. HLAscan can be reliably applied for determination of HLA type across the whole-genome, exome, and target sequences. HLAscan software is a freely available public tool for academic purposes, and requires a license for commercial HLA typing[40]. Default settings were used to determine HLA alleles from the 24 genomes.

On the other hand, HLA-HD[41,42] is also a freely available tool for academic use, to accurately determine HLA alleles from NGS data (fastq format). Additionally, HLA-HD[41,42] may use RNA-Seq data for HLA imputation. The tool by default ignores any reads less than 100 base pairs (bp), and considers HLA exonic and intronic read coverage. The tool firstly generates an HLA library of HLA genes in the IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla/docs/release.html) using the latest release (3.34.0 of October 2018)[5]. The number of reads mapped to the HLA

dictionary determines the weighting score of a potential allele at that locus. Default settings were modified in the HLA dictionary to type additional HLA ~DRB5, HLA ~T, ~W and ~Y as per HLA_gene.split.3.32.0.txt in the version 1.2.0.1 July 11, 2018 release[41,42])

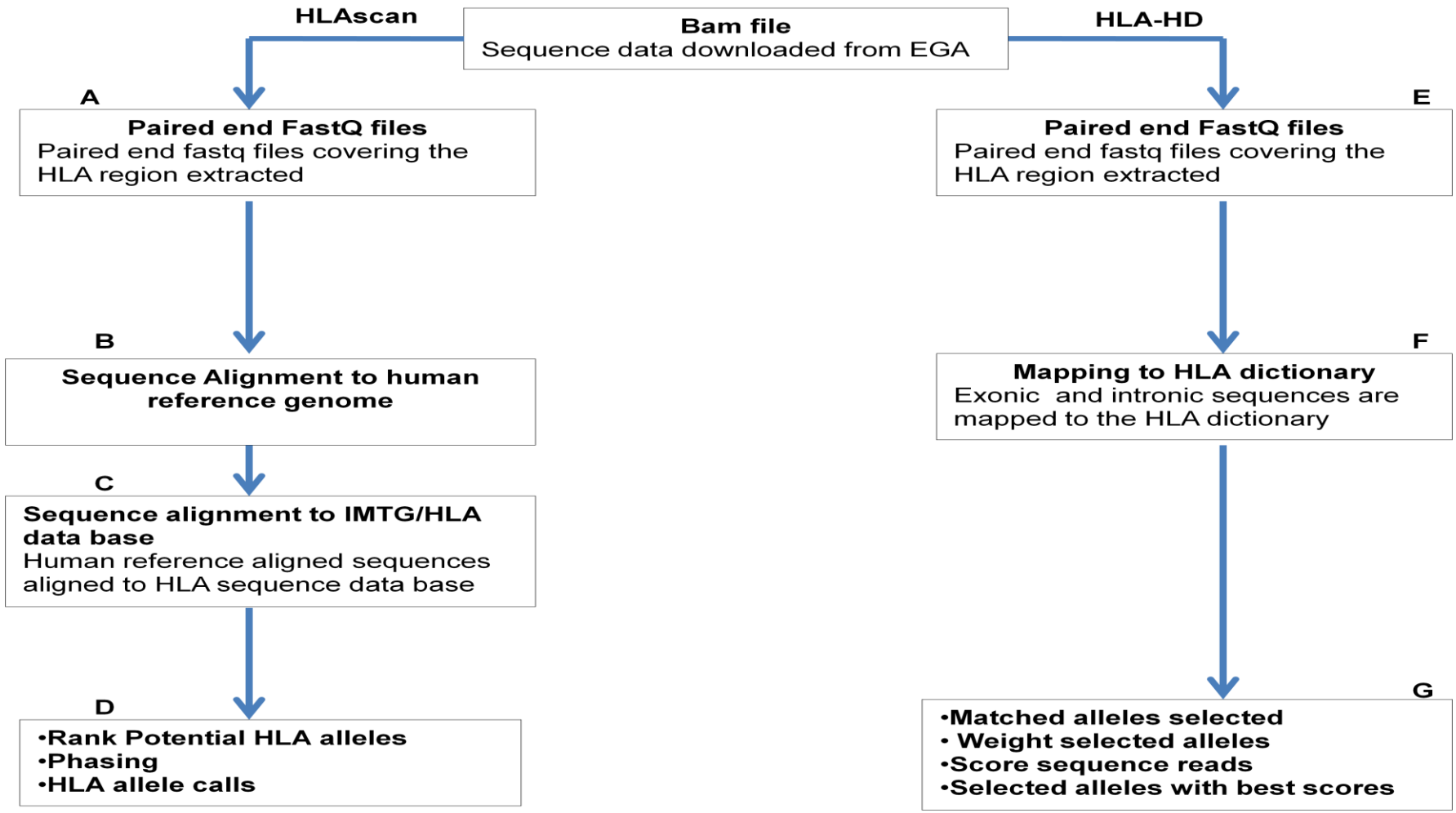

Figure 5.1 In silico HLA typing using HLA scan and HLA –HD tools

HLAscan imputation. Sections A-D summarise the steps. Briefly: paired end fastq files covering the HLA region are used as input for imputation (A). The sequence reads are aligned to the human reference genome sequence (B) and HLA allele sequences in the IMGT HLA data base (C). Potential alleles are scored based on read coverage, and resolving phasing issues (D). HLA-HD imputation (E-G), briefly: paired end fastq sequence reads (E) are mapped onto exons and introns of all the alleles recorded in the HLA dictionary (F). Matched reads are assigned to HLA alleles for the allele pair score calculation, with each read being weighted. The score of the weighted sum of reads is calculated for potential allele pairs, and the pair yielding the highest score is selected (G). This flow diagram was adapted from HLA scan and HLA-HD methods[40-42].

5.3.4 Assessing concordance of Imputation tools

We describe concordance as the total number of similar alleles called by the two methods per locus X [typing resolution is taken into account, for example A*29:02:01by HLA HD and A*29:02:01:02 by HLAscan are considered similar as one method gives a higher resolution of the same allele (Table 5.1)]. The X is divided by the total number of alleles/loci *Y* (excluding ambiguous typing the default is 4, which is two alleles per imputation method). Therefore concordance is given by:

$$\frac{X}{Y} * 100\%$$

5.4 Results

The two HLA imputation tools successfully determined classical (HLA class 1 and class II) and non classical (HLA class III) HLA alleles from whole genome sequences of 24 individuals. Supplementary Tables 5.1 (S5.1) and 5.2 (S5.2) summarize the imputed alleles using HLA-HD and HLAscan methods[40-42] respectively. Generally, HLA-HD[41,42] determined HLA alleles in 28 loci (S5.1) while HLAscan[40] used 17 HLA loci plus 4 non HLA loci (S5.2). The highest HLA typing resolution from HLA-HD[41,42] was 6 digits, for example the genotype of individual 1 (Table 5.1) is HLA ~ A*24:02:01/A*25:01:01. On the other hand, HLAscan[40] gave up to 8 digit typing resolution (for example individual 12 in Table 5.1 is HLA-A*30:02:01:03/A*68:01:01:01). Tables 5.1 to 5.6 summarise classical HLA typing results generated by the two *in silico* methods[40-42].

HLA ~B (Table 5.2) and ~C (Table 5.3) loci had the highest concordance between the two HLA imputation methods[40-42], with 100% concordance in 21/24 individuals for both loci. Additionally, 100% concordance for HLA ~DRB1 in 19/24 individuals (Table 5.4), for HLA ~A in 15/24 individuals (Table 5.1), for HLA ~DAQ1 in 11/24 individuals (Table 5.5) and for HLA ~DQB1 in 10/24 individuals (Table 5.6). Zero (0%) concordance between the two methods used[40-42] was observed for HLA ~A in 4/24 individuals (Table 5.1), for HLA ~B in 1/24 individuals (Table 5.2), HLA ~DQA1 in 5/24 individuals (Table 5.5) and for HLA ~DQB1 in 5/24 individuals (Table 5.6). No

concordance (0%) was observed for 4/24 individuals (HLA ~A Table 5.1), 1/24 individuals (HLA ~B Table 5.2), 4/24 individuals (HLA ~DQA1 Table 5.5) and 5/24 individuals (HLA ~DQB1 Table 5.6). There was generally higher concordance in class I alleles (HLA ~A, ~B and ~C) compared to class II alleles (HLA ~DRB1, ~DQA1 and ~DQB1). In some cases one method gave a higher resolution of the same allele e.g. A*29:02:01 for HLA-HD[41,42] and A*29:02:01:02 for HLAscan (Table 5.1).

HLAscan[40] could not determine HLA ~DRB5 alleles in some (18/24) individuals (S5.2). On the other hand HLA-HD[41,42] gave ambiguous typing results in HLA ~DOB, ~DRB4, ~H and ~K loci in some individuals (Table 5.7). No ambiguous typing was obtained for *in silico* classical HLA alleles (Tables S5.1 and S5.2), but in some cases imputation methods[40-42] could not determine HLA alleles (Tables 5.2, 5.4, 5.5 and 5.6). Unfortunately, the 24 individuals in this study were not HLA typed experimentally or for any medical reasons; hence we could not compare the *in silico* determined HLA alleles to HLA typing results. The two HLA imputation tools used HLA-HD and HLAscan methods[40-42] in this study were evaluated on public datasets including the 1000 Genomes[22-25] with 100% accuracy. Imputation results described in this study highlights the feasibility of leveraging from existing sequence data from African populations to better understand HLA diversity in these populations.

Table 5.1 *In silico* HLA –A determination using HLA scan and HLA-HD tools

Sample ID	HLA $A_{HLA\text{-}HD}$[41,42]		HLA $A_{HLA\text{-}SCAN}$[40]		%
1	A*24:02:01	A*25:01:01	A*25:01:01	A*24:02:01:03	100
2	A*23:17:01	A*30:04:01	A*30:04:01	A*23:01:01	50
3	A*02:05:01	A*02:603	A*02:14	A*02:02:01	0
4	A*32:01:01	A*30:04:01	A*30:04:01	A*32:01:01	100
5	A*30:01:01	A*03:01:01	A*30:01:01	A*03:01:01:03	100
6	A*29:02:01	A*30:02:01	A*29:02:01:01	A*30:02:01:02	100
7	A*30:02:01	A*02:01:01	A*30:02:01:02	A*02:01:01:02L	100
8	A*29:02:01	A*23:17:01	A*29:02:01:02	A*23:01:01	50
9	A*02:01:01	A*30:18	A*02:09	A*30:01:01	0
10	A*33:03:01	A*34:01:01	A*33:03:01	A*34:01:01	100
11	A*43:01	A*02:05:01	A*43:01	A*02:05:01	100
12	A*68:01:01	A*30:02:01	A*30:02:01:03	A*68:01:01:01	100
13	A*03:01:01	A*74:01:01	A*03:01:01:03	A*74:02:01:02	100
14	A*23:01:01	A*02:02:01	A*02:02:01	A*23:01:01	100
15	A*02:01:18	A*01:01:01	A*02:01:15	A*01:01:01:01	50
16	A*24:02:01	A*25:01:01	A*24:02:01:03	A*25:01:01	100
17	A*23:17:01	A*02:01:01	A*23:01:01	A*02:01:01:02L	50
18	A*68:02:01	A*66:01:01	A*66:01:01	A*68:02:01:03	100
19	A*26:01:01	A*29:01:01	A*26:01:01:01	A*29:01:01:02N	100
20	A*68:02:02	A*66:03:01	A*68:02:01:03	A*66:02	0
21	A*29:02:01	A*26:01:01	A*29:02:01:02	A*26:01:07	50
22	A*01:01:01	A*11:01:01	A*01:04N	A*11:01:47	0
23	A*68:02:01	A*03:01:01	A*68:02:01:02	A*03:01:01:03	100
24	A*02:05:01	A*30:02:01	A*02:05:01	A*30:02:01:03	100

percentage (%) concordance between the two methods. The difference in typing resolution of the same allele is ignored (the two methods are considered concordant in predicting that allele)

Table 5.2 *In silico* HLA –B determination using HLA scan and HLA-HD tools

Sample ID	HLA $B_{HLA\text{-}HD}$[41,42]		HLA $B_{HLA\text{-}SCAN}$[40]		%
1	B*07:02:01	B*37:01:01	B*07:02:01	B*37:01:01	100
2	B*58:02:01	B*44:03:01	B*58:02	B*44:03:01	100
3	B*44:03:01	B*57:03:01	B*44:03:01	B*57:03:01	100
4	B*15:01:01	-	B*15:01:01:03	B*15:01:01:03	50
5	B*42:02:01	B*44:03:02	B*42:02:01:02	B*44:03:02	100
6	B*42:01:01	B*15:03:01	B*42:01:01	B*15:03:01	100
7	B*08:01:01	B*40:01:02	B*40:01:01	B*08:01:01	50
8	B*44:37:02	B*58:07	B*44:03:02	B*58:02	0
9	B*81:01:01	B*45:01:01	B*45:01:01	B*81:01	100
10	B*15:21:01	B*44:03:02	B*44:03:02	B*15:02:01	50
11	B*15:10:01	B*44:03:01	B*15:10:01	B*44:03:01	100
12	B*07:02:01	B*14:02:01	B*07:02:01	B*14:02:01	100
13	B*18:01:01	B*57:03:01	B*18:01:01:02	B*57:03:01	100
14	B*15:10:01	B*08:01:01	B*15:10:01	B*08:01:01	100
15	B*81:01:01	B*45:01:01	B*81:01	B*45:01:01	100
16	B*55:01:01	B*18:01:01	B*55:01:01	B*18:01:01:01	100
17	B*07:02:01	B*44:03:01	B*44:03:01	B*07:02:01	100
18	B*15:10:01	B*58:02:01	B*15:10:01	B*58:02	100
19	B*41:01:01	B*18:01:01	B*18:01:01:01	B*41:01:01	100
20	B*15:03:01	B*53:01:01	B*53:01:01	B*15:03:01	100
21	B*44:03:02	B*51:01:01	B*44:03:02	B*51:01:01:02	100
22	B*35:03:01	B*37:01:01	B*35:03:01	B*37:01:01	100
23	B*58:02:01	B*18:01:01	B*58:02	B*18:01:01:02	100
24	B*08:01:01	B*50:01:01	B*50:01:01	B*08:01:01	100

percentage (%) concordance between the two methods. The difference in typing resolution of the same allele is ignored (the two methods are considered concordant in predicting that allele) '-:' Tool could not determine the HLA allele

Table 5.3 *In silico* HLA –C determination using HLA scan and HLA-HD tools

Sample ID	HLA $C_{HLA\text{-}HD}$[41,42]		HLA $C_{HLA\text{-}SCAN}$[40]		%
1	C*07:02:01	C*06:02:01	C*06:02:01:02	C*07:02:01:01	100
2	C*04:01:01	C*06:02:01	C*04:01:01:06	C*06:02:01:03	100
3	C*07:01:02	C*04:01:01	C*04:01:01:02	C*07:01:02	100
4	C*04:01:01	C*03:03:01	C*03:03:01	C*04:01:01:01	100
5	C*07:06:01	C*17:01:01	C*17:01:01:02	C*07:06	100
6	C*02:10:01	C*17:01:01	C*17:03	C*02:10	50
7	C*07:01:01	C*03:04:01	C*03:04:43	C*07:01:01:02	50
8	C*06:02:01	C*07:06:01	C*07:06	C*06:02:01:02	100
9	C*04:01:01	C*16:01:01	C*04:01:01:02	C*16:01:01	100
10	C*04:03:01	C*07:06:01	C*07:06	C*04:03:01	100
11	C*08:04:01	C*02:10:01	C*08:04:01	C*02:10	100
12	C*07:02:01	C*08:02:01	C*08:02:01:02	C*07:02:01:03	100
13	C*18:02	C*07:01:01	C*07:01:01:03	C*18:02	100
14	C*07:01:01	C*16:01:01	C*07:01:01:03	C*16:01:01	100
15	C*16:01:01	C*18:01	C*16:01:01	C*18:01	100
16	C*12:03:01	C*03:03:01	C*12:03:01:01	C*03:03:01	100
17	C*07:02:01	C*02:10:01	C*02:10	C*07:02:01:03	100
18	C*03:04:02	C*06:02:01	C*06:02:01:01	C*03:04:02	100
19	C*07:04:01	C*17:01:01	C*07:04:01	C*17:03	50
20	C*02:10:01	C*04:01:01	C*04:01:01:04	C*02:10	100
21	C*07:01:01	C*07:06:01	C*07:01:01:03	C*07:06	100
22	C*06:02:01	C*04:01:01	C*06:02:01:01	C*04:01:01:06	100
23	C*06:02:01	C*05:01:01	C*05:01:01:01	C*06:02:01:01	100
24	C*07:01:01	C*06:02:01	C*06:02:01:03	C*07:01:01:03	100

% percentage concordance between the two methods. The difference in typing resolution of the same allele is ignored (the two methods are considered concordant in predicting that allele)

Table 5.4 *In silico* HLA –DRB1 determination using HLA scan and HLA-HD tools

Sample ID	HLA DRB1$_{HLA-HD}$[41,42]		HLA DRB1$_{HLA-SCAN}$[40]		%
1	DRB1*15:01:01	DRB1*10:01:01	DRB1*15:01:01:02	DRB1*10:01:01	100
2	DRB1*03:01:01	DRB1*04:04:01	DRB1*04:04:01	DRB1*03:01:01:02	100
3	DRB1*01:02:13	DRB1*03:02:01	DRB1*03:02:01	DRB1*01:02:01	50
4	DRB1*01:03:01	DRB1*04:01:01	DRB1*04:01:01	DRB1*01:03	100
5	DRB1*11:01:02	DRB1*13:02:01	DRB1*13:02:01	DRB1*11:01:02	100
6	DRB1*13:02:01	DRB1*08:04:01	DRB1*13:02:01	DRB1*08:04:01	100
7	DRB1*03:02:01	DRB1*04:01:01	DRB1*04:01:01	DRB1*03:02:01	100
8	DRB1*04:04:01	DRB1*11:01:02	DRB1*11:01:02	DRB1*11:01:02	50
9	DRB1*03:02:01	DRB1*15:03:01	DRB1*03:02:01	DRB1*15:03:01:01	100
10	DRB1*15:02:01	DRB1*07:01:01	DRB1*15:02:01	DRB1*07:01:01:01	100
11	DRB1*13:01:01	DRB1*04:01:01	DRB1*13:01:01	DRB1*04:01:01	100
12	DRB1*15:03:01	DRB1*09:01:02	DRB1*15:03:01:01	DRB1*09:01:02	100
13	DRB1*13:01:01	DRB1*11:04:01	DRB1*11:04:01	DRB1*13:01:01	100
14	DRB1*11:01:02	DRB1*03:01:01	DRB1*11:01:02	DRB1*03:01:01:01	100
15	DRB1*11:01:02	-	DRB1*11:01:02	DRB1*11:01:02	50
16	DRB1*14:54:01	DRB1*07:01:01	DRB1*07:01:01:02	DRB1*14:10	50
17	DRB1*15:01:01	DRB1*04:01:01	DRB1*15:01:01:02	DRB1*04:01:01	100
18	DRB1*07:01:01	DRB1*10:01:01	DRB1*10:01:01	DRB1*07:01:01:02	100
19	DRB1*13:01:01	DRB1*13:02:01	DRB1*13:02:01	DRB1*13:01:01	100
20	DRB1*07:01:01	DRB1*01:02:13	DRB1*01:02:01	DRB1*07:01:01:01	50
21	DRB1*11:01:02	DRB1*15:01:01	DRB1*11:01:02	DRB1*15:01:01:04	100
22	DRB1*13:01:01	DRB1*01:01:01	DRB1*13:01:01	DRB1*01:01:01	100
23	DRB1*08:04:01	DRB1*11:01:02	DRB1*08:04:01	DRB1*11:01:02	100

24	DRB1*03:01:01	DRB1*07:01:01	DRB1*03:01:01:02	DRB1*07:01:01:01	100

% percentage concordance between the two methods. The difference in typing resolution of the same allele is ignored (the two methods are considered concordant in predicting that allele). '-:' Tool could not determine the HLA allele

Table 5.5 *In silico* HLA –DQA1 determination using HLA scan and HLA-HD tools

Sample ID	HLA DQA1$_{HLA-HD}$[41,42]		HLA DQA1$_{HLA-SCAN}$[40]		%
1	DQA1*01:02:01	DQA1*01:12	DQA1*01:02:01:04	DQA1*01:05:02	50
2	DQA1*05:05:01	DQA1*03:03:01	DQA1*05:09	DQA1*05:09	0
3	DQA1*04:01:01	DQA1*01:01:02	DQA1*01:01:02	DQA1*04:01:01	100
4	DQA1*01:01:01	DQA1*03:03:01	DQA1*01:01:02	DQA1*01:04:01:02	0
5	DQA1*01:02:01	-	DQA1*01:02:01:04	DQA1*01:02:01:04	50
6	DQA1*01:02:01	DQA1*05:05:01	DQA1*01:02:01:04	DQA1*05:05:01:02	100
7	DQA1*04:01:01	DQA1*03:03:01	DQA1*04:01:01	DQA1*04:01:01	50
8	DQA1*03:03:01	DQA1*01:02:01	DQA1*01:02:01:02	DQA1*01:02:04	50
9	DQA1*04:01:01	DQA1*01:02:01	DQA1*01:02:01:04	DQA1*04:01:01	100
10	DQA1*02:01:01	DQA1*01:01:01	DQA1*01:01:01	DQA1*02:01	100
11	DQA1*01:03:01	DQA1*03:03:01	DQA1*01:03:01:02	DQA1*03:03:01	100
12	DQA1*01:02:01	DQA1*03:03:01	DQA1*01:02:01:02	DQA1*01:02:04	50
13	DQA1*05:05:01	DQA1*01:02:01	DQA1*01:02:01:04	DQA1*05:05:01:02	100
14	DQA1*05:02	DQA1*05:05:01	DQA1*05:09	DQA1*05:01:01:01	0
15	DQA1*05:05:01	-	DQA1*05:09	DQA1*05:09	0
16	DQA1*02:01:01	DQA1*01:04:01	DQA1*01:04:01:01	DQA1*02:01	100
17	DQA1*01:02:01	DQA1*03:03:01	DQA1*01:02:02	DQA1*01:11	0
18	DQA1*01:05:01	DQA1*03:03:01	DQA1*01:01:01	DQA1*03:03:01	100
19	DQA1*01:03:01	DQA1*01:02:01	DQA1*01:03:01:01	DQA1*01:02:01:04	100
20	DQA1*02:01:01	DQA1*01:01:02	DQA1*01:01:02	DQA1*02:01	100
21	DQA1*05:05:01	DQA1*01:02:01	DQA1*01:02:01:04	DQA1*05:05:01:01	100
22	DQA1*01:01:01	DQA1*01:03:01	DQA1*01:03:01:01	DQA1*01:01:02	50
23	DQA1*04:01:02	DQA1*05:05:01	DQA1*05:02	DQA1*04:01:02:02	50
24	DQA1*02:01:01	DQA1*05:01:01	DQA1*02:01	DQA1*05:01:01:01	50

% percentage concordance between the two methods. The difference in typing resolution of the same allele is ignored (the two methods are considered concordant in predicting that allele). '-:' Tool could not determine the HLA allele

Table 5.6 *In silico* HLA –DQB1 determination using HLA scan and HLA-HD tools

Sample ID	HLA DQB1$_{HLA-HD}$ [41,42]		HLA DQB1$_{HLA-SCAN}$ [40]		%
1	DQB1*06:02:01	DQB1*05:01:01	DQB1*06:02:01	DQB1*05:01:01:02	100
2	DQB1*04:23	DQB1*03:01:01	DQB1*04:02:01	DQB1*03:01:01:03	50
3	DQB1*05:01:01	DQB1*04:02:01	DQB1*04:02:01	DQB1*04:02:01	50
4	DQB1*03:02:01	DQB1*05:01:01	DQB1*03:02:01	DQB1*03:02:12	50
5	DQB1*06:09:01	DQB1*06:02:01	DQB1*06:09:01	DQB1*06:02:01	100
6	DQB1*06:09:01	DQB1*03:01:04	DQB1*06:09:01	DQB1*03:01:01:01	50
7	DQB1*04:23	DQB1*03:01:01	DQB1*04:02:01	DQB1*04:02:01	0
8	DQB1*06:02:01	DQB1*04:02:01	DQB1*06:02:01	DQB1*04:13	50
9	DQB1*06:02:01	DQB1*04:02:01	DQB1*06:02:01	DQB1*04:02:01	100
10	DQB1*05:01:24	DQB1*02:02:01	DQB1*05:01:01:03	DQB1*05:01:01:03	0
11	DQB1*06:03:01	DQB1*03:02:01	DQB1*06:03:01	DQB1*03:02:01	100
12	DQB1*06:02:01	DQB1*02:02:01	DQB1*06:02:01	DQB1*02:02:01	100
13	DQB1*03:01:01	DQB1*05:01:01	DQB1*03:01:01:01	DQB1*05:01:01:03	100
14	DQB1*02:01:01	DQB1*03:19:01	DQB1*02:01:01	DQB1*03:01:01:01	50
15	DQB1*03:19:01	DQB1*03:22	DQB1*03:01:01:03	DQB1*03:01:01:03	0
16	DQB1*03:03:02	DQB1*05:03:01	DQB1*03:03:02:01	DQB1*05:03:01:01	100
17	DQB1*06:02:01	DQB1*03:02:01	DQB1*03:02:01	DQB1*06:02:01	100
18	DQB1*05:01:01	DQB1*02:02:03	DQB1*02:12	DQB1*05:01:01:03	50
19	DQB1*06:03:01	DQB1*06:09:01	DQB1*06:09:01	DQB1*06:03:01	100
20	DQB1*05:01:01	DQB1*02:02:01	DQB1*05:01:01:03	DQB1*02:12	50
21	DQB1*06:02:01	DQB1*03:19:01	DQB1*06:02:01	DQB1*03:01:01:01	50
22	DQB1*06:03:01	DQB1*05:01:01	DQB1*06:03:01	DQB1*05:01:01:01	100
23	DQB1*03:19:01	-	DQB1*03:01:01:03	DQB1*03:01:01:03	0
24	DQB1*02:02:01	DQB1*02:01:08	DQB1*02:12	DQB1*02:12	0

% percentage concordance between the two methods. The difference in typing resolution of the same allele is ignored (the two methods are considered concordant in predicting that allele). '-:' Tool could not determine the HLA allele

Table 5.7 Ambiguous typing results generated by HLA –HD tool

Sample ID	HLA $A_{HLA\text{-}HD}$[41,42] ambiguous typing results	
3	H*01:01:01	H*02:03
	H*01:01:01	H*01:02
4	K*01:01:01	K*01:03
	K*01:01:01	K*01:01:01
7	K*01:02	K*01:01:01
	K*01:02	K*01:03
9	K*01:02	K*01:01:01
	K*01:02	K*01:03
10	K*01:03	K*01:02
	K*01:01:01	K*01:02
	K*01:01:01	K*01:03
	K*01:01:01	K*01:01:01
12	K*01:01:01	K*01:03
	K*01:01:01	K*01:01:01
	K*01:02	K*01:03
	K*01:02	K*01:01:01
13	K*01:01:01	K*01:02
	K*01:01:01	K*01:01:01
14	H*02:03	H*01:01:01
	H*01:02	H*01:01:01
18	K*01:01:01	K*01:02
	K*01:01:01	K*01:01:01
19	DOB*01:01:03	DOB*01:01:01
	DOB*01:01:03	DOB*01:02:01
	DOB*01:01:03	DOB*01:03
20	DRB4*01:03:01	DRB4*01:02
	DRB4*01:03:01	DRB4*01:03:01
23	K*01:01:01	K*01:02
	K*01:01:01	K*01:01:01
24	K*01:02	K*01:01:01
	K*01:02	K*01:03

5.5 Discussion

This chapter highlights the potential of using bioinformatics tools to understand HLA diversity in populations with limited HLA data. Despite the small sample size (24 WGS), HLA-HD and HLAscan[40-42] predicted high resolution HLA alleles in the South Africans assessed. Accurate high resolution (up to 8 digits) HLA imputation from WGS, WES and SNPs has become feasible with improved accuracy. Most imputation tools use mostly "non African" populations as references; as a result, accurate HLA imputation in African populations might be compromised due to the documented genetic diversity in Africans[44]. We sought to describe HLA genotypes from 24 genomes from the SAHGP as a bench mark for a larger project to describe HLA diversity in South Africans. HLAscan[40] and HLA-HD[41,42] tools predicted class I, II and non HLA genes from high coverage (50X) whole genome sequences.

South Africa has a unique demographic, ethnic and cultural diversity coupled with a high disease burden. The pilot SAHGP study demonstrated higher genetic variability amongst the eastern Bantu speakers of South Africa[28] than previosuly thought. HLA imputation from this dataset provides an essential bioresource for future population genetics studies, HLA-disease association studies and general human genetic diversity. The ability to use *in silico* methods to determine high resolution HLA typing results in South Africans benchmarks future application of using bionformatic approaches to understand HLA diversity. The successful application of HLA imputation to the SAHGP sequence data[28] provides a motivation to increase sample size to augument HLA typing results from these populations. There is generally limited HLA diversity data from southern Africans (reviewed in Chapter 2[9]). *In silico* HLA typing methods borrowing from existing data sets like the SAHGP sequence data[28] might help better understand HLA diversity in these populations.

Clinical HLA typing using sequencing based methods is still considered the gold standard, due to its high accuracy and ability to detect genetic differences across the HLA genes. However, these methods are still not accessible in most resource limited settings, and are generally expensive, hence limit in the number of individuals who have been typed (reviewed in[45]). Additionally, advances in NGS HLA typing enables high thoroughput high resolution typing. Generally, NGS generates a vast amount of

short read sequences that may be used for *in silico* HLA allele determination[8,46]. The main challenge in using short read sequences in HLA imputation is the polymorphic nature of the HLA gene region[47]. It is computationally challenging to accurately map or align the many short NGS reads to HLA allele reference sequences[5,48]. Short reads generated by most NGS technologies are difficult to use for HLA imputation owing to many potential candidate alleles and thereby leading to high sequence noise in imputation experiments. Most algorithms (tools) filter out the less common alleles before giving the final HLA result; for example, OptiType[49] only considers alleles reported in the allele frequency database (AFND)[50] and HLA-VBSeq only considers 100 possible HLA alleles[51]. This highlights the variability of HLA imputation tools. As a result, targeted sequence HLA typing remains the gold standard in clinical applications. It is generally difficult to know which allele(s) is represented by short sequencing read(s) considering the high similarity amongst different HLA alleles and the presence of pseudo-genes. Additionally, most reference alleles in the IMGT/HLA database do not have full length sequences[48], making it difficult to accurately call HLA alleles from short read sequence data. Additionally, the human genome reference does not fully cover HLA diversity, thereby confounding alignment of reads to the reference (reviewed in[52]).

HLA-HD and HLAscan[40-42] methods used in this chapter are alignment based imputation methods. Only reads aligning to human reference and HLA regions are used by the alignment based imputation methods like HLAscan[40] and HLA-HD[41,42]. A lot of potentially useful data (unmapped reads) is lost or not used, hence it might be beneficial to use assembly based methods to impute HLA genotypes from these individuals in future. HLA HD[41,42] considers sequence reads outside the antigen binding domain to determine HLA allele pair, unlike other tools like OptiType[49] and HLAreporter[53], which are restricted to the antigen binding domain. HLAscan[40] addresses the chromosome phasing problem in NGS HLA imputation. From previously typed data sets, HLAscan[40] outperformed (100% accuracy) PHLAT[16] (95% accuracy) and HLAreporter[53] (98% accuracy) in the four digit HLA typing of 1000 Genomes data set[16]. The HLAscan tool may be used for clinical purposes, but a minimum coverage depth over 90× is recommended[40] by the software developers. Generally, read depth (coverage) directly impacts the sensitivity and specificity of HLA allele calls[54,55]. The current study used Illumina HiSeq2000 generated whole

genome sequences with 50X coverage[28]. Read coverage and unmapped reads might have contributed to failure to predict some alleles and ambiguous typing results in this study (Tables S5.1 and S5.2).

WGS HLA imputation gives more information than SNP and WES based imputation; even gene regulatory elements and non coding elements like untranslated regions (UTR) and introns are covered. Basically, HLA imputation from short NGS reads can be classified into assembly based and alignment based methods. Assembly approaches assemble the short NGS reads into long contigs, which are then used for HLA imputation. Assembly methods are however time and computationally challenging as reported by HLAminer[56], HLAreporter[53] and ATHLATES[57] assembly tools. Alignment based approaches align the short reads to known HLA allele sequences in the IMGT HLA database[5,43]. SNP imputation needs an *apriori* reference panel with information on SNPs associated with HLA alleles in that population. There is currently no reference panel for South Africa, or Africans in general (reviewed in[58]).

5.6 Conclusions

Despite the limited sample size (24 whole genome sequences), this chapter highlights the potential of HLA imputation tools in understanding HLA diversity. The key highlight is the ability to impute high resolution (up to 8 digit typing resolution) from a population with limited HLA diversity data (reviewed in[9]). This provides a future framework to use more sequencing (whole exome, RNAseq, whole genome and SNP) datasets to fully understand HLA diversity in South Africans. Although HLA imputation results may not be ideally applicable to clinical applications like transplantation, they provide an economically feasible opportunity to screen potential donors without actually doing the high resolution HLA typing[58]. Additionally, despite the ability to use HLA imputation tools to accurately determine HLA alleles, the challenge of limited full length sequences of many alleles in the IMGT/HLA database[5] cannot be ignored. Despite the high resolution typing results from the *in silico* methods, standard HLA typing remains the gold standard for clinical applications. Imputation might benefit disease association, population genetics and

anthropological studies[59]. Unfortunately, the 24 individuals in this study were not HLA typed experimentally or for any medical reasons; hence we could not compare the *in silico* determined HLA alleles to HLA typing results. The two HLA imputation tools used in this study, namely HLA-HD and HLAscan[40-42] were evaluated on public datasets including the 1000 Genomes[22-25] with 100% accuracy. Imputation results described in this study highlights the feasibility of leveraging from existing sequence data from African populations to better understand HLA diversity in these populations.

5.7 Supplementary Information

Supplementary Tables are available in Addendum 1 as Table S5.1 and Table S5.2

Supplementary Table 5.1 (S5.1)

HLA alleles for 24 whole genome sequences from individuals enrolled in the SAHGP pilot study[28] determined by *in silico* HLA HD[41,42] method

Supplementary Table 5.2 (S5.2)

HLA alleles for 24 whole genome sequences from individuals enrolled in the SAHGP pilot study[28] determined by *in silico* HLAscan [40] method.

5.8 References

1. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood. 2007;110(13):4576-83.
2. Kunze-Schumacher H, Blasczyk R, Bade-Doeding C. Soluble HLA technology as a strategy to evaluate the impact of HLA mismatches. J Immunol Res. 2014;246171(10):1.
3. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L, et al. The DNA sequence and analysis of human chromosome 6. Nature. 2003;425:805-11.
4. Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, et al. Deep whole-genome sequencing of 100 southeast Asian Malays. Am J Hum Genet. 2013;92(1):52-66.
5. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43(Database issue):20.
6. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet. 2002;71(4):759-76.
7. Lam TH, Tay MZ, Wang B, Xiao Z, Ren EC. Intrahaplotypic Variants Differentiate Complex Linkage Disequilibrium within Human MHC Haplotypes. Scientific Reports. 2015;5:16972-.
8. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16(th) IHIW : review of HLA typing by NGS. Int J Immunogenet. 2013;40(1):72-6.
9. Tshabalala M, Mellet J, Pepper MS. Human Leukocyte Antigen Diversity: A Southern African Perspective. J Immunol Res. 2015;746151(10):12.
10. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet. 2009;54(1):15-39.
11. Pidala J, Wang T, Haagenson M, Spellman SR, Askar M, Battiwalla M, et al. Amino acid substitution at peptide-binding pockets of HLA class I molecules increases risk of severe acute GVHD and mortality. Blood. 2013;122(22):3651-8.
12. Cereb N, Kim HR, Ryu J, Yang SY. Advances in DNA sequencing technologies for high resolution HLA typing. Hum Immunol. 2015;76(12):923-7.

13. Profaizer T, Kumanovics A. Human Leukocyte Antigen Typing by Next-Generation Sequencing. Clin Lab Med. 2018;38(4):565-78.

14. Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. J Hum Genet. 2015;60(11):665-73.

15. Hayashi S, Yamaguchi R, Mizuno S, Komura M, Miyano S, Nakagawa H, et al. ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data. BMC Genomics. 2018;19(1):018-5169.

16. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics. 2014;15(325):1471-2164.

17. Nariai N, Kojima K, Saito S, Mimori T, Sato Y, Kawai Y, et al. HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. BMC Genomics. 2015;16(2):1471-2164.

18. Lee H, Kingsford C. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. Genome Biol. 2018;19(1):018-1388.

19. Lee H, Kingsford C. Accurate Assembly and Typing of HLA using a Graph-Guided Assembler Kourami. Methods Mol Biol. 2018;1802:235-247.

20. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, et al. Imputing amino acid polymorphisms in human leukocyte antigens. PloS ONE. 2013;8(6):e64683.

21. Dilthey AT, Gourraud PA, Mentzer AJ, Cereb N, Iqbal Z, McVean G. High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. PLoS Comput Biol. 2016;12(10):e1005151.

22. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.

23. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

24. The Genomes Project C, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526:68.

25. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81.

26. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517(7534):327-32.

27. Pepper MS. Launch of the Southern African Human Genome Programme. S Afr Med J. 2011;101(5):287-8.

28. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, et al. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nat Commun. 2017;8(1):017-00663.

29. Disotell TR. Archaic human genomics. Am J Phys Anthropol. 2012;55:24-39.

30. Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo JM, et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell. 2012;150(3):457-69.

31. Marks SJ, Montinaro F, Levy H, Brisighelli F, Ferri G, Bertoncini S, et al. Static and moving frontiers: the genetic landscape of Southern African Bantu-speaking populations. Mol Biol Evol. 2015;32(1):29-43.

32. Henn BM, Botigue LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci U S A. 2016;113(4):28.

33. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, et al. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A. 2011;108(13):5154-62.

34. WHO. Global Health Report. Geneva2013.

35. Stewart JR, Stringer CB. Human evolution out of Africa: the role of refugia and climate change. Science. 2012;335(6074):1317-21.

36. Relethford JH. Genetic evidence and the modern human origins debate. Heredity. 2008;100(6):555-63.

37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

38. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013;29(16):2041-3.

39. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767-71.

40. Ka S, Lee S, Hong J, Cho Y, Sung J, Kim H-N, et al. HLAscan: genotyping of the HLA region using next-generation sequencing data. BMC Bioinformatics. 2017;18(1):258.

41. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. Human Mutation. 2017;38(7):788-97.

42. Kawaguchi S, Higasa K, Yamada R, Matsuda F. Comprehensive HLA Typing from a Current Allele Database Using Next-Generation Sequencing Data. Methods Mol Biol. 2018;1802:225-33.

43. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. Nucleic Acids Research. 2013;41(D1):D1222-D1227.

44. Disotell TR. Archaic human genomics. . Am J Phys Anthropol 2012;55:24-39.

45. Erlich H. HLA DNA typing: past, present, and future. Tissue Antigens. 2012;80(1):1-11.

46. Gabriel C, Furst D, Fae I, Wenda S, Zollikofer C, Mytilineos J, et al. HLA typing by next-generation sequencing - getting closer to reality. Tissue Antigens. 2014;83(2):65-75.

47. Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, et al. HLA typing from RNA-Seq sequence reads. Genome Medicine. [journal article]. 2012 December 22;4(12):102.

48. Robinson J, Soormally AR, Hayhurst JD, Marsh SGE. The IPD-IMGT/HLA Database - New developments in reporting HLA variation. Hum Immunol. 2016;77(3):233-7.

49. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014;30(23):3310-6.

50. González-Galarza Faviel F, Takeshita Louise YC, Santos Eduardo JM, Kempson F, Maia Maria Helena T, Silva Andrea Luciana Soares d, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Research. 2015;43(D1):D784-D8.

51. Nariai N, Kojima K, Saito S, Mimori T, Sato Y, Kawai Y, et al. HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. BMC Genomics. 2015;16(2):S7-S.

52. Boegel S, Scholtalbers J, Löwer M, Sahin U, Castle JC. In silico HLA typing using standard RNA-seq sequence reads. Methods in Molecular Biology. 2015; 1310: 247-58.

53. Huang Y, Yang J, Ying D, Zhang Y, Shotelersuk V, Hirankarn N, et al. HLAreporter: a tool for HLA typing from next generation sequencing data. Genome Medicine. 2015;7(1):25.

54. Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q. Impact of three Illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol. 2015;76(2-3):166-75.

55. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121-32.

56. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, et al. Derivation of HLA types from shotgun sequence datasets. Genome Medicine. 2012;4(12):95.

57. Liu C, Yang X, Duffy B, Mohanakumar T, Mitra RD, Zody MC, et al. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Research. 2013;41(14):e142.

58. Meyer D, Nunes K. HLA imputation, what is it good for? Human Immunology. 2017;78(3):239-41.

59. Bauer DC, Zadoorian A, Wilson LOW, Melbourne Genomics Health A, Thorne NP. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. Briefings in bioinformatics. 2016;19(2):179-87.

CHAPTER 6

GENERAL DISCUSSION AND CONCLUSION

6.1 General Discussion

The human leukocyte antigen (HLA) region on the short arm of chromosome 6 in humans is highly polymorphic, currently with 20 088 alleles being described in the IMGT HLA database (3.34.0 release October 2018) (https://www.ebi.ac.uk/ipd/imgt/hla/stats.html)[1]. HLA molecules bind to endogenous antigenic epitopes (HLA class I) and present them to $CD8^+$ T lymphocytes while HLA class II molecules present antigenic peptides to $CD4^+$ T lymphocytes. The polymorphic nature of HLA genes allows the presentation of a wide range of peptides to the immune system. In addition to the polymorphic nature of the HLA region, each offspring has unique HLA alleles inherited from both parents. HLA typing methods have evolved from low resolution serology to high resolution sequencing based methods. Individuals' HLA genotypes can now be determined to protein level (digit typing) owing to advancement in sequencing technologies. There are additionally an increasing number of studies generating next generation sequencing data that may be used for high resolution HLA typing (through HLA imputation).

There is generally a marked difference in HLA diversity distribution globally, with geographically separated regions showing varying degrees of diversity. Most HLA loci, except for HLA-DPB1, show high allele numbers across populations[2,3]. The global distribution of HLA diversity provides insight into human migration patterns, and could help understand past pathogen exposures[4] and trace human evolution[5]. South Africa has a heterogeneous population, whose HLA genetic diversity has not been well described, despite the immunological significance of HLA. Previous studies have identified novel alleles in South African populations[6,7], suggesting high HLA diversity in these populations. HLA diversity in South African populations is

generally not conclusively known. Despite global efforts in understanding human genetic diversity through projects like Hap Map Project[8], 1000 Genomes Project [9], the African Genome Variation Project[10] and the Southern African Human Genome Programme (SAHGP)[11], there is limited information available on South African populations. Of particular note is the limited data on HLA genetic diversity from South African populations. The hypothesis was driven by that poor understanding of HLA genetic diversity amongst South Africans and how this might impact clinical applications including vaccine development, disease association and transplantation. This thesis sought to define the extent of HLA diversity in South African populations. The approach was divided into three sections:

i) An extensive literature search for South African HLA diversity studies to highlight the paucity of information;
ii) Documentation of HLA diversity from the South African Bone Marrow Registry (SABMR), the National Health Laboratory Services (NHLS) and the South African National Blood Transfusion Services (SANBS). These three institutions provide most of public health HLA typing service in South Africa;
iii) The use of computational methods to determine HLA alleles from existing whole genome sequencing data.

6.2 Summary of the key findings

Chapter 2: There is limited HLA diversity data in the public domain for South Africans and southern Africans in general. Most South African studies have HLA data generated from disease association studies, or have low resolution typing results despite improvements in typing methods. The paucity of information on HLA genotypic data for southern African populations' impacts on disease association studies, population based vaccine design and transplantation outcomes.

Chapter 3: The South African Bone Marrow Registry (SABMR) is the only active bone marrow donor registry in Africa supporting transplantation programs. Hapl-o-Mat software was used to compute allele and haplotype frequencies from 237 volunteer bone marrow donors typed at various resolutions, with some alleles in

multiple allele code (MAC) format. Four hundred and thirty eight (438) HLA ~A, 235 HLA ~B, 234 HLA ~DRB1, 41 HLA ~DQB1 and 29 HLA ~C alleles are reported. The most frequent alleles were A*02:02g (0.096), B*07:02g (0.082), C*07:02g (0.180), DQB1*06:02 (0.157) and DRB1*15:01 (0.072). Additionally, the most common haplotype A*03:01g~B*07:02g~C*07:02g~DQB1*06:02~DRB1*15:01 (0.067) was previously reported in other global populations at varying frequencies. Despite the small sample size (237), these results form a key resource for future population studies, disease association studies and support donor recruitment strategies into the SABMR.

Chapter 4: This chapter describes high resolution typing (HLA ~A, HLA ~B, HLA ~C, HLA ~DRB1, HLA ~DQA1and HLA ~DQB1) in 3007 individuals, and low resolution typing (HLA ~A, HLA ~B, HLA ~C, HLA ~DRB1, HLA ~DQA1, HLA ~DQB1 and HLA ~DPB1) in 51 891 individuals. These individuals were previously typed by SANBS or NHLS as part of a routine clinical service. The South African HLA data showed genetic distinctness compared to other global populations using non metric multidimensional scaling. Additionally, principal component analysis showed genetic relatedness of South Africans with other sub Sahara African populations. The large HLA data sample size from South Africans might be a useful resource to support anthropological studies, disease association studies, population based vaccine development and donor recruitment programs.

Chapter 5: HLA typing services are generally centralized and inaccessible in most resource limited settings. However, with an increase in population based NGS data sets, it is increasingly feasible to determine HLA alleles from these datasets using i*n silico* methods. This chapter describes high resolution (up to 8 digit) determination of HLA alleles from 24 whole genome sequences generated from SAHGP (a government funded initiative to understand human genetic diversity) using *in silico* methods. The *in silico* HLA imputation methods used predicted high resolution HLA alleles including HLA genes from the 24 genomes. Despite the small sample size, this chapter highlights the potential of HLA imputation tools in understanding HLA diversity. Additionally, the chapter highlights the need for full length sequences for HLA alleles in the IMGT/HLA database to support accurate HLA imputation tools. Although *in silico* methods successfully predicted high resolution HLA typing results,

standard HLA typing remains the gold standard for clinical applications. HLA imputation might benefit disease association studies, population genetics and anthropological studies.

6.3 Conclusions

It is thus important to fully understand HLA diversity in South African populations, to establish HLA-disease associations, and to use this data for the informed design of population-specific vaccines against the many diseases, and to improve on donor-recipient matching. There is generally limited HLA diversity data for South African populations which impacts on clinical applications including transplantation. Continued documentation and research on HLA diversity in clinical settings like in the SABMR, SANBS and NHLS, might provide a future resource to better understand HLA diversity in these populations. Additionally, HLA imputation tools may be used to better understand HLA diversity in settings where HLA data is limited. With improvements in NGS and a reduction in sequencing costs, HLA imputation offers an economically viable approach to obtain HLA genotypes from a large pool of individuals without additional cost. The lack of HLA data for South African populations has limited our understanding of disease association studies, population based vaccine development, transplantation clinical outcomes. Generally, correlates of protective immunity for many diseases affecting South Africans are poorly understood.

6.4 Limitations of the study

- There is limited publicly available HLA diversity data from southern African populations for extensive comparison with South African datasets;
- The study relied heavily on public data sets, which might not be exhaustive (representative of the population). Conclusions on HLA diversity data for South Africans are thus to be interpreted with caution;
- Demographic data of participants could not be accessed due to ethical considerations. Lack of ethnic data for South Africans was a major limitation in

understanding HLA diversity in these populations; considering the genetic differences amongst ethnic groups reported in other genetic studies.

- Additional demographic information that could not be accessed due to ethical considerations, that could impact on interpretation of HLA diversity described in this thesis include the disease state of participants and familial relatedness of some participant(s)
- The retrospective nature of the study resulted in mixed resolution data with ambiguous typing which could not be corrected. Conclusions based on mixed resolution typing results should be cautiously interpreted.
- Only reads mapping to human reference genome and the HLA genes are used by alignment based imputation tools like HLAscan and HLA HD[12,13]. Unmapped sequence reads are discarded (a limitation since some of these reads might highlight further information of the genetic structure of the HLA region).

6.5 Future research directions

- There is a need to build an HLA diversity resource for southern Africa (and Africa as a whole) copying from the HLA-net[14] example. HLA-net is a European network focusing on HLA diversity and its applications include histocompatibility, transplantation, epidemiology and population genetics. This network has developed analysis pipelines, tools and guidelines for HLA diversity data for mostly European populations[14,15]. An African HLA resource might be useful for future studies including donor recruitment strategies[16], population studies[4,5,15] and disease association studies[17-20];
- Analyse a larger SABMR sample size and compare its HLA diversity data to other registries globally;
- In addition to an African reference panel to improve imputation accuracy, the fact that Africans are genetically diverse makes it difficult to identify novel HLA alleles using alignment based imputation approaches. Computationally intensive assembly based imputation is proposed to fully understand the HLA diversity in the 24 South African genomes.

6.6 References

1. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43(Database issue):D423-D31.
2. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Research. 2011;39(1):D913-D9.
3. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PloS ONE. 2011;6(2):e14643.
4. Sanchez-Mazas A, Fernandez-Vina M, Middleton D, Hollenbach JA, Buhler S, Di D, et al. Immunogenetics as a tool in anthropological studies. Immunology. 2011;133(2):143-64.
5. Sanchez-Mazas A, Thorsby E. HLA in anthropology: the enigma of Easter Island. Clin Transpl. 2013:167-73.
6. Hayhurst JD, du Toit ED, Borrill V, Schlaphoff TEA, Brosnan N, Marsh SGE. Two novel HLA alleles, HLA-A*30:02:01:03 and HLA-C*08:113, identified in a South African bone marrow donor. Tissue Antigens. 2015;85(4):291-3.
7. Paximadis M, Mathebula TY, Gentle NL, Vardas E, Colvin M, Gray CM, et al. Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population. Human Immunol. 2012;73:80-92.
8. HapMapProject. The International HapMap Project. Nature. 2003;426(6968):789-96.
9. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
10. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517(7534):327-32.
11. Pepper MS. Launch of the Southern African Human Genome Programme. S Afr Med J. 2011;101(5):287-8.

12. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. Human Mutation. 2017;38(7):788-97.

13. Kawaguchi S, Higasa K, Yamada R, Matsuda F. Comprehensive HLA Typing from a Current Allele Database Using Next-Generation Sequencing Data. Methods Mol Biol. 2018;1802:225-33.

14. Nunes JM, Buhler S, Roessli D, Sanchez-Mazas A. The HLA-net GENE[RATE] pipeline for effective HLA data analysis and its application to 145 population samples from Europe and neighbouring areas. Tissue Antigens. 2014;83(5):307-23.

15. Sanchez-Mazas A, Vidan-Jeras B, Nunes JM, Fischer G, Little AM, Bekmane U, et al. Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA-NET methodological recommendations. International Journal of Immunogenetics. 2012;39(6):459-76.

16. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Med J. 2007;48(1):11-23.

17. Garamszegi LZ. Global distribution of malaria-resistant MHC-HLA alleles: the number and frequencies of alleles and malaria risk. Malar J. 2014;13(349):1475-2875.

18. Ramsay M. Africa: continent of genome contrasts with implications for biomedical research and health. FEBS Lett. 2012;586:2813-9.

19. Sanchez A, Wagoner KE, Rollin PE. Sequence-based human leukocyte antigen-B typing of patients infected with Ebola virus in Uganda in 2000: identification of alleles associated with fatal and nonfatal disease outcomes. J Infect Dis. 2007;196(2):S329-36.

20. Yim JJ, Selvaraj P. Genetic susceptibility in tuberculosis. Respirology. 2010;15(2):241-56.

APPENDICES

Appendix 1 University of Pretoria Ethics Approval

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 22 May 2002 and Expires 20 Oct 2016.
- IRB 0000 2235 IORG0001762 Approved dd 22/04/2014 and Expires 22/04/2017.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences Research Ethics Committee

16/07/2015

Approval Certificate
New Application

Ethics Reference No.: 220/2015

Title: Human Leukocyte Antigen (HLA) genetic diversity in South African populations

Dear Mqondisi Tshabalala

The **New Application** as supported by documents specified in your cover letter dated 19/05/2015 for your research received on the 25/05/2015, was approved by the Faculty of Health Sciences Research Ethics Committee on its quorate meeting of 15/07/2015.

Please note the following about your ethics approval:
- Ethics Approval is valid for 3 years
- Please remember to use your protocol number (**220/2015**) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, or monitor the conduct of your research.

Ethics approval is subject to the following:
- The ethics approval is conditional on the receipt of 6 monthly written Progress Reports, and
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

Dr R Sommers; MBChB; MMed (Int); MPharMed.
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes 2004 (Department of Health).

☎ 012 354 1677 🖷 0866516047 deepeka.behari@up.ac.za http://www.healthethics-up.co.za
✉ Private Bag X323, Arcadia, 0007 - 31 Bophelo Road, HW Snyman South Building, Level 2, Room 2.33, Gezina, Pretoria

Appendix 2 University of Pretoria Ethics amendment certificate

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 22 May 2002 and Expires 20 Oct 2016.
- IRB 0000 2235 IORG0001762 Approved dd 22/04/2014 and Expires 22/04/2017.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences Research Ethics Committee

26/11/2015

Approval Certificate
Amendment
(to be read in conjunction with the main approval certificate)

Ethics Reference No.: 220/2015

Title: Human Leukocyte Antigen (HLA) genetic diversity in South African populations

Dear Mqondisi Tshabalala

The **Amendment** as described in your documents specified in your cover letter dated 1/11/2015 received on 2/11/2015 was approved by the Faculty of Health Sciences Research Ethics Committee on its quorate meeting of 25/11/2015.

Please note the following about your ethics amendment:
- Please remember to use your protocol number (**220/2015**) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committe may ask further questions, seek additional information, require further modification, or monitor the conduct of your research.

Ethics amendment is subject to the following:
- The ethics approval is conditional on the receipt of 6 monthly written Progress Reports, and
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

*** Kindly collect your original signed approval certificate from our offices, Faculty of Health Sciences, Research Ethics Committee, H W Snyman South Building, Room 2.33 / 2.34.*

Dr R Sommers; MBChB; MMed (Int); MPharMed.
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes 2004 (Department of Health).

♦ Tel:012-3541330 ♦ Fax:012-3541367 Fax2Email: 0866515924 ♦ E-Mail: fhsethics@up.ac.za
♦ Web: //www.healthethics-up.co.za ♦ H W Snyman Bld (South) Level 2-34 ♦ Private Bag x 323, Arcadia, Pta, S.A., 0007

Appendix 3 University of Pretoria Ethics Extension

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences Research Ethics Committee

26/04/2018

Mqondisi Tshabalala
Department of Immunology
University of Pretoria

Dear Mqondisi Tshabalala

RE.: 220/2015 ~ Letter dated 9 April 2018

Protocol Number	**220/2015**
Protocol Title	Human Leukocyte Antigen (HLA) genetic diversity in South African populations
Principal Investigator	Mqondisi Tshabalala **Tel: Email:** mtshabaz@gmail.com **Dept:** Immunology

We hereby acknowledge receipt of the following document:

- Extension from 16 July 2018 till 16 January 2019

which has been approved at 25 April 2018 meeting.

With regards

Dr R Sommers; MBChB; MMed (Int); MPharMed; PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

012 356 3085 **fhsethics.up.ac.za** http://www.up.ac.za/healthethics
Private Bag X323, Arcadia, 0007 - Tswelopele Building, Level 4-59, Gezina, Pretoria

Appendix 4 SANBS Ethics Approval

SOUTH AFRICAN NATIONAL BLOOD SERVICE NPC

Human Research Ethics Committee

SANBS
South African National Blood Service

Association Incorporated Under Section 21
Registration No. 2000/026390/08

OHRP Number : IORG0006278
FWA Registration Number : IRB00007553
SA NHREC Registration Number : REC-270606-013

Secretariat: Tel: 011 761 9135 | Fax: 011 761 9137 | Cell: 082 523 8523 | thandiwe.matsoso@sanbs.org.za

To: Mqondisi Tshabalala
E-mail:mtshabaz@gmail.com

Dear Mqondisi Tshabalala

DATE OF COMMITTEE MEETING:	**10 November 2015**
PROJECT TITLE:	**Human Leukocyte Antigen (HLA) genetic diversity in South African populations**
DECISION OF THE COMMITTEE:	**Approved**
CLEARANCE CERTIFICATE NO:	**2015/31**

- Execution of the study must be compliant with applicable guidelines and policies.
- Any amendment, extension or other modifications to the protocol must be submitted to this Ethics Committee for approval prior to implementation.
- The Committee must be informed of any serious adverse event, planned and unplanned termination of the study.
- A progress report should be submitted yearly for long-term studies and a final report at completion of both short term and long term studies.
- Kindly refer to the SANBS HREC clearance certificate number on all future correspondence on this study to the HREC secretariat.
- This approval is valid for 5 years from the date stated above.

COMMITTEE GUIDANCE DOCUMENTS:

- International Conference on Harmonization (ICH) Good Clinical Practices (GCP) Guideline (ICH, 1996), Ethics in Health Research: Principles, Structures and Procedures (SA Department of Health, 2004); Guidelines for Good Practice in the Conduct of Clinical Trials in Human Participants in South Africa (SA Department of Health, 2006); Ethical Principles for Medical Research Involving Human: Declaration of Helsinki (World Medical Association, 2013); Reviewing Clinical trials: A Guide For Ethics Committees (Karlberg and Speers, 2010)

1 December 2015

CHAIRPERSON: Prof J.N. Mahlangu **DATE**

Appendix 5 NHLS Ethics Approval

NATIONAL HEALTH LABORATORY SERVICE

Academic Affairs and Research
Modderfontein Road, Sandringham, 2031
Tel: +27 (0)11 386 6142
Fax: +27 (0)11 386 6296
Email: babatyi.kgokong@nhls.ac.za
Web: www.nhls.ac.za

25 January 2016

Applicant: Mqondisi Tshabalala
Institution: University of Pretoria
Department: Immunology
Email: mtshabaz@gmail.com
Tel: 012 319 2107

Re: Approval to access National Health Laboratory Service (NHLS) Data

Your application to undertake a research project "**Human Leukocyte Antigen (HLA) Genetic Diversity in South African Populations**" using data from the NHLS database has been reviewed. This letter serves to advise that the application has been approved and the required data will be made available to you to conduct the proposed study as outlined in the submitted application.

Please note that the approval is granted on your compliance with the NHLS conditions of service and that the study can only be undertaken provided that the following conditions have been met.

- Ethics approval is obtained from a recognised SA Health Research Ethics Committee.
- Processes are discussed with the relevant NHLS departments (i.e. Information Management Unit and Operations Department) and are agreed upon.
- Confidentiality is maintained at participant and institutional level and there is no disclosure of personal information or confidential information as described by the NHLS policy.
- A final report of the research study and any published paper resulting from this study are submitted and addressed to the NHLS Academic Affairs and Research office and the NHLS has been acknowledged appropriately.
- Initialise collaboration with a preferred NHLS Researcher.

Please note that this letter constitutes approval by the NHLS Academic Affairs and Research. Any data related queries may be directed to Sue Candy, manager NHLS Corporate Data Warehouse, Tel: (011) 386 6036. Email: sue.candy@nhls.ac.za.

Yours sincerely,

Dr Babatyi Malope-Kgokong
National Manager: Academic Affairs and Research

Appendix 6 SAHGP Data Access Approval

Southern African Human Genome Programme

Prof M S Pepper
University of Pretoria
South Africa

19 March 2018

Dear Prof Michael S Pepper,

Re: Request to access the SAHGP data – whole genome sequence data on 24 South African individuals – Request code: SAHGP004

Project title: HLA diversity in the Southern African Human Genome Project (SAHGP) data set using imputation methods

Thank you for your application and request.
The SAHGP Data Access Committee has reached the following conclusions:

Approval on condition that the applicant addresses the following:
1. Provide assurance that ethics approvals will be kept up to date for the duration of the project.

Please address the queries in a letter and provide the additional information if requested. To gain access to the data, the next step is to complete the Data Access Agreement and have it signed by the relevant individuals. Please find the form attached to the email.

Once we receive the completed form and have assessed your response. We will send a message to the European Genome Phenome Archive to contact you with regard to data transfer.

Please contact the undersigned should you require further clarification.

Yours sincerely,

M. Ramsay

Michele Ramsay

On behalf of the SAHGP Data Access Committee
Michele.ramsay@wits.ac.za

Appendix 7 EGA SAHGP Data Access Procedure

a) Requesting the data set

java -jar EgaDemoClient.jar –p demo@test.org '123pass' -rfd EGAD00001003791 -re abc -label request_EGAD00010000498

demo@test.org and '123pass' is the email address and password of the individual approved to the access the data. “abc” is user defined decryption key. Data is downloaded in encrypted format for security reasons.

b) Downloading Request

java -jar EgaDemoClient.jar –p demo@test.org '123pass' -dr request_EGAD00001003791 -nt 7

The optional parameter '-nt' specifies the number of parallel download streams (7 in this case), -dr lists the download request

c) Decrypt downloaded data

java -jar EgaDemoClient.jar –p ***demo@test.org*** ‘123pass’ –dc <path to downloaded data> -dck <abc>

The decryption password ‘abc’ is used to decrypt all the downloaded files

Appendix 8 Customised script for HLA imputation

```
#!/usr/bin/env python

import sys
import os
import re

indir = sys.argv[1]

all_dirs = os.listdir(indir)
for dir in all_dirs:
    ind = dir
    if 'HLA' not in dir:
        bam_files = os.listdir(indir + '/' + dir + '/Assembly')#location of BAM files
        for bam_file in bam_files:
            if bam_file.endswith('bam'):
                command = 'samtools view -h -b ' + indir + '/' + dir + '/Assembly/' + bam_file + ' "chr6:28866528-33775446" > ' + ind + '_HLA.bam'#extraction of HLA region BAM file
                print command
                os.system(command)

                command2 = '/apps/jdk-8u162/bin/java -jar /apps/picard-2.17.11/picard.jar SamToFastq INPUT=' + ind + '_HLA.bam' + ' FASTQ=' + ind + '_1.fastq SECOND_END_FASTQ=' + ind + '_2.fastq UNPAIRED_FASTQ=' + ind + '_U.fastq VALIDATION_STRINGENCY=LENIENT'
                print command2
                os.system(command2)#conversion to fastq file formats

#explains the command
```