Genetic Structure of East African Populations Based on HLA Class I Genes

Aamer, Waleed Hussein Omer

Doctor of Philosophy

Department of Genetics
School of Life Science
SOKENDAI (The Graduate University for Advanced Studies)

Genetic Structure of East African Populations Based on HLA Class I Genes

PhD thesis

Aamer, Waleed Hussein Omer

The Graduate University for Advanced Studies (SOKENDAI)

School of Life Science

Department of Genetics

December 2016

Acknowledgments

I would like to thank everyone who made this work possible especially:

- My supervisor **Professor Ituro INOUE** for his patient, encouragement, guidance and personal support.

- Dr. Hirofumi Nakaoka for the unlimited support and continuous advice.

- Participants who gave their samples in this study are much appreciated.

- My colleagues at the division of human genetics and all of the other colleagues, thank you for providing the good scientific and friendly environment that was very helpful for me to do the work.

- Last not least, I would like to express my worm and deep gratitude to my lovely family. I really feel indebted to them for their continuous support and patient; my wife **Huda**, my son **Sanad**, my father **Hussein**, my mother **Fatima**, my brother **Adil** and my sisters **Safa, Samar** and **Iman,** I can't even find the words to express my love and care, but thank you for everything.

I also would like to thank MEXT for providing the scholarship during my stay in Japan, which made my dream possible.

Contents

List of Tables

List of Figures

Summary

East Africa is characterized by high levels of genetic, linguistic, and cultural diversity. Studies have shown that, in Africans, there is a strong correlation between genetic and linguistic families. Focusing on the Sudanic region of East Africa, a recent study based on genome-wide SNP data found two substructures in the Sudanese population. These two substructures were related to the Afro-Asiatic and Nilo-Saharan/Niger-Congo linguistic families. Furthermore, the Afro-Asiatic family contains several groups with different genetic backgrounds, which requires further differentiation. Here I studied the genetic diversity of HLA class I genes in eight ethnic groups from East Africa. The samples I used represent three countries in East Africa: Sudan, South Sudan and Ethiopia. The samples from Sudan have a wide geographical distribution and all samples belong to one of three known African linguistic families: Afro-Asiatic, Nilo-Saharan, and Niger-Congo.

I first sequenced the samples using next-generation sequencing on the Miseq platform. Then, I analyzed the sequence data and identified *HLA-A*, *HLA-B*, and *HLA-C* genotypes. While identifying the genotypes, I found four mutations not previously described in the sequence database, which can be considered to be new HLA alleles. I then confirmed all of the new mutations by Sanger sequencing. The HLA genotypes were then used to calculate allele frequencies and determine the distribution of allele frequencies in each group. To find out how diverse the study populations are, I calculated heterozygosity and F_{st} values, which indicated a high level of heterozygosity and population differentiation.

To examine the population structure, I performed principal component analysis (PCA) using the calculated allele frequencies. I found that most of the variances were explained by the first and second principal components. The PCA identified three clusters that correlated with the linguistic families of the study groups. The first cluster was defined by alleles common to

the Nilo-Saharan groups, the second cluster was defined by alleles that are common in Afro-Asiatic groups, and the third cluster was defined by alleles of the Sudanese Arab groups. Interestingly, the division in cluster 3 between Arabs and other groups within the Afro-Asiatic family was not previously seen in SNP data, suggesting the possibility of further differentiation within this family.

For further analysis, I focused on cluster 3 and investigated whether the identified alleles form haplotypes. I estimated haplotype frequencies and compared them to the PCA clustering patterns. I found that alleles of the most common haplotypes are also found in the same cluster and those alleles are informative for differentiation between groups because they have similar patterns of allele frequency distribution.

Having identified the informative haplotypes in cluster 3, I checked whether they are tightly linked by estimating pairwise linkage disequilibrium (LD) between alleles of the three HLA loci. I observed that the *HLA-B* and *HLA-C* alleles are in perfect LD most of the time. On the other hand, the *HLA-A* alleles show lower LD values, which are expected as LD decays with an increase in physical distance along the chromosome.

Furthermore, I decided to trace the ancestry of cluster 3 alleles, so I searched for the haplotypes common in the Sudanese Arabs in other populations in the Allele Frequency Net Database (AFND), which is the largest database for global distribution of HLA alleles. Interestingly, the most common two-locus haplotypes (*HLA-B* and *HLA-C*) among Arab groups were not found in all Sub-Saharan African populations and were found mainly in populations from Asia, Europe, and North America. More interestingly, I found that the most common B-C haplotype among Sudanese Arabs (B*51:01-C*15:02) has a frequency of 4.7% in the Saudi Arabian population.

Finally, finding cluster 3 alleles exclusively outside of Africa strongly suggested that the Sudanese Arab groups have experienced gene flow from non-African sources. To understand further the population structure of the Sudanese groups in the context of other populations, I downloaded data from the AFND database, including on populations from Sub-Saharan Africa, the Middle East, and Europe, and merged them with the data from East Africa. PCA on the integrated data sets confirmed the previously seen clustering patterns in our own data. Intriguingly, the PCA of the merged data sets showed that Arab groups from Sudan are the closest to all non-African populations, particularly populations from the Middle East like the Saudi one.

In conclusion, this study identified four new HLA alleles and established a map of HLA class I allele and haplotype frequencies in Sudan. I also found a substructure within Afro-Asiatic groups, which separates Sudanese Arabs from non-Arab groups. The identified substructure seems to have been affected by gene flow from West Asia or the Middle East.

Chapter One

INTRODUCTION

1.1 Background

The human major histocompatibility complex (MHC) [in humans also called the human leukocyte antigen (HLA); I will use these two terms interchangeably] is a genomic complex located on 6p21.31. Spanning 3.6 Mb of the human genome, HLA loci have been associated with many diseases (Gough *et al*. 2007), making HLA the loci with the highest number of disease associations. The numerous diseases associated with HLA loci include type I diabetes (TID), rheumatoid arthritis (RA), and many other diseases related to the immune system. Two main features make HLA interesting genomic loci. The first is the higher gene density compared with the rest of the genome (224 genes according to the MHC Sequencing Consortium 2009, see section **1.2**). The second, and most important, is the highly polymorphic nature of many genes located in this region (the IPD-IMGT/HLA database currently contains over 4000 *HLA-B* alleles, Robinson *et al.* 2015). Several types of variation have been observed in the HLA complex, including single-nucleotide polymorphisms (SNPs), insertion-deletions (INDELs), large segmental duplications (>1000 bp), and copy number variations (Gaudieri *et al*. 2000). The accumulation of these variations in the HLA genes ultimately led to an increase in genetic diversity at the population level. Such increased diversity provides valuable markers for population genetics and disease association studies. Furthermore, comparative analysis of HLA genes in the same/different populations offers an ideal tool for tracing past demographic events, understanding variations in disease susceptibility, and studying variability in the effects of drugs between different groups/populations.

1.1 Genomic organization of MHC loci

In 1999, the MHC Sequencing Consortium published the first complete sequence and map of all genes within the MHC loci (MHC Sequencing Consortium 1999), in which 224 genes were identified. Historically, MHC loci have been categorized into three classes based on gene localization and functional similarities between genes in the same cluster. The first cluster of genes is the MHC class I genes, which is located toward the telomere and includes the classical *HLA-A*, *HLA-B*, and *HLA-C* genes and non-classical genes such as HLA-G and HLA-E (**Figure 1.1**). The second cluster is class II genes, which includes classical genes such as HLA-DPA, HLA-DPB, HLA-DRA, HLA-DRB, HLA-DQA, and HLA-DQB, in addition to other non-classical genes. Both class I and class II have numerous pseudogenes; although there is no evidence for their expression, they may play a role in MHC evolution and the emergence of new HLA genes/alleles (MHC Sequencing Consortium 1999). In contrast to class I and II genes, the third cluster, class III genes, does not contain many pseudogenes, but it is very dense with genes related to the immune system. The genes located in this cluster (class III) include several complement genes and the tumor necrosis factor alpha (*TNF-α)* gene (Horton *et al*. 2004).

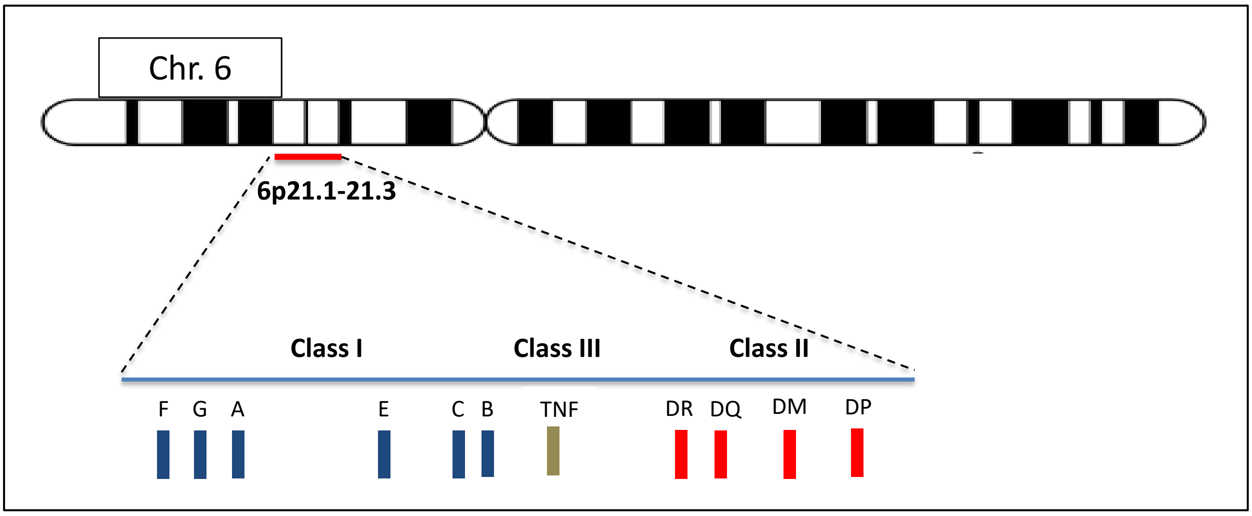

Figure 1.1: Genomic organization of the HLA loci

1.2 Structure of MHC molecules

The protein products of MHC class I & II genes (here called “molecules”) are transmembrane glycoproteins present on the surface of all nucleated cells and antigen-presenting cells (APCs), respectively. Although MHC class I and II molecules perform similar functions in antigen presentation, there are some differences in their structures. Class I MHC molecules are heterodimer proteins composed of α and β chains. The α chain is a polypeptide encoded by one of the classical class I genes (i.e., *HLA-A*, *HLA-B*, or *HLA-C*). The β chain is encoded by the β2 microglobulin gene, which is not part of the MHC loci and located on chromosome 15. The α chain itself is organized into three domains: an extracellular domain composed of α1 and α2 and part of the α3 domain, a transmembrane domain, and a cytosolic domain contained in α3. The extracellular position of the MHC proteins facilitates interaction with T cells as part of their function. Such interaction is possible due to the formation of what is called a peptide-binding pocket (also known as a peptide-binding cleft), which in class I is found at the top of the molecule between the α1 and α2 domains. Furthermore, intracellular interaction is also possible through the transmembrane part of the α3 domain. At the genomic level, the α1 and α2 domains are encoded by exons 2 and 3 in all class I genes and the α3 domain is encoded by exon 4 of these genes. On the other hand, class II molecules consist of two polypeptides, the α and β chains, each of which has two extracellular domains (α1 and α2 for the α chain and β1 and β2 for the β chain). The overall organization is mainly similar to that of class I molecules; however, in class II molecules, the two polypeptides are encoded by genes located in the MHC loci. The peptide-binding cleft in class II molecules is found between the α1 and β1 domains, and also at the top of the molecule. The α2 and β2 parts non-covalently interact and also bind to the membrane (Owen *et al*. 2013).

1.2 Function of MHC molecules

As noted previously, more than 200 genes are located in the MHC region, many of which encode proteins that function as part of the immune system. Class I and II MHC molecules principally exert the same function by presenting antigens, but these two types of molecules differ in the way they acquire these antigens in the first place and the type of cells they interact with. Class I molecules obtain the processed antigens intracellularly in the endoplasmic reticulum (Peaper *et al*. 2008). The loaded antigens are then transported to the cell membrane and presented on the molecules' surfaces to CD8$^+$ T cells. Class II molecules, however, specialize in extracellular antigens that are captured by immune cells and processed by their lysosomal machinery. Loading of the processed antigens occurs after the fusion of endosomes with other ones carrying class II molecules. The loaded antigens are finally exported to the cell membrane where they are presented to CD4$^+$ T cells (**Figure 1.2**) (Owen *et al*. 2009).

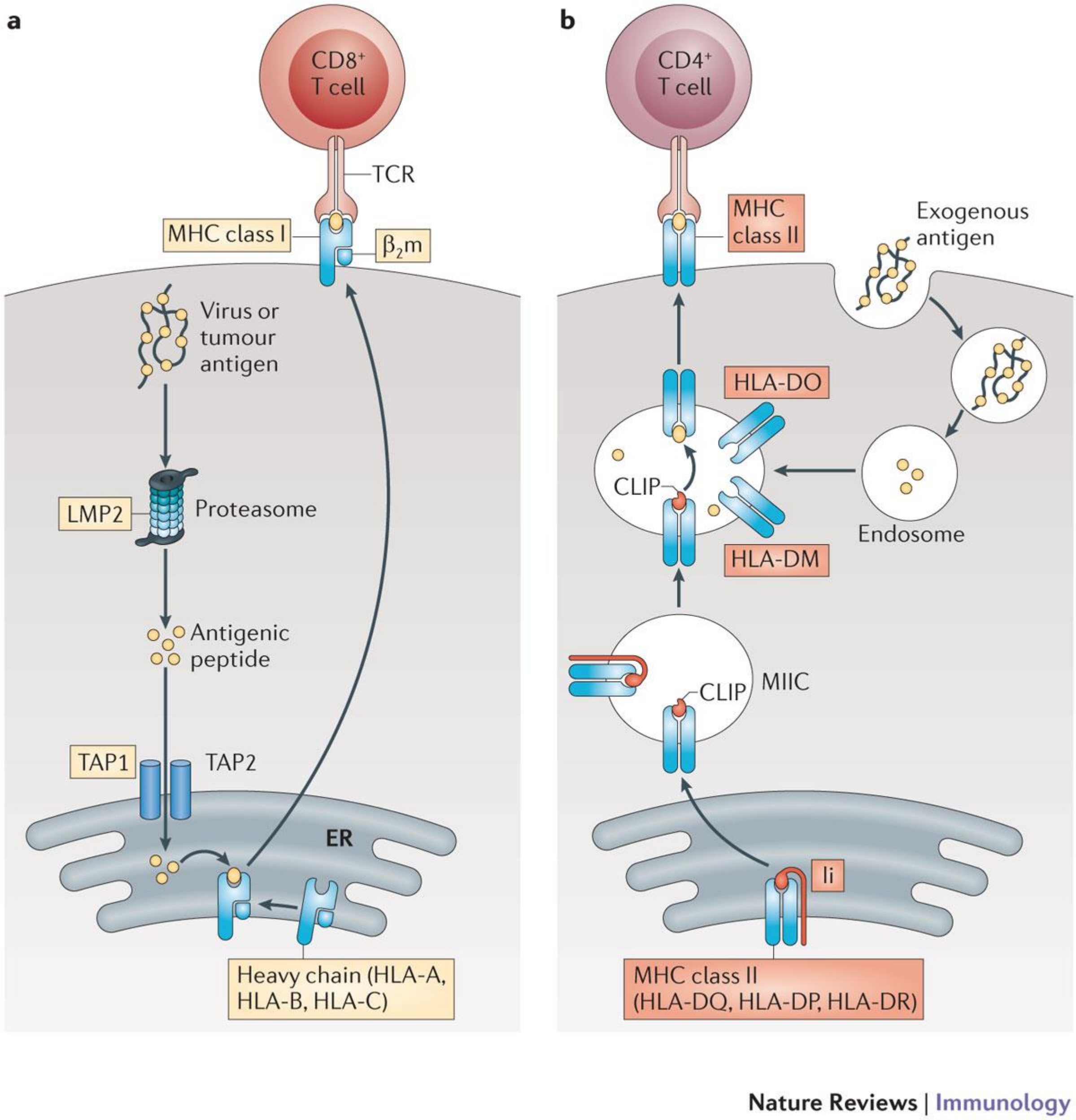

Figure 1.2: Antigen presentation pathway of MHC class I & II molecules

Antigen presentation pathways in (a) class I and (b) class II molecules. Reprinted with permission from Macmillan Publishers Ltd. [Nature Reviews Immunology] (Neefjes *et al.* 2011), copyright (2011).

1.3 HLA and disease associations

The significance of the HLA region is clearly evident from the number of disease associations that are linked to genetic variants located in the loci. This relatively small region occupies 3.6 Mb of the human genome (i.e., 0.12% of the genome), but contains 6.4% of the NHGRI-reported GWAS associations (Ripke *et al.* 2013).

Genetic variability in the HLA loci is associated with common/complex diseases. Common diseases that have been associated with HLA genes include TID, RA, and multiple sclerosis (MS). The one element shared between these diseases is the presence of what is called "autoantibodies," whose presence triggers an immune reaction against self-proteins; for this reason, the mentioned diseases are known as autoimmune diseases. One of the candidates that have been associated with RA is the HLA-DRB1 locus, although some of the reported associations were specific to populations with European ancestry (Kochi *et al.* 2010). Another class II locus (HLA-DQ) has also been linked to TID. Early studies of genetic susceptibility to TID suggested that having two haplotypes of HL-DQ loci (DQA1*03:01-DQB1*03:02 and DQA1*03:01-DQB1*03:01) confers increased risk and protection against the disease, respectively (Sanjeevi *et al.* 1995). Furthermore, recently, more DRB loci have been linked to TID, which included DRB3 (DRB3*01:01:02 and DRB3*13:01:01) and DRB4 (DRB4*01:03:01) (Zhao *et al.* 2016). The factor linking all of these diseases is the central role played by HLA genes in immunity.

HLA loci also contain several genes associated with infectious diseases. Again, the involvement of MHC molecules in antigen presentation to T helper cells places them in the center of the defense against pathogens and infectious agents. The list of such infectious diseases includes tuberculosis (TB), malaria, human immunodeficiency virus (HIV) infection, and hepatitis. Previous studies showed that individuals with heterozygote genotypes in class

II and I HLA genes have greater ability to clear infection or slow disease progression in hepatitis B virus infection and HIV/AIDS, respectively (Thursz *et al.* 1997; Carrington *et al.* 1999).

Contrary to the TID and RA diseases, in which HLA class I and II were the main candidate loci, a gene in class III (tumor necrosis factor alpha: *TNF-α*) seems to be associated with the severity of malaria in the Gambian population (McGuire *et al.* 1999), although a later GWAS failed to replicate this association (Jallow *et al.* 2009). Furthermore, the associations of HLA class I and II alleles with malaria have also been reported, although their strength is questionable (Blackwell *et al.* 2009).

1.3 Evolution of MHC loci

The question of how MHC genes evolved has yet to be fully addressed; some evidence has been presented showing that MHC genes (class I and class II) are as old as most of the other adaptive immune system genes, including T-cell receptor (TCR) genes (Flajnik and Kasahara, 2001). In the evolutionary tree, the MHC genes first appeared after the divergence between jawed and jawless vertebrates, so these genes have been identified in all major jawed vertebrates (Flajnik *et al.* 1999) (**Figure 1.3)**.

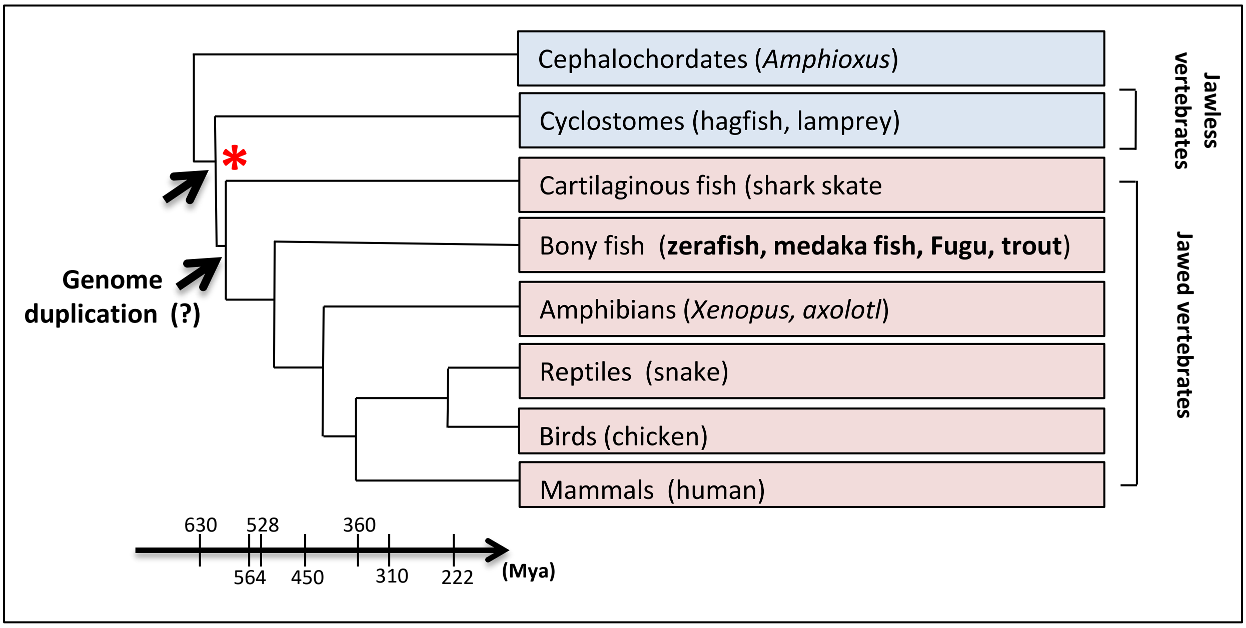
Cephalochordates (*Amphioxus*)
Cyclostomes (hagfish, lamprey)
Cartilaginous fish (shark skate
Bony fish (**zerafish, medaka fish, Fugu, trout**)
Amphibians (*Xenopus, axolotl*)
Reptiles (snake)
Birds (chicken)
Mammals (human)
Jawless vertebrates
Jawed vertebrates
Genome duplication (?)
*
630
528
360
564
450
310
222
(Mya)

Figure 1.3: Evolution of adaptive immune system.

First appearance of MHC genes in this evolutionary tree is indicated by the red asterisk at the divergence between jawless and jawed vertebrates.

Adapted from Immunity 15 (3), Flajnik *et al*., Comparative genomics of the MHC glimpses into the evolution of the adaptive immune system, 351-362, copyright (2001), with permission from Elsevier.

1.3.1 HLA genes evolve through segmental duplications and insertion/deletions

The high variability of HLA loci is featured in the over representation of segmental duplications, which are very common along the loci (Bailey *et al.* 2001) (**Figure 1.4**). Nei *et al.* (1997) studied the patterns of phylogenetic variations in the MHC class I genes of many vertebrates and found that these patterns fit the birth and death model of evolution more than the concerted model one. In the birth and death model, gene families as in class I genes evolve mainly through duplications, which lead to the emergence of new genes or sometimes the disappearance of others due to the accumulation of deleterious mutations. As they showed, many MHC class I genes in humans and mice have experienced duplication events or became non-functional (pseudogenes). Furthermore, Anzai *et al*. (2003) compared 1.7 Mb of human and chimpanzee MHC loci and found a large 95-kb deletion in the chimpanzee sequence, which led to the chimpanzee having a single *MIC* gene. Because the detected deletion seemed to be in a functionally significant position, the authors postulated that INDELs have probably played a role in the MHC evolution of primate species.

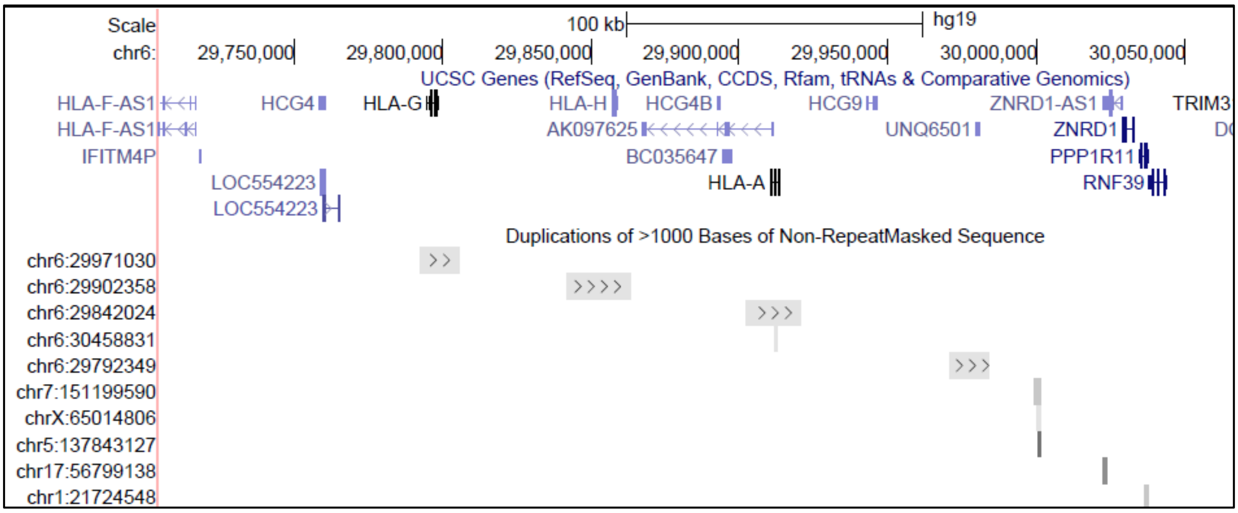

Figure1.4: **A screen shot from UCSC genome browser showing segmental duplication in HLA class I region**

The grey rectangles showing duplications extending in regions include HLA-A, HLA-H, and HLA-G. Segmental duplications data are based on Bailey *et al.* (2001)

1.3.2 Natural selection in the HLA loci

Selection acts on a genetic locus in several ways. For example, an increase or decrease in the specific allele frequency at a locus could be interpreted as directional selection of that allele; whether it is considered to be positive or negative selection depends on the direction of the shift in allele frequency. In some other cases, the change in allele frequency takes the form of an increase in allele number, with many alleles maintained at intermediate frequencies. The result of this general increase in allele frequencies is also an increase in heterozygosity at the population level. This last pattern of selection is known as overdominant selection (also called balancing selection).

The driving force in the evolution of MHC genes is probably highly relevant to the functions exerted by their encoded proteins. It is generally believed that MHC genes evolved as a result of interaction between the host (organism) and infectious pathogens.

1.3.1.1 Balancing selection as a driving force for HLA gene evolution

In almost all populations, it has consistently been observed that HLA genes have a high level of heterozygosity (Solberg *et al.* 2008). Based on this observation, it was proposed that balancing selection maintains this increased level of heterozygosity, although other types of selection, such as frequency-dependent selection, cannot be ruled out. Regarding the mechanism underlying this maintenance, it was suggested that individuals with heterozygote genotypes might have better fitness than those with homozygote ones. This is because heterozygote genotype carriers have greater capacity to recognize and present antigenic peptides to the immune system, and thus have a wider pathogen-combating capacity than individuals with homozygote genotypes. The increased fitness of heterozygote genotypes is called the "heterozygote advantage," which represents the basic concept of the heterozygote advantage model. Although this heterozygote advantage hypothesis is widely accepted, validating it experimentally is not easy. Penn *et al*. (2002) tested this hypothesis on mouse strains after challenging them with different bacterial strains. Their findings indicated a difference in survival and infection clearance between heterozygote and homozygote mice, with the former having a higher survival rate and greater likelihood of clearing the infection.

Apart from the observation of increased heterozygosity, studies of DNA sequence variations also provided clues regarding balancing selection acting on HLA loci. Hughes and Nei (1988) studied the pattern of nucleotide substitutions in the antigen recognition site (ARS) of HLA class I genes. They first speculated whether overdominant selection is acting on these genes; then, the ratio of nonsynonymous substitutions (d_N) to synonymous substitutions (d_S) in the ARS would be high. Interestingly, their findings came to support this hypothesis, so they concluded that overdominant selection plays a major role in the evolution of HLA genes.

Recently, a different viewpoint was proposed by Lau *et al.* (2015), who studied balancing selection using the HLA-DRB1 locus as a model. Based on their previously established phylogenetic analysis of the PBR, they found that DRB1 alleles cluster into two groups, only one of which supports the heterozygote advantage model. They suggested that the divergent allele advantage model is unlikely to explain all of the diversity in the HLA-DRB1 locus and other mechanisms might also be working in combination.

Evidence for other types of selection has also been documented. Directional selection of a specific allele is expected to increase or decrease the frequency of that allele in the population. As a consequence of such selection, a favorable outcome might occur if the selected allele confers protection from certain diseases or vice versa if the allele increases the risk of diseases. An interesting example of such selection is the strong selection of HLA-DPB1*04:01, which was shown to have reached an allele frequency of 6.1% in the Japanese population, although evidence of its association with a specific phenotype/disease is still lacking (Kawashima *et al.* 2012).

1.4 Nomenclature of HLA alleles/genes

Dealing with the complexity of HLA loci necessitated an organized effort to cope with it. The World Health Organization therefore established a designated committee for managing and maintaining the HLA system nomenclature. This committee is currently known as the WHO Nomenclature Committee for Factors of the HLA System. Criteria for naming HLA alleles depend on several factors regarding the PBR's serological activity, sequence length considered, and the type of variations in this sequence. The name of an HLA allele starts with "HLA" followed by a hyphen and capital letters to indicate the gene name (**Figure 1.5**). The gene name is followed by a separator "*" and then four sets of digits separated by a colon ":" The first set of digits represents the allelic group, which is determined by the antigenic reactivity. The second set of digits corresponds to the specific HLA protein and any two alleles differing in this set must have at least one missense variant. On the other hand, synonymous variants are represented by the third set of digits. The fourth and last set of digits distinguishes alleles with variants in the untranslated regions of the gene (i.e., introns, 5′ and 3′ untranslated regions). An additional suffix is sometimes added at the end of the allele name to give additional information; for example, a suffix with the letter N indicates that this allele is not expressed (also called "Null") (Marsh *et al.* 2010). As of October 2016, there are 15,635 HLA alleles registered in the IPD-IMGT/HLA database (**Figure 1.6**).

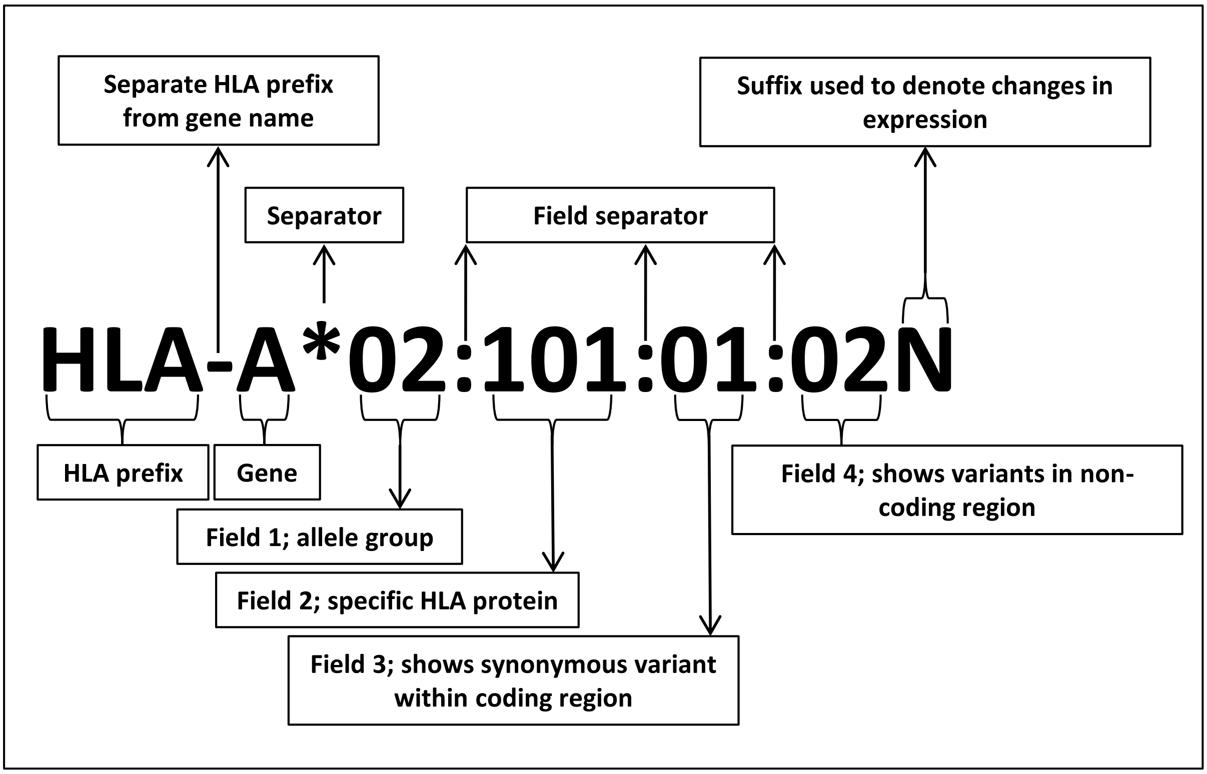
Separate HLA prefix from gene name
Suffix used to denote changes in expression
Separator
Field separator
HLA-A*02:101:01:02N
HLA prefix
Gene
Field 1; allele group
Field 2; specific HLA protein
Field 3; shows synonymous variant within coding region
Field 4; shows variants in non-coding region

Figure 1.5: Criteria for nomenclature of HLA alleles/genes (March *et al.* 2010)

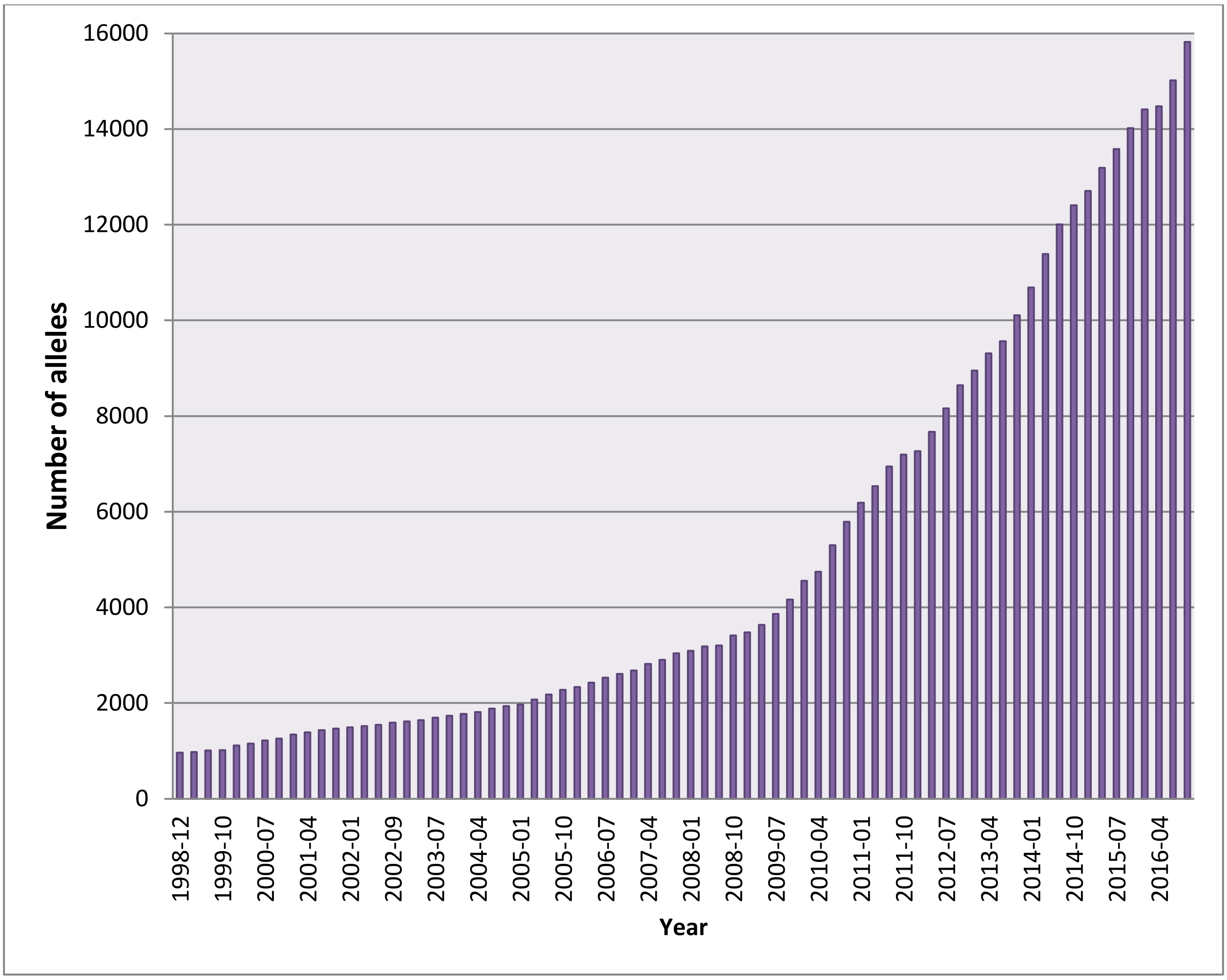

Figure 1.6: Number of HLA class I & II alleles in the IPD-IMGT/HLA database (October 2016)

1.4 HLA typing

The increased interest in HLA typing in the last decade is due to the high demand for tissue matching between donors and recipients in transplantation therapy. This is supported by the large number of new HLA alleles that were found based on bone marrow donor registries around the world. Interestingly, a single study conducted on potential stem cell donors found 2127 new HLA alleles in class I genes (Hernandez-Frederick *et al.* 2014), which clearly indicates how much effort will be required to obtain all of the variability in the HLA genes.

1.4.1 HLA typing methods

HLA typing methods can be classified into two main categories based on the type of molecules used in the test. The first category is based on serological identification of the HLA allelic group, which is based on antigen-antibody reaction. This is the first approach used for HLA typing and, because the test is based on cellular interaction, only the functional part of the protein is targeted. For this last reason, this approach only captures variability in the peptide-binding region, which is why the resolution of typing is considered to be low (defined at two digits). More recently, with the advancement of molecular biology techniques, especially after the establishment of PCR, sequence-based typing started to be used extensively, which is the second category here. Sequence-based typing (SBT) includes several methods that all share the use of a DNA sample as the test material; therefore, the objective of this approach is to identify variations in a DNA sequence. Because all SBT methods use DNA as the starting material, they also share a PCR amplification step at the beginning. If the primers used in this amplification are specific to some alleles, then typing could be achieved in this step (this is known as PCR-sequence-specific primer, PCR-SSP). In another method called PCR-sequence-specific oligonucleotide probe (PCR-SSO or PCR-SSOP), the PCR amplification product is immobilized onto a nylon membrane, followed by hybridization to a specific labeled probe. Detection of the signal from the bound probe will determine the sample’s genotype. The target regions in both PCR-SSP and PCR-SSO are exons 2 and 3 in class I genes and exon 2 in class II genes; hence, the typing of these methods is at four-digit resolution. A third SBT method is based on PCR followed by Sanger sequencing, which has the capacity to determine the full sequence of the HLA gene of interest, permitting a maximum typing resolution of eight digits.

1.4.2 Ambiguity in HLA typing

The complexity of HLA loci (e.g., homology between loci) and the highly polymorphic nature of some genes may make the identification of unique genotypes a challenging task. Without exception, all of the HLA typing methods sometimes provide ambiguous genotyping results and such ambiguity is related to the principle of the method itself. In serology-based typing, because only PBR is considered, several alleles share the same antigenic epitope; therefore, they are placed in the same allelic group. The problem, however, is that distinguishing between these alleles is essential for tissue matching in transplantation because they are functionally different and might trigger an immunological response and increase the risk of graft rejection. Sequence-based typing methods targeting specific exons such as PCR-SSP and PCR-SSO also have ambiguous typing when more than one allele shares the same sequence in the targeted exons. In such cases, it is necessary to include additional exons to obtain a unique typing result, which obviously requires further laboratory work. Finally, although Sanger sequencing can achieve the best typing resolution up to eight digits, it is also not free from ambiguity in some cases. Here, reading the chromatogram in positions that have excessive heterozygosity is very difficult because traces of fluorescence signal overlap with each other, preventing a unique nucleotide call. As a consequence of such ambiguity, identification of the two HLA alleles will not be possible because the sequence of each allele cannot be uniquely identified: a problem known as haplotype phasing problem. Dealing with such ambiguity requires experienced personnel, if not additional experiments.

1.4.3 HLA typing using next-generation sequencing

Next-generation sequencing (NGS) has had a tremendous impact on many fields in biology and medicine and is increasingly being used for HLA typing. Furthermore, several NGS-based protocols have recently been developed, which fall into two main categories: methods based on PCR amplification followed by MiSeq and methods based on capture and analysis of HLA sequence data.

1.5 HLA and genetic diversity in human populations

There are several approaches to studying genetic diversity in a population. One of these is to use genetic markers that are highly polymorphic in the population under study. The fact that HLA genes are exceptionally diverse among populations worldwide makes them good candidates for genetic diversity studies. Although other markers such as mitochondrial DNA, the Y chromosome, and genome-wide SNP data are also used, the relevance of HLA goes beyond the issue of diversity to include disease associations, drug effects, and transplantation therapy.

1.5.1 Genetic diversity of East African populations

Until recently, the genetic diversity of African populations had not received much attention in genetic studies addressing global human diversity. Thanks to the efforts of Dr. Tishkoff and her group, the great level of diversity in this part of the world has been revealed. The African origin model of modern human evolution suggests that all non-African groups are descendants of a small group that left Africa probably more than 40,000 years ago (Relethford 2001). Haplotype analysis of autosomal genes (*CD4*, *DM1*, and *PLAT*) in different populations showed that non-African groups, in addition to their shared patterns of variation, also share a subset of variations with East African groups from Ethiopia (Tishkoff *et al*. 2002). This suggests that migration of early modern humans took a route from East Africa and a group of early East Africans, who migrated out of Africa, could be the ancestors of all current non-African populations. Evidence from the non-recombining region of the Y chromosome also showed that current haplogroups from East Africa (including Sudan and Ethiopia) and the Khoisan of South Africa are the most ancestral in the world (Underhill *et al*. 2000). Moreover, East African groups have the greatest level of regional genetic substructure among African populations (Tishkoff *et al.* 2009).

1.5.2 History, linguistic, and genetic diversity of Sudanese population

Sudan is a country in northeast Africa located in 15°00 N and 30°00 E coordinates. The history of Sudan is largely undocumented and sources on its early history are limited and mostly from elsewhere, such as Egypt. The name "Sudan" comes from the Arabic word "Sud," which means black people; it was used to describe the area south of Aswan City where the Nubian people currently live. Prehistorically, the land north of the first cataract of the Nile was known as the "Cush" (currently north Sudan). From 800 BC and several centuries thereafter, Cush was a powerful kingdom that ruled north Sudan and sometimes extended its political influence over Egypt (Metz 1991).

Linguistically, Sudan is one of the most diverse countries in Africa. According to Ethnologue (Ethnologue: Languages of the world, http://www.ethnologue.com), there are currently 76 living languages in Sudan, although many of them are classified as threatened or endangered (Lewis *et al.* 2016). Despite the huge linguistic diversity seen in Sudan, Sudanese languages fall into three main linguistic families (**Figure 1.8**): Afro-Asiatic, Nilo-Saharan, and Niger-Congo/Kordofanian (Greenberg 1963). The most commonly spoken language is Arabic, which is currently the official language and belongs to the Afro-Asiatic family. This family is widely spread across the northern and eastern parts of the country and it includes Arabic and Bedawiyet (spoken by Beja). Furthermore, the Afro-Asiatic family also extends to neighboring countries such as Eritrea and Ethiopia, as well as North Africa. The second family is Nilo-Saharan, which has a vast representation in Sudan; in fact, most Sudanese languages are categorized into this linguistic family (Greenberg 1963). Languages of the Nilo-Saharan family include those spoken by Nubians in the north, Darfurians in the west, and Nilotes in the south (which is now a separate country). The last major linguistic family in Sudan is the Niger-Congo family (also called Niger-Kordofanian), which is, in contrast to the Nilo-Saharan family, restricted to a few groups from the Nuba mountains in the south.

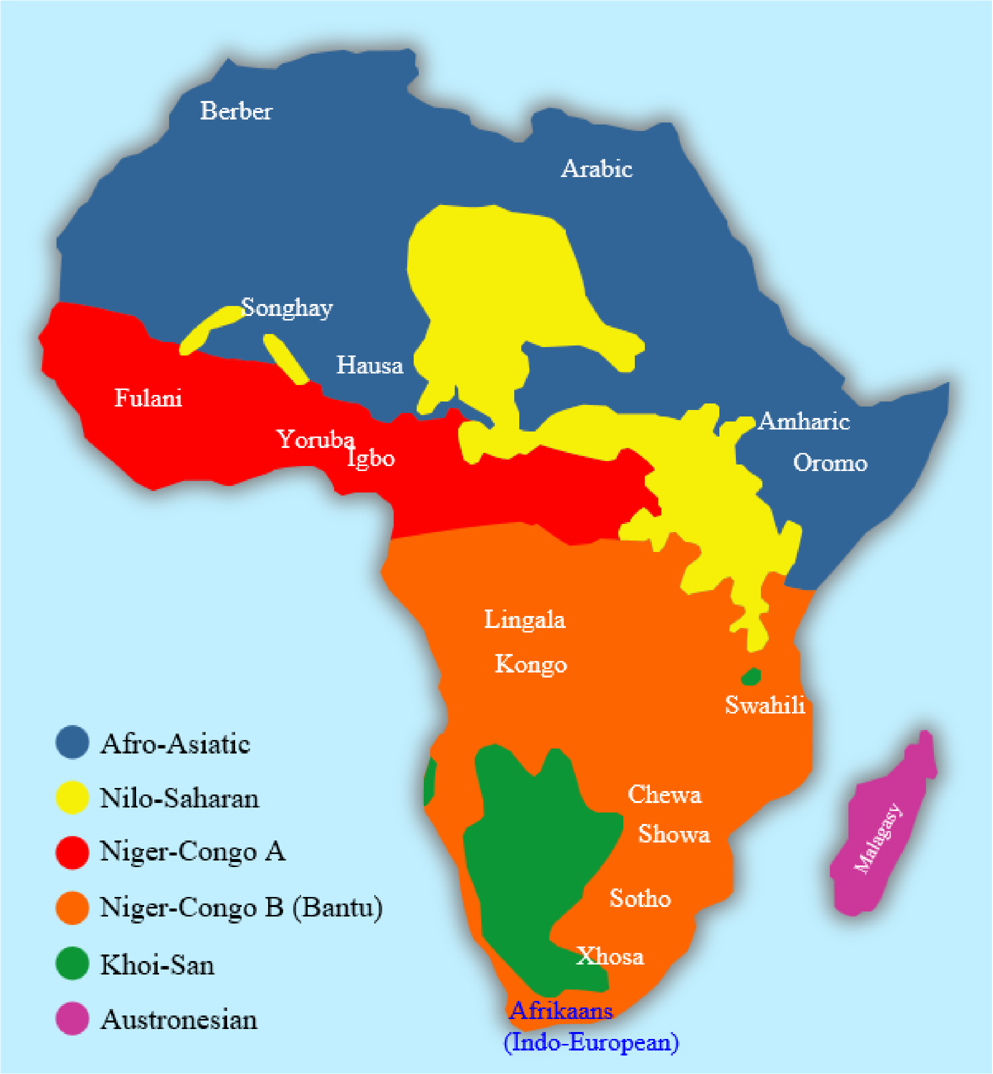

Figure 1.7: Major linguistic families in Africa

Classification is based on Greenberg (1963). Downloaded from (https://commons.wikimedia.org/wiki/File:African_language_families_en.svg). Permission to reprint is obtained under GNU Free Documentation License.

The genetic diversity of Sudanese groups was basically unknown until recently. However, lately, studies based on Y-chromosome variations and SNP data were conducted and provided valuable information for understanding the genetic diversity in Sudan. A study of Y-chromosome variations in different Sudanese groups found haplogroups A, B, and E, which are found mainly in Africa, are very common among Nilo-Saharan-speaking groups such as Nilotes and Fur (Darfurians) (Hassan *et al*. 2008). In this study, the same groups exhibited limited evidence for gene flow from outside Africa. Paradoxically, the genetic diversity of groups inhabiting north and east Sudan seems to have been shaped by waves of migration and invasion that passed through their areas. This is evident from the high frequency of haplogroup J-12f2 observed in almost all Afro-Asiatic groups and surprisingly the Nilo-Saharan Nubian group. It is largely accepted that this haplogroup originated in West Asia, which explains its high frequency among Middle Eastern and European populations (Hassan *et al*. 2008). Moreover, recently, genome-wide SNP data in several Sudanese groups identified a Nilo-Saharan component that was not seen in Tishkoff's study of 2009, adding one additional component to the 14 ones identified in African populations by her group (Dobon *et al*. 2015, Tishkoff *et al.* 2009). Furthermore, the deep history of Sudanese and East Africans was revealed by studying the mitochondrial cytochrome C oxidase subunit II (*MT-CD2*) gene, which indicated a large effective population size (N_e) for populations in East Africa (Elhassan *et al*. 2014). A phylogenetic analysis based on the same data showed that East Africans are the closest groups to the root of the human evolutionary tree, clearly pointing to East Africa as a potential site for the origin of modern humans.

Rationale

The HLA genes have been extensively studied in many populations around the world; however, the HLA genes of African populations have not received the same amount of study (Appendix **Figure A.1**). Several factors are responsible for this, including a lack of established transplantation programs. Sudan is one of the African countries affected by this situation. Despite the diversity of the Sudanese population, little is known about the distribution of HLA genes in this country. Few studies have been conducted, reflected in there being only five HLA data sets from Sudan in the AFND database (**Table 1.1**). Furthermore, Dafalla *et al*. (2011) recently studied the HLA diversity of samples from renal transplant donors from north Sudan. All of these previous studies either had a low typing resolution or did not include a wide representation of the Sudanese population in addition to some of them were never published. Therefore, a comprehensive map of the diversity of HLA genes would serve several purposes.

Furthermore, the diversity of the Sudanese population has recently been studied by Dobon *et al.* (2015) using genome-wide SNP data, which revealed two main substructures and a shared genetic component with North African and Middle Eastern populations from the Arabian Peninsula (Dobon *et al*. 2015). This shared component was found in the Sudanese Afro-Asiatic groups and Ethiopians.

In light of the previous points, the objectives of this study are:

- To study HLA class I diversity in seven Sudanese ethnic groups and samples from the neighboring country of Ethiopia.
- To establish a wide map of HLA allele distribution in Sudan.
- To identify HLA class I haplotypes that are common among the study groups.

- To examine the genetic structure of Sudanese groups based on HLA data and identify further substructures if found.

Table 1.1: Previous HLA studies conducted in Sudan

Gene	Study	Sample size	Year	Reference	Ethnic group	Study place
DRB1 & DPB1	Disease association	157	1992	Magzoub *et al.*	mixed	north
A, B, C, & DRB1	Anthropology	36	1997	12th IHIWS*	mixed	north
A, B, C, & DRB1	Anthropology	27	1995	12th IHIWS*	Rashaida	east
A, B, C, DPB1, DQB1, & DRB1	Disease association	200	2006	AFND§	mixed	north
A, B, C, & DRB1	Anthropology	46	1995	AFND§	Nuba	South

* IHIWS, International HLA and Immunogenetics Workshop. §AFND (González-Galarza *et al*. 2015), the two data sets are unpublished and were directly submitted to the AFND.

Chapter Two

MATERIALS & METHODS

2.1 Sample information

Saliva samples from 329 individuals were collected between July-September 2010 from different geographical areas, representing groups from Sudan, South Sudan, and Ethiopia. All of sample collection was done by Dr. Hisham Y. Hassan from the University of Medical Science and Technology (UMST). I personally participated in the collection of samples from three groups (Beja, Gaalien and Shokrya). The South Sudanese samples were collected before the separation of Sudan into two countries and the Ethiopians samples were collected from people who currently live in Sudan. In this study, groups were defined based on the linguistic classification of African populations. Three of the major linguistic groups known in Africa are included in these samples: Afro-Asiatic (Beja, Gaalien, Shokrya, and Ethiopians), Nilo-Saharan (Nubians, Nilotes, and Darfurians), and Nilo-Saharan/Niger-Kordofanian (Nuba). The ethnic group affiliations for all individuals were self-reported, which along with their linguistic classification and sample collection locations are shown in **Table 2.1**. Furthermore, the geographic locations of sampling sites are also shown in **Figure 2.1**. Both sample collection and DNA isolation were performed using the Oragene™ collection kit (OG-500) (DNA Genotek, Ontario, Canada), following the recommended protocol. In brief, participants were first asked to wash their mouth with water for any food remnants then 10ml of saliva were collected in collection tube. DNA was isolated by taking 500µl from the collected sample in 1.5ml tube, gently mixing the tube and then incubation in a water bath at 50°C for 60 minutes. The sample was transferred to 1.5ml tube and 20µl of PT-L2P buffer were added and the tube was mixed by vortex for several seconds, incubated on ice for 10 minutes and then centrifuged for 5 minutes at 15,000g. All of the supernatant was then transferred to fresh 1.5ml tube and 600µl of 95% ethanol were added and then mixed by inversion 10 times. After allowing the DNA to precipitate for 10 minutes, the tube was centrifuged for 2 minutes at 15,000g and the supernatant was discarded. The sample was washed using 70% ethanol before drying the tube and adding 100µl of TE to dissolve the DNA pellet. The isolated DNA was stored at -20°C till used for the experiments.

Ethical considerations were made when collecting the samples, as all of the participants gave informed consent and their anonymity was protected all the time. In addition, the study protocol was approved by the ethical committee of Sudan Medical and Scientific Research Institute, University of Medical Science and Technology, Khartoum, Sudan (SUM 2010/7), and the Ethics Committee of the National Institute of Genetics, Mishima, Japan (nig1508, 2015.11.30).

For the sake of comparing HLA genetic diversity between populations in this study and other African and Middle Eastern populations, I downloaded data from the Allele Frequency Net Database (AFND) (González-Galarza *et al.* 2015). The downloaded data included populations from Sub-Saharan Africa (West and Central Africa), the Middle East, and two European populations (in total, I downloaded 30 data sets and the total sample size for all populations was 9119 samples). Details of these downloaded data are shown in appendix **Table A.1**. As a measure quality, I excluded any data set when the total allele frequency did not add up to one. Although the only data set I found from Saudi Arabia did not meet the last condition, I retained this data set as it is the only one available for comparing the study populations to Arabs from the Middle East.

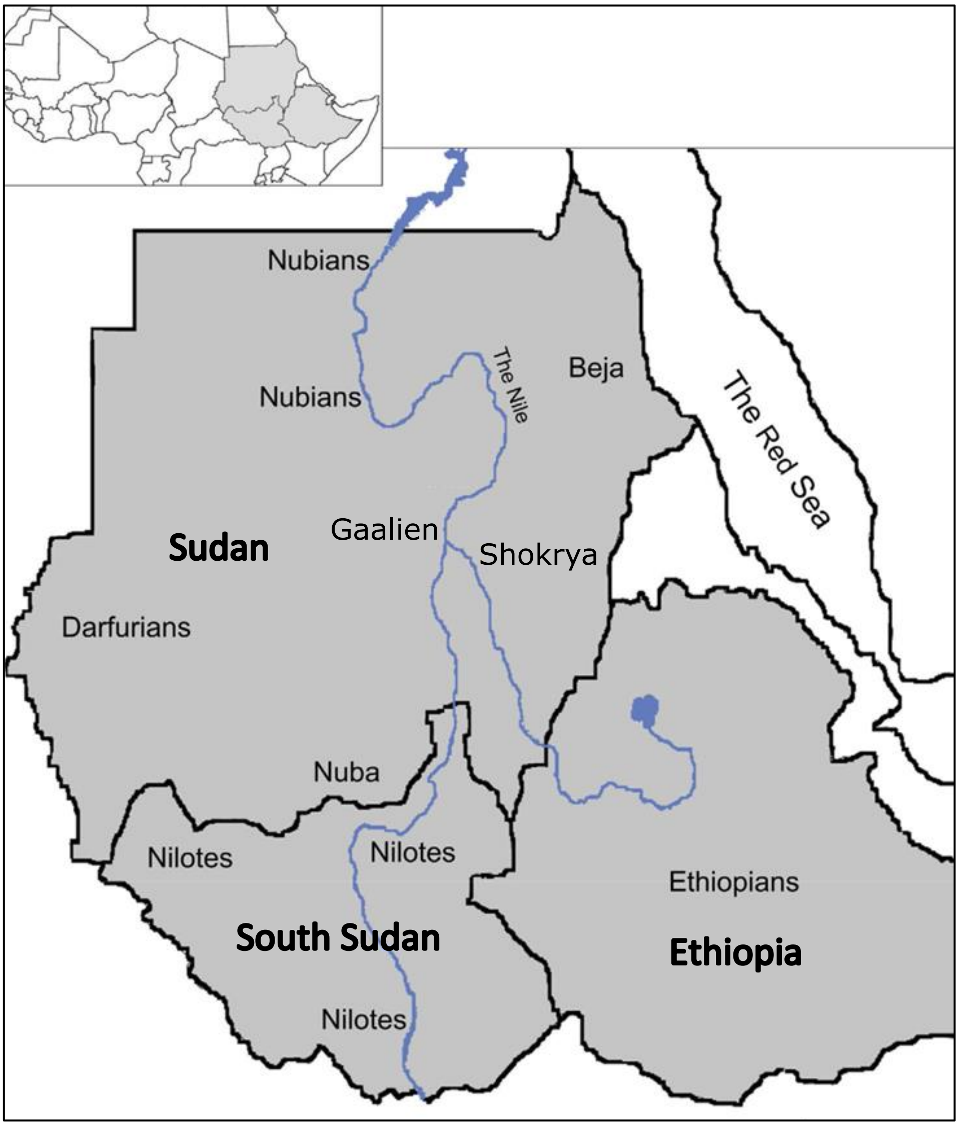

Figure 2.1: Approximate locations of populations in this study

The inset at the top left shows the locations of Sudan, South Sudan and Ethiopia

Table 2.1: Geographic distribution and linguistic affiliation of 8 East African populations

Ethnic group	**Size**	**Socio-economical activities**	**Country**	**Sampling location**	**Coordinates**	**Linguistic family**	Linguistic subfamily
Gaalien	40	Agriculturist	Sudan	Shendi	16N 33E	Afro-Asiatic	Semitic
Shokrya	40	Pastoralist	Sudan	New Halfa	15N 35E	Afro-Asiatic	Semitic
Beja	40	Pastoralist	Sudan	Sinkat	18N 36E	Afro-Asiatic	Cushitic
Ethiopian	40	Agropastoralist	Ethiopia	Khartoum	15N 32E	Afro-Asiatic	Cushitic/Semitic
Nubian	40	Agriculturist	Sudan	Wadi Halfa	21N 31E	Nilo-Saharan	Eastern Sudanic
Darfurians	40	Agriculturist	Sudan	El-Fashir	13N 25E	Nilo-Saharan	Eastern Sudanic, Fur, Mapan, and Saharan
Nuba	40	Agropastoralist	Sudan	Kadugli	11N 29E	Nilo-Saharan/Niger-Congo	Kordofanian, Kadugli-Krongo, and Eastern Sudanic
Nilotes	49	Pastoralist	South Sudan	Juba	4N 31E	Nilo-Saharan	Central Sudanic, Eastern Sudanic

N, number of collected samples

2.2 Targeted sequencing of HLA genes

2.2.1 NGS library preparation and target enrichment protocol

The NGS library preparation was based on two main principles; the first is pooling of multiple samples using unique combination of indices for each one of them and the second is capturing the target region of interest using hybridization probes labelled with magnetic beads. The first part of library preparation was done using SureSelectQXT Library Prep Kit (Agilent Technologies) and last part of target enrichment was done using SeqCap EZ choice system (Roche Diagnostics). The full NGS library protocol is as described by Ahmadloo *et al*. (in press), but I will also give the details here:

2.2.1.1 Fragmentation and purification

I first measured the concentration of the isolated DNA using Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific Inc.) on FilterMax F5 Multi-Mode Microplate Readers (Molecular Devices). Then I adjusted the concentration of each DNA sample to a final concentration of 20ng/µl. For each of the 328 samples, 20ng of the adjusted DNA were placed into a well of 96-well plate and 1 µl of 0.5x diluted SureSelect QXT Buffer in addition to 8 µl of SureSelect QXT Enzyme Mix were added. The plates were then sealed and vortexed thoroughly for 20 seconds before spun down. The sealed plated were incubated in a thermal cycler using the following protocol; 10 minutes at 45ºC, 1 minute at 4ºC and final hold at 4ºC. Following that a 16 µl of SureSelect Stop Solution were added to stop the reaction.

The products from previous steps are fragmented DNA, which were then purified by incubation with magnetically labeled AMPure XP beads (Beckman Coulter). The purification process included incubation with AMPure XP beads for 5 minutes and several washing steps with 80% ethanol. In the final step plates were placed on a magnetic stand and the supernatant was discarded to dry up the wells before adding 10 µl of nuclease-free water and sealing the plates.

2.2.1.2 Amplification and adapter ligation

The fragmented and purified DNA was subject to PCR amplification and indices ligation. The PCR reaction contained 5 µl Herculase II Reaction Buffer, 0.5 µl Herculase II Fusion DNA Polymerase, 0.25 µl 100 mM dNTP Mix, 1.25 µl DMSO (100%) and 6 µl PCR-grade water. For each well in the 96-well plate, I added 13 µl of the prepared mix plus 1 µl of dual index primers that are specific to each sample. Therefore, each sample can be identified after pooling and sequencing reaction using the unique set of two indices. The prepared plates were placed in a thermal cycler and the following program was used;

First incubation: 2 minutes at 68ºC

Denaturation: 2 minutes at 98ºC

7 cycles of:

30 seconds denaturation at 98ºC

30 seconds primer annealing at 57ºC

1 minute primer extension at 72ºC

Final incubation for 5 minute at 72ºC

Hold at 4ºC

Following the amplification, the PCR products were purified using the previously mentioned purification steps and in the last step the products were dissolved in 10 µl of nuclease-free water and transferred to a new plate.

Up to this step, the prepared DNA library contained the entire genomic DNA of the sample and to capture the target regions of interest (i.e. HLA loci) I used a hybridization based protocol. Furthermore, to reduce the cost and labor I pooled all of 96 samples together in a single tube, an approach previously established in our lab and proved to be efficient and cost-effective (Hosomichi *et al*. 2013). For the purpose of pooling, I measured DNA concentration of the library and took equal amount of DNA from each well to get a final concentration of 1 µg from the pre-capture library.

2.2.1.3 Hybridization step SeqCap EZ protocol

The entire pooled library was used for the hybridization reaction, which was done following these steps:

Hybridization enhancing (HE) oligo was prepared and 2 µl (2 pmol) were put into new tube and then added 5 µl of COT Human DNA (1 mg/ml) (Sigma-Aldrich).

The tube content was evaporated using Centrifugal Concentrator CC-105 (TOMY Digital Biology).

A master mix of 7.5 µl of 2x Hybridization Buffer and 3 µl of Hybridization Component A was prepared, mixed and 10.5 µl of it were transferred to a new tube.

To the master mix, I added 4.5 µl of SeqCap EZ probe and vortexed then spun down.

The final mixture was placed on the thermal cycler and ran the following PCR program:

Denaturing at 95ºC for 5 minutes

Incubation at 47ºC for 20 hours

The post-capture library was washed using the following steps:

A 100 µl of room temperature adjusted Dynabeads M-270 Streptavidin (Thermo Fisher Scientific) were put into new tube and placed on a magnetic stand for 5 minutes till the solution became clear; then the supernatant was discarded.

The beads were washed 3 times by placing out of the magnetic stand, adding 200 µl of 1x Beads Wash Buffer, mixing thoroughly by vortex, placing on the magnetic stand again till a clear solution was obtained and finally discarding all the supernatant.

To the beads containing tube, 15 µl of post-capture library were added and thoroughly mixed by pipetting.

The tubes were incubated in a thermal cycler for 45 minutes at 47ºC with intermittent vortexing every 15 minutes.

After removing the tubes from the thermal cycler, 100 µl of 1x Wash Buffer I were added and the tubes were mixed then placed on a magnetic stand till a clear solution was obtained, which was then discarded.

The captured library was washed twice by adding 200 µl 1x Stringent Wash Buffer, mixing and incubation in a thermal cycler at 47ºC for 5 minutes. The tubes were then placed on a magnetic stand to capture all the beads and the clear supernatant was discarded.

Finally, the library was washed 3 times using Wash Buffers I, II and III respectively. These washing steps involved adding 200 µl of the buffer, mixing thoroughly and placing the plate on a magnetic stand till the solution became clear, which was later discarded. The final obtained library was dissolved in 20 µl of PCR-grade water.

2.2.1.4 Amplification of post-capture library

The post-capture library was amplified using KAPA HiFi HotStart ReadyMix (Kapa Biosystems). To do that, I prepared a master mix, which contained 50 µl of 2x KAPA ready mix, 2 µl of 100 µM TS-PCR oligo 1 & 2 and 26 µl of PCR-grade water. The master mix was added to 20 µl from the post-capture library and amplified on the thermal cycler using the following program:

Initial denaturing at 98ºC for 30 seconds

18 cycles of:

Denaturation at 98ºC for 10 seconds

Primer annealing at 60ºC for 30 seconds

Primer extension at 72ºC for 30 seconds

Final extension at 72ºC for 5 minutes

Final hold at 4ºC

The amplified post-capture was purified using AMPure XP beads as previously described in section 2.2.1.1 of the QXT library preparation. After washing, the library was dissolved in 52 µl of nuclease-free water. The quality of the final library was checked using DNA 1000 kit on Bioanalyzer 2100 and Qubit (BR).

2.2.2 Sequencing of the prepared library

The prepared DNA libraries for all samples (328 samples pooled in four 96-well plates) were sequenced using Illumina MiSeq® platform. Before doing the sequencing, I adjusted the libraries' concentrations to 4 nM; then using 5 µl of that, I denatured it by adding 5 µl of 0.2 N NaOH, mixed and centrifuged at 280×g for 1 minute and incubated for 5 minutes at room temperature. In the final steps, the libraries were diluted twice using pre-chilled HT1; the first time to 20 pM by adding 990 µl of HT1 and the second time to 12 pM by combining 360 µl of the 20 pM library to 240 µl of the pre-chilled HT1. Finally the diluted DNA libraries were loaded into flow cells of MiSeq® reagent cartridge along with custom sequencing primers provided in SureSelect QXT Library Prep Kit and Illumina TruSeq primers. I designed the sequencing protocol to be paired-end type with forward read length of 350bp and reverse one of 250bp, so the expected insert size between the read pairs was around 600bp.

2.3 Bioinformatics analysis

The data were generated from the MiSeq® platform in “fastq” format, which is the standard format currently used in NGS technology. A fastq file contains the sequence generated from the run in the form of short sequences, called reads, and some quality information such as base qualities. An overview of the NGS data analysis is presented in **Figure 2.2**.

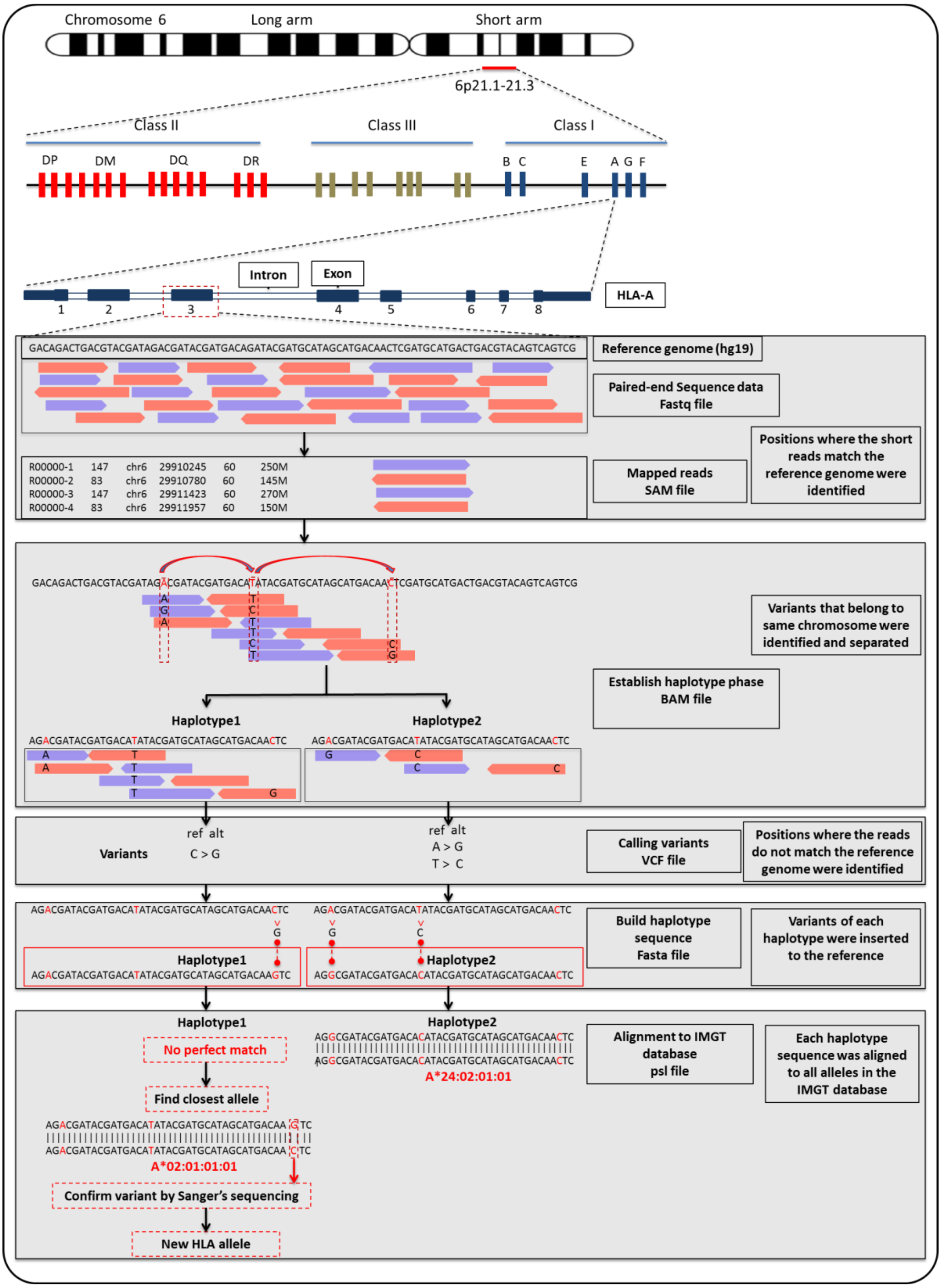
Chromosome 6
Long arm
Short arm
6p21.1-21.3
Class II
Class III
Class I
DP
DM
DQ
DR
B
C
E
A
G
F
Intron
Exon
HLA-A
1
2
3
4
5
6
7
8
GACAGACTGACGTACGATAGACGATACGATGACAGATACGATGCATAGCATGACAACTCGATGCATGACTGACGTACAGTCAGTCG
Reference genome (hg19)
Paired-end Sequence data Fastq file
R00000-1 147 chr6 29910245 60 250M
R00000-2 83 chr6 29910780 60 145M
R00000-3 147 chr6 29911423 60 270M
R00000-4 83 chr6 29911957 60 150M
Mapped reads SAM file
Positions where the short reads match the reference genome were identified
GACAGACTGACGTACGATAGACGATACGATGACATATACGATGCATAGCATGACAACTCGATGCATGACTGACGTACAGTCAGTCG
A
G
A
T
C
T
T
C
T
C
G
Variants that belong to same chromosome were identified and separated
Establish haplotype phase BAM file
Haplotype1
Haplotype2
AGACGATACGATGACATATACGATGCATAGCATGACAACTC
AGACGATACGATGACATATACGATGCATAGCATGACAACTC
A
T
A
T
T
T
G
G
C
C
C
Variants
ref alt
C > G
ref alt
A > G
T > C
Calling variants VCF file
Positions where the reads do not match the reference genome were identified
AGACGATACGATGACATATACGATGCATAGCATGACAACTC
AGACGATACGATGACATATACGATGCATAGCATGACAACTC
G
G
C
Haplotype1
AGACGATACGATGACATATACGATGCATAGCATGACAAGTC
Haplotype2
AGGCGATACGATGACACATACGATGCATAGCATGACAACTC
Build haplotype sequence Fasta file
Variants of each haplotype were inserted to the reference
Haplotype1
No perfect match
Find closest allele
AGACGATACGATGACATATACGATGCATAGCATGACAAGTC
AGACGATACGATGACATATACGATGCATAGCATGACAACTC
A*02:01:01:01
Confirm variant by Sanger's sequencing
New HLA allele
Haplotype2
AGGCGATACGATGACACATACGATGCATAGCATGACAACTC
AGGCGATACGATGACACATACGATGCATAGCATGACAACTC
A*24:02:01:01
Alignment to IMGT database psl file
Each haplotype sequence was aligned to all alleles in the IMGT database

Figure 2.2: Overview of the analytical pipeline used in this study

2.3.1 Quality control of NGS data

NGS platforms are known to introduce several errors during sequencing runs. One of these systematic errors is the decrease in base quality as cycling of sequencing progresses, which is known to be the case for the Mi-Seq platform. Such biases can be clearly revealed by simply plotting base quality per sequencing cycle. One of the very common solutions to this problem is the trimming of reads at the point where base quality drops to less than 20. Although this threshold is an arbitrary one, it is widely accepted; therefore, I trimmed reads in the fastq files using Trimmomatic (Bolger *et al*. 2014), allowing only reads with ends of base quality 20 or more.

2.3.2 Mapping sequencing reads to the reference genome (hg19, GRCh37)

The data in fastq files are short sequences (reads) without any information about their position in the genome: therefore, reads need to be aligned to specific genomic positions. Several sequence alignment programs are available, of which Burrows-Wheeler Aligner (BWA) is most commonly used (Li and Durbin 2009). I used BWA version 0.7.12 to align the trimmed fastq files to the reference genome hg19 (GRCh37, Genome Reference Consortium Human Reference 37). One of the considerations I made during the mapping step is the high degree of homology between HLA genes, which may cause mapping errors when reads from other highly similar positions are mistakenly mapped to the wrong locus (e.g., reads from pseudogenes such as HLA-J align to HLA-A). In my experience, mismapped reads have a higher number of mismatches than reads with a true mapping position. To reduce the effect of this type of mismapping, I utilized these mismatch differences by applying the parameter "–B" in the BWA-MEM algorithm. This parameter increases the mismatch penalty during mapping, so it leads to low mapping quality for reads with increased mismatches (like the mismapped reads in this case). Although this solution improved the mapping efficiency, some reads from highly similar positions would still be mistakenly mapped to the target HLA genes. The fact that I used paired-end sequencing (i.e. DNA fragments sequenced from both ends) means that, in some cases, the forward and reverse reads mapped to different positions which makes the insert size (the distance between forward and reverse reads) higher than the expected size of 600 bp. For this reason, I also removed reads that had large insert sizes because they are not informative for further analysis (an insert size threshold of 2000 bp was used). The output from the BWA program is a file called SAM file (stands for Sequence Alignment/Map format), which contains the sequence and information about read positions and mapping quality. SAM files were converted to a binary format (BAM file), sorted, and indexed using SAMtools version 1.3.1 (Li *et al*. 2009).

2.3.3 Phasing haplotypes from BAM files

Autosomal cells in humans are diploid, which means that they have two copies from each chromosome, except for the sex chromosomes XY in males. The differentiation between these two copies sometimes provides valuable information and is necessary to trace specific chromosomes in a population. Furthermore, the ability to identify two adjacent variants, whether they are in the same or different chromosomes, is called haplotype phasing. Genotyping approaches, such as Sanger sequencing, have no phase information because genotypes are called nominally as two alleles. One of the advantages of using NGS sequencing over other methods is the ability to identify haplotype phase by utilizing information from paired-end reads. Therefore, to obtain two completely phased haplotypes, I used the SAMtools phase algorithm, which takes a single BAM file as an input and outputs two separate BAM files, each with one haplotype.

2.3.4 Variant calling and building haplotype sequence

From this step, each of the two new BAM files (two haplotypes) was processed by sorting, indexing, and adding read groups as in the previous step. Using the two haplotype files, I called variants using HaplotypeCaller of the Genome Analysis Toolkit (GATK) (DePristo *et al*. 2011) to obtain candidate variants in VCF file format.

Since alignment programs can sometimes not accurately map reads to their true positions, in such cases some samples may have an ambiguous haplotype phase; therefore, not all of the called variants are expected to be true ones. To reduce the effect of these false positive variants on the pipeline performance, I filtered the VCF files by selecting variants with an allele frequency of ≥0.5. I based this filtering criterion on the assumption that mismapped reads are less frequent than the accurately mapped ones, so the majority of reads in the problematic positions are for the true allele and the minor allele is the false one. The filtered VCF files of the two haplotypes were used to build haplotype sequences in the next step.

To obtain the exact sequence of each haplotype, I created a new Fasta file using the GATK "FastaAlternateReferenceMaker" algorithm. This tool uses variants from the filtered VCF file and the sequence from the reference genome (hg19) to output a new sequence with those variants inserted into it; this sequence represents the actual sample's haplotype. In this step, to obtain the complete CDS sequence, I only extracted exonic sequences of the target genes and then concatenated them together.

2.3.5 Determining genotypes of HLA genes

The identification of HLA genotypes is performed by comparing samples' sequences to a list of known HLA alleles, which is provided by the IPD-IMGT/HLA database (Immune Polymorphism Database - International ImMunoGeneTics Information System®/Human

Leukocyte Antigen) (Robinson *et al*. 2015). This database serves as a public repository for storing and maintaining sequences of the HLA genes. Unfortunately, most of the sequences deposited in the database are for exons 2 and 3 in class I HLA genes and exon 2 in class II genes. Therefore, the available data sets of HLA allele frequencies for many populations are based on four-digit resolution. This limited sequence information is due to the significance of the mentioned exons in HLA matching because they contain the PBR, so typing was restricted to these exons; however, this is now changing. On the other hand, the sequencing method I am using in this study can provide complete sequences of HLA genes including not only exons but also introns (i.e., up to eight-digit resolution). To compare the samples' sequences to the ones from the database, I downloaded the latest release of all HLA alleles (3.26.0, October 2016) from the IMGT/HLA website in Fasta file format. To obtain the HLA genotypes from the data, I aligned the obtained CDS sequences of the two haplotypes to the list of all known HLA alleles from the IMGT database. The sequence alignment was performed using BLAT (Kent 2002) and the downloaded fasta files were used as a reference. I considered I have a genotype call when the sample's sequence perfectly matched a previously known allele from the IMGT/HLA list with no gaps or mismatches. However, two possibilities arose when the sample did not match any known allele. First, when the mismatch was due to low coverage in the sample, I re-sequenced that sample to improve the read depth, or when the mismatch position had poor phasing quality due to mismapped reads, I performed the phasing manually by looking at the BAM file and deciding on the most likely haplotype. Second, when the mismatch position had sufficient read depth (>20) and no phasing issues were present, I considered this as a potential new allele and performed Sanger sequencing to confirm that.

2.4 Sanger sequencing of new HLA alleles

As mentioned previously, mismatches in some samples were candidates for new HLA alleles, so they needed to be confirmed by other methods. For each mismatch, I designed specific primers flanking the variant and amplified that position using PCR. I further performed Sanger sequencing using the PCR products and the BigDye® Terminator V3.1 Cycle Sequencing kit (Life Technologies), so I sequenced both forward and reverse primers on the ABI 3130xl Genetic Analyzer (Applied Biosystems).

2.5 Population genetics analysis

2.5.1 Allele frequencies and heterozygosities

The genetic diversity indicators such as the number of alleles in each locus and the degree of heterozygosity were calculated by using Python for Population Genomics (PyPop v.0.7.0) (Lancaster *et al*. 2007). The expected heterozygosity ($\hat{H}$), assuming Hardy-Weinberg equilibrium (HWE), was estimated using the following formula:

$$\hat{\mathrm{H}} = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} {P_i}^2\right)$$

,where P$_i$ is the frequency of the *i*th allele and *n* is the number of samples for *k* number of alleles.

2.5.2 Hardy-Weinberg equilibrium test

Many population genetic tests assume that genotype frequencies are under HWE, so I also tested whether the identified genotypes follow or deviate from HWE using Arlequin, which has implementation of an exact test that follows Guo and Thompson's procedure (1992). When the gametic phase between loci is unknown, HWE is only tested between loci. The test is performed by creating a contingency table of the observed allele frequencies and calculating the number of alleles. The probability of observing the same table under null-expectations is found by Levene's sampling distribution. The P-value of this test is determined by finding the proportion of tables with a probability equal to or less than that of the previously observed table.

2.5.3 Measuring genetic distances

One of the most common genetic distance measures is the classic F_{ST} statistic. F_{ST} compares the proportion of variance within populations relative to the total variance among all populations. To understand the genetic relatedness among study populations, I computed the pair-wise F_{ST} statistic between all pairs of populations using Arlequin. The statistical significance (P value < 0.05) is determined after permuting haplotypes (10,000 permutations) under the assumption of no difference between populations. The pairwise genetic distances were used to build a neighbor-joining phylogenetic tree (Saitou and Nei 1987) using MEGA v.6 (Tamura *et al.* 2013).

2.5.4 Ewens-Watterson homozygosity test of neutrality

To test whether natural selection is operating on any HLA loci in the study populations, I performed Ewens-Watterson homozygosity test of neutrality (Ewens 1972 ; Watterson 1978) as implemented in PyPop, v.0.7.0 (Lancaster *et al.* 2007). In this test, the observed homozygosity is compared to the expected homozygosity, which is computed by simulation under neutrality expectations for the same sample size and number of alleles. The homozygosity F statistic is given by:

$$F = \sum_{i=1}^{k} P_i^2$$

,where P_i is the frequency of the *i*th for *k* number of unique alleles. The normalized deviate of homozygosity (Fnd) is calculated as the difference between observed and expected homozygosity divided by the square root of the variance of the expected homozygosity (Salamon *et al*. 1999). The reported *P*-values in this test are the probability of obtaining a homozygosity F statistic under neutrality assumptions that is less than or equal to the observed one. The implementation of the test is based on the exact test written by Slatkin (1994), which uses a Markov-Chain Monte-Carlo (MCMC) method to obtain the null distribution of homozygosity. A negative significant Fnd value indicates the observed homozygosity is deviated in the direction of balancing selection, while a significant positive value indicates directional selection.

2.5.5 Estimation of haplotype frequencies

To estimate haplotype frequencies between HLA loci, I used the maximum likelihood (ML) estimation approach as implemented in Arlequin (Excoffier and Slatkin 1995). This procedure is used when the genotypes has no phase information. Furthermore, ML method determines haplotype frequencies through an iterative process to reach the maximum likelihood estimation of candidate haplotypes. Although there is no guarantee that all of the obtained haplotypes are true ones, it was shown that, at least for common alleles, ML achieved relatively accurate prediction compared with haplotypes estimated from data with known gametic phase.

2.5.6 Detection of ancestry-informative alleles

To determine the ancestry-informative alleles, I used principal component analysis (PCA), a mathematical technique that reduces the dimensionality of the data while retaining most of its variations. Conceptually, PCA works by finding the projections that explain maximum variations in the data. I performed PCA on a covariance matrix of allele frequencies in all groups using the function (prcomp) from the STATS package in R version 3.2.1 (R Foundation for Statistical Computing, Vienna, Austria), which uses a singular value decomposition method. The two significant estimates in PCA analysis are the principal components (PCs, also called eigenvectors) and PC scores (eigenvalues). I selected the PCs that explain most of the variations in allele frequencies among populations (usually the first few PCs). Clustering patterns were then determined by plotting these PCs. I assumed that the ancestry-informative alleles are the ones associated with clustering patterns in PC plots; therefore, they have more discriminatory power. This approach proved to be effective in tracing ancestral haplotypes even in admixed populations like the Japanese population (Nakaoka *et al*. 2013). I therefore selected the alleles with PC scores greater than one standard deviation as ancestry-informative alleles. To determine whether the identified alleles are significantly different between populations, I used Fisher's exact test as implemented in R.

2.5.7 Estimation of linkage disequilibrium

Linkage disequilibrium (LD) measures the non-random association between alleles at two different loci in the same chromosome: in other words, it is a measure of how these two loci are associated so they are transmitted to the next generation as a single unit. Using the PyPop program, I calculated pairwise LD between different HLA loci (i.e., A, B, and C) based on the normalized measure Lewontin's D' defined as: $D'_{ij} = D_{ij}/D(max)$.

Chapter Three

RESULTS

3.1 Measures of genetic diversity

3.1.1 Number of alleles and heterozygosity

The number of observed alleles and the heterozygosity for each locus in the studied populations are presented in **Table 3.1**. As was found previously in Sub-Saharan African populations (Vina *et al*. 2012), all groups in this study showed a relatively large number of alleles compared with other non-African populations. The number of alleles in *HLA-A* for all groups was between the values for *HLA-B* and *HLA-C*, except for *HLA-A* in Nubians, which was the minimum number among the groups. In *HLA-A*, Nubians had the minimum number of alleles and Gaalien had the largest one. On the other hand, Nuba had the least number of alleles in *HLA-B* and Nilotes had the largest. Finally, the number of alleles in *HLA-C* was minimum in Darfurians while Nilotes showed the greatest number of alleles.

In addition to the large number of alleles, the observed heterozygosity in all loci was high, but did not deviate from the expected heterozygosity under HWE (**Table 3.1**). However, only one locus (*HLA-B* in Nilotes) showed a low heterozygosity value, which also deviated from the expected heterozygosity (*P* value <0.05).

Table 3.1: Number of alleles and heterozygosity in HLA class I genes of eight East African groups

Group	**2N**	**No. of alleles (Heterozygosity, %)**		
		HLA-A	***HLA-B***	***HLA-C***
Beja	78	23 (92.3)	27 (92.3)	20 (94.9)
Ethiopians	66	21 (90.9)	28 (100)	21 (100)
Nubians	72	15 (80.6)	32 (88.9)	20 (94.4)
Gaalien	70	26 (97.1)	30 (100)	17 (94.3)
Shokrya	60	23 (93.3)	25 (93.3)	17 (83.3)
Darfurians	70	20 (100)	27 (100)	15 (94.3)
Nuba	72	20 (97.2)	24 (91.7)	19 (91.7)
Nilotes	92	25 (97.8)	34 (84.8)*	23 (91.3)

2N, number of gene copies; * indicates significant deviation from Hardy-Weinberg expectations (*P* < 0.05).

3.1.2 Population differentiation

Pairwise comparisons between the study populations showed marked differences in F_{st} values (**Table 3.2)**. In general, the genetic distances, as measured by the F_{st} statistic, were larger between Afro-Asiatic and Nilo-Saharan/Niger-Congo families than between groups within the same linguistic family. In almost all pairwise comparisons, populations were significantly differentiated from each other; however some populations had close genetic distances. Populations that had close genetic distances include the Nubian-Ethiopian pair, which showed the closest, and the Darfurian-Nuba pair from the Nilo-Saharan/Niger-Congo families.

Phylogenetic analysis based on the calculated F_{st} statistics indicated that the most distant populations were Darfurians (Nilo-Saharan) on one side and the two Arab groups (Gaalien and Shokrya) from the other (**Figure 3.1**).

Table 3.2: F_{st} indices for pairs of eight East African populations

	Beja	Gaalien	Nubians	Ethiopians	Nuba	Darfurians	Nilotes
Gaalien	0.01694*						
Nubians	0.00738*	0.0098*					
Ethiopians	0.00641*	0.00807*	0				
Nuba	0.02094*	0.02557*	0.01041*	0.01171*			
Darfurians	0.02689*	0.0312*	0.01982*	0.01658*	0.00058		
Nilotes	0.02124*	0.02411*	0.01267*	0.01311*	0.0052*	0.01278*	
Shokrya	0.01324*	0.01459*	0.01065*	0.01309*	0.02928*	0.03406*	0.02877*

* Significant at a *P*-value < 0.05.

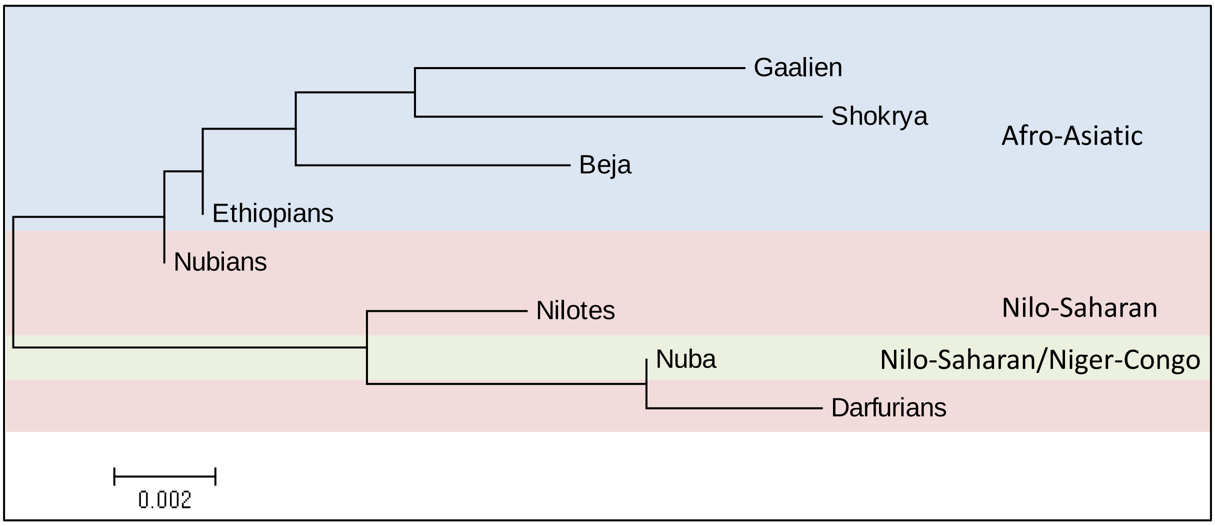

Figure 3.1: Neighbor-joining phylogenetic tree based on F_{st} indices in Table 3.2

3.2 Test of Natural selection

To test whether any of the HLA loci is under selection, I performed Ewens-Watterson test of selective neutrality which revealed, in few populations, the observed homozygosity significantly deviated from the expected homozygosity under neutrality assumption for the same sample size and number of alleles (**Table 3.3**). The loci that showed significant deviations from expected homozygosity were: *HLA-A* in Nubians, *HLA-B* in Ethiopians and Nuba, and *HLA-C* in Nuba. All of these loci showed a significant negative *Fnd* value, which suggests that balancing selection is operating on these loci.

Table 3.3: Ewens-Watterson test of selective neutrality for eight East African groups

Group	locus	Observed F	Expected F	Normalized deviate of F (Fnd)	P value
Gaalien	A	0.06	0.08	-0.92	0.112
	B	0.07	0.06	0.38	0.736
	C	0.09	0.13	-1.07	0.066
Shokrya	A	0.07	0.09	-0.79	0.180
	B	0.08	0.07	0.86	0.851
	C	0.11	0.13	-0.44	0.383
Beja	A	0.10	0.09	0.35	0.731
	B	0.06	0.07	-0.67	0.246
	C	0.10	0.11	-0.50	0.346
Ethiopians	A	0.07	0.10	-1.13	0.042
	B	0.05	0.07	-1.23	0.024*
	C	0.10	0.11	-0.16	0.540
Nubians	A	0.09	0.15	-1.37	0.007*
	B	0.05	0.06	-0.53	0.319
	C	0.08	0.11	-1.10	0.050
Nuba	A	0.11	0.11	-0.01	0.608
	B	0.06	0.09	-1.42	0.005**
	C	0.08	0.12	-1.25	0.017*
Darfurians	A	0.11	0.11	-0.01	0.608
	B	0.06	0.07	-0.70	0.227
	C	0.13	0.17	-0.73	0.227
Nilotes	A	0.08	0.09	-0.66	0.259
	B	0.07	0.06	0.72	0.820
	C	0.08	0.11	-0.88	0.141

3.3 Identification of new HLA alleles

As mentioned previously in the introduction, finding a new variant(s) in an HLA gene that is not previously reported is considered new allele for that gene. Although most of the time the new variant is previously known, it is not seen in the background of accompanying variants (i.e., the haplotype). As I was expecting, because HLA genes in Africans are relatively not well studied and genetic diversity in African populations is thought to be high, I found four new HLA alleles in the Nubian and Darfurian groups, the details of which are presented in **Table 3.4** and **Figures 3.2** and **3.3**.

The first new allele was found in *HLA-C* of three individuals from the Nubian group. This new allele differs from the closest one (C*14:02:01) by two single-nucleotide variants in exon 1 (**Table 3.4** and **Figure 3.2 A**), both of which are nonsynonymous. Two of the three individuals are homozygous for both variants, while the third one has a heterozygote genotype. The second new allele was found in *HLA-B* of one individual from the Nubian group. This allele has one nucleotide variant from the closest B*51:01:01:01 allele, which does not change the amino acid codon (synonymous variant) (**Figure 3.2 B**).

The third and fourth new alleles were found in *HLA-B* of three individuals from the Darfurian group. The third allele has a single synonymous variant from B*35:01:01:01 and was found in one individual (**Figure 3.3 B**), while in the fourth allele, a nonsynonymous variant in B*39:10:01 background changed the amino acid codon from serine to cysteine (**Figure 3.3 A**).

Table 3.4: New HLA alleles found in Nubian and Darfurian groups

Sample ID	Group	Gene	Zygosity	Closest allele	Position	Variant type	AA change
31	Nubians	C	Heterozygote	C*14:02:01	31239821	Nonsynonymous	L > I
32	Nubians	C	Homozygote	C*14:02:01	31239821	Nonsynonymous	L > I
38	Nubians	C	Homozygote	C*14:02:01	31239821	Nonsynonymous	L > I
31	Nubians	C	Heterozygote	C*14:02:01	31239824	Nonsynonymous	L > V
32	Nubians	C	Homozygote	C*14:02:01	31239824	Nonsynonymous	L > V
38	Nubians	C	Homozygote	C*14:02:01	31239824	Nonsynonymous	L > V
78	Nubians	B	Heterozygote	B*51:01:01:01	31323116	Synonymous	P > P
4	Darfurians	B	Heterozygote	B*35:01:01:01	31324902	Synonymous	L > L
18	Darfurians	B	Heterozygote	B*39:10:01	31324371	Nonsynonymous	S > C
57	Darfurians	B	Heterozygote	B*39:10:01	31324371	Nonsynonymous	S > C

AA change, amino acid change

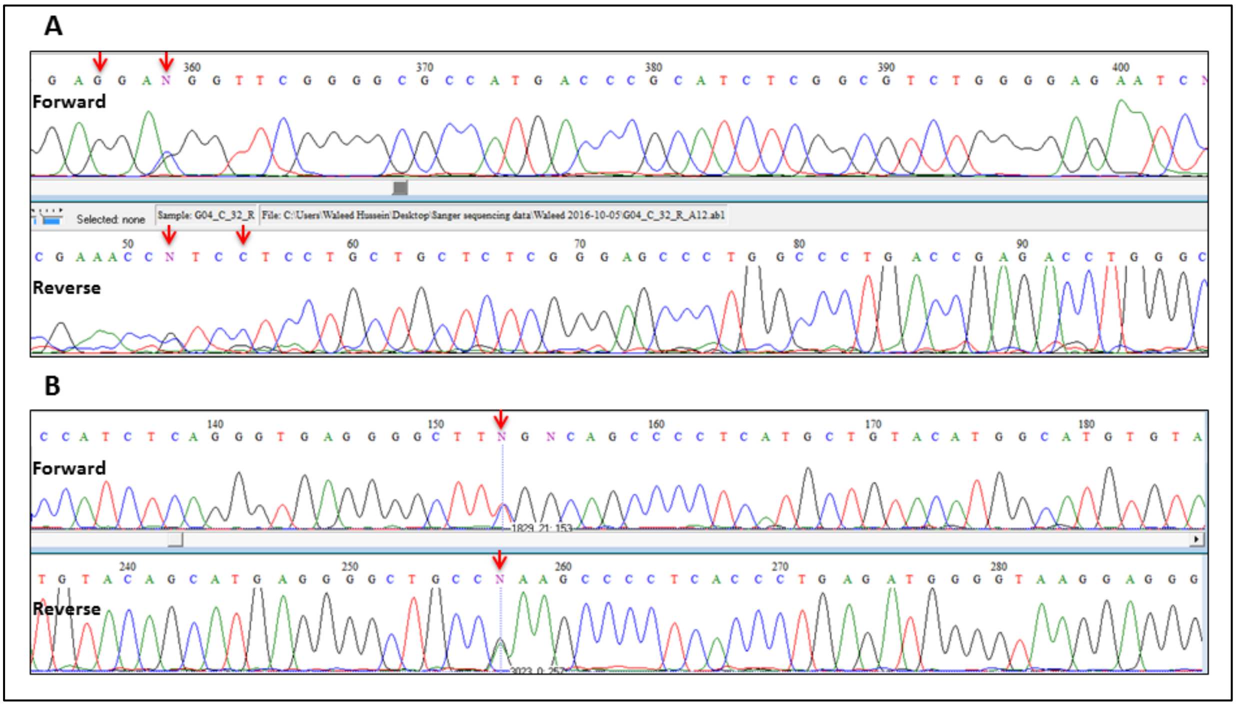

Figure 3.2: New HLA class I alleles found in Nubian group

(A) A new *HLA-C* allele has two variants from the closest C*14:02:01 allele. (B) A new *HLA-B* allele differs by a single variant from the B*51:01:01:01 allele. Variant positions in both alleles are indicated by red arrows.

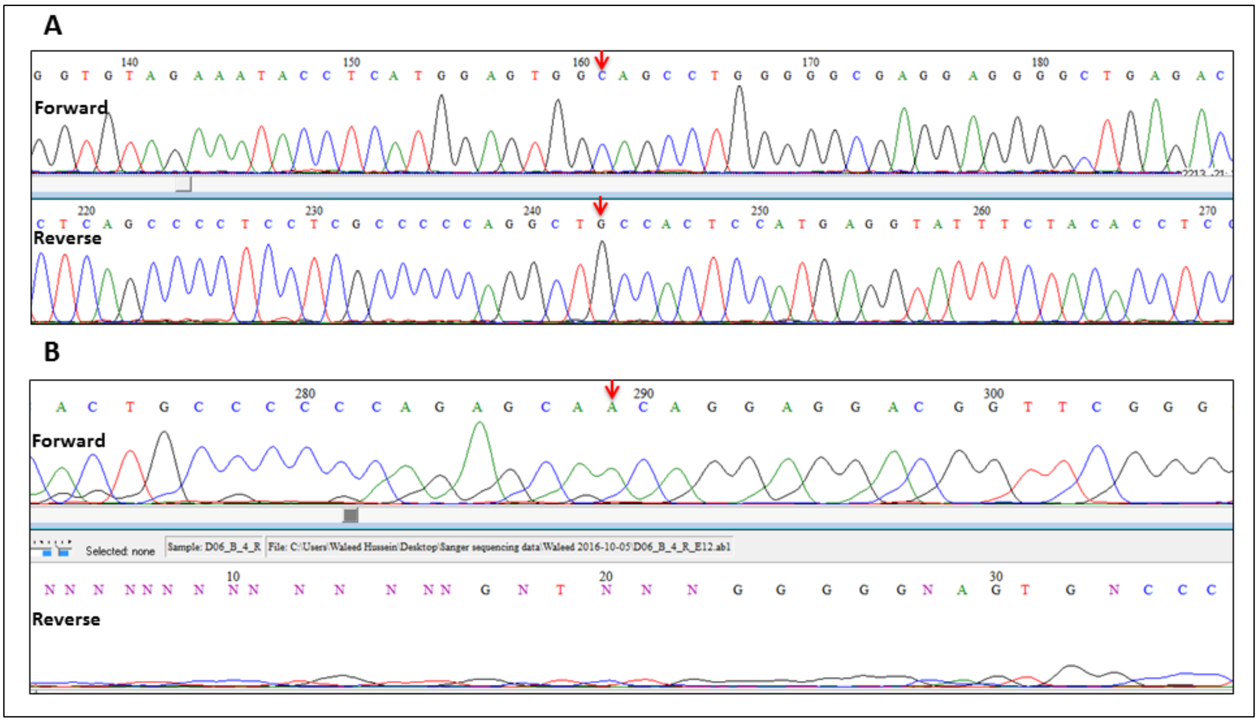

Figure 3.3: New HLA class I alleles found in Darfurian group

(A) A new *HLA-B* allele that has a single variant from its closest B*39:10:01 allele. (B) A new *HLA-B* allele differs by a single variant from the B*35:01:01:01 allele. Variant positions in both alleles are indicated by red arrows.

3.4 Common HLA alleles in the study populations

The most frequent HLA class I alleles for all groups are presented in **Table 3.5, Table 3.6** and **Table 3.7**. As shown in these tables, the most frequent *HLA-A* allele among all populations, with the exception of the Arab groups, was A*02:01, which reached 25% in Nuba. Furthermore, the most frequent *HLA-A* alleles among Arabs (Gaalien and Shokrya) were A*30:02 and A*01:01. *HLA-B,* however, had a different situation as the most frequent allele was not shared between the groups and Arab groups had B*51:01 and B*50:01 as the most frequent alleles in Gaalien and Shokrya, respectively. In *HLA-C*, the C*06:02 and C*07:01 alleles were the most frequent among the Afro-Asiatic groups and the Nilo-Saharan group Nubians as well. On the other hand, the most frequent alleles among the other Nilo-Saharan and Niger-Congo groups were C*04:01 and C*07:18.

Table 3.5: Most common HLA class I alleles among Afro-Asiatic groups (≥5%)

HLA-A		***HLA-B***		***HLA-C***	
Allele	**Frequency**	**Allele**	**Frequency**	**Allele**	**Frequency**
Beja					
A*02:01	0.205	B*13:02	0.125	C*06:02	0.225
A*03:01	0.192	B*07:02	0.088	C*07:02	0.100
A*30:02	0.077	B*41:01	0.088	C*17:01	0.100
A*32:01	0.051	B*47:01	0.088	C*07:01	0.088
A*68:01	0.051	B*51:01	0.088	C*16:04	0.075
		B*14:02	0.063	C*04:01	0.063
		B*08:01	0.050	C*08:02	0.063
Ethiopians					
A*02:01	0.162	B*50:01	0.097	C*07:01	0.184
A*01:01	0.095	B*49:01	0.083	C*04:01	0.158
A*03:01	0.081	B*07:02	0.069	C*06:02	0.145
A*30:02	0.081	B*15:220	0.069	C*07:02	0.079
A*68:01	0.068	B*41:01	0.069	C*15:05	0.066
A*01:03	0.054	B*41:02	0.056	C*17:01	0.053
A*02:02	0.054	B*57:03	0.056		
A*02:05	0.054				
A*23:01	0.054				
Gaalien					
A*30:02	0.129	B*51:01	0.158	C*07:01	0.145
A*02:01	0.114	B*07:02	0.105	C*07:02	0.132
A*03:02	0.071	B*49:01	0.092	C*15:02	0.132
A*01:01	0.057	B*14:02	0.066	C*12:03	0.092
A*30:01	0.057	B*52:01	0.066	C*06:02	0.079
A*31:01	0.057			C*12:02	0.079
				C*04:01	0.066
				C*08:02	0.053
Shokrya					
A*01:01	0.149	B*50:01	0.171	C*06:02	0.237
A*02:01	0.122	B*51:01	0.157	C*04:01	0.118
A*33:03	0.081	B*13:02	0.086	C*07:01	0.118
A*02:02	0.068	B*52:01	0.086	C*15:02	0.105
A*31:01	0.068	B*41:01	0.071	C*17:01	0.079
A*02:05	0.054	B*35:08	0.057	C*12:02	0.053
A*03:01	0.054				
A*11:01	0.054				

Table 3.6: Most common HLA class I alleles among Nilo-Saharan groups (≥5%)

HLA-A		***HLA-B***		***HLA-C***	
Allele	**Frequency**	**Allele**	**Frequency**	**Allele**	**Frequency**
Darfurians					
A*02:01	0.250	B*53:01	0.125	C*04:01	0.256
A*68:02	0.125	B*58:01	0.113	C*07:18	0.146
A*30:04	0.097	B*15:220	0.100	C*06:02	0.122
A*30:01	0.083	B*45:01	0.063	C*16:01	0.098
A*03:01	0.056	B*07:02	0.050	C*12:03	0.061
A*29:02	0.056	B*15:10	0.050		
		B*51:01	0.050		
Nilotes					
A*02:01	0.149	B*58:01	0.177	C*06:02	0.133
A*01:01	0.128	B*08:01	0.115	C*07:01	0.133
A*68:02	0.106	B*35:01	0.063	C*07:18	0.133
A*02:05	0.074	B*47:03	0.063	C*07:02	0.082
A*30:01	0.074	B*53:01	0.063	C*07:04	0.082
A*03:01	0.064	B*07:02	0.052	C*04:01	0.071
				C*03:02	0.061
Nubians					
A*02:01	0.154	B*49:01	0.097	C*06:02	0.145
A*01:01	0.115	B*50:01	0.083	C*07:01	0.145
A*02:05	0.103	B*51:01	0.083	C*04:01	0.105
A*03:01	0.103	B*07:02	0.069	C*07:02	0.066
A*02:02	0.090	B*15:16	0.069	C*07:18	0.053
A*24:02	0.077	B*41:01	0.069	C*12:03	0.053
A*32:01	0.077			C*16:01	0.053
A*30:01	0.064			C*17:01	0.053

Table 3.7: Most common HLA class I alleles among Nilo-Saharan/Niger-Congo groups (≥5%)

HLA-A		***HLA-B***		***HLA-C***	
Allele	**Frequency**	**Allele**	**Frequency**	**Allele**	**Frequency**
Nuba					
A*02:01	0.250	B*53:01	0.090	C*04:01	0.128
A*30:04	0.111	B*58:01	0.090	C*07:18	0.128
A*30:01	0.097	B*15:20	0.077	C*07:01	0.103
A*32:01	0.083	B*39:10	0.077	C*16:01	0.103
A*02:05	0.069	B*51:01	0.077	C*06:02	0.077
A*01:01	0.056	B*08:01	0.064	C*07:04	0.064
A*23:01	0.056	B*35:01	0.064	C*12:03	0.064
		B*40:12	0.051		
		B*41:01	0.051		

3.5 Identification of ancestry-informative alleles

To trace the ancestry of populations, reliable markers need to be identified. In this study, I performed principal component analysis (PCA) on a covariance matrix of class I HLA allele frequencies. This analysis revealed that the first two principal components (PCs) explained 75.1% of the variance in the data (60.5% and 14.6% contributions from the first and second components, respectively) (**Figure 3.4**). In the first principal component, Arab groups (Gaalien and Shokrya) clustered away from the other groups in the Afro-Asiatic family (Beja and Ethiopians) (**Figure 3.5**). I also found the Nilo-Saharan group Nubians in the same cluster with Ethiopians and Beja and away from the other Nilo-Saharan/Niger-Congo groups (Darfurians, Nilotes, and Nuba). These findings are consistent with the previous genetic distance analysis based on F_{st} statistic. Moreover, the Nilo-Saharan/Niger-Congo groups, apart from the Nubians, were clustered as a single group. The second principal component divided Afro-Asiatic groups with Nubians from Nilo-Saharan/Niger-Congo groups.

To further investigate the clustering patterns and to identify the alleles informative for discrimination between populations, I plotted the principal component scores (PCSs), which are the projections of alleles in relation to each principal component (**Figure 3.6**). These projections reflect the contributions of alleles to the total variance explained by the PCs. Alleles, whose projections are deviate markedly from the PCs make greater contributions and thus are informative for discrimination between populations. I selected the alleles that had more than one standard deviation from the mean of the first or the second principal components.

I found that the selected alleles exhibited patterns of allele frequencies that could be categorized into three clusters. The first cluster of alleles included B*58:01, C*07:18, B*47:03, B*35:01, B*58:01, B*42:0 and several other alleles (cluster 1 in **Figure 3.6**).

Alleles of cluster I have a high frequency in Nilo-Saharan and Niger Congo groups. Meanwhile, the second cluster features the A*30:02, B*41:01, B*49:01, and B*50:01 alleles , which have high frequencies in the Afro-Asiatic groups, in addition to Nubians from the Nilo-Saharan family (**Figure 3.7**). Alleles of the last cluster (cluster 3) are specific to the Arab groups (Gaalian and Shokrya) and characterized by a high frequency among them. This last cluster contains the A*31:01, B*52:01, C*12:02, and C*15:02 alleles (**Figure 3.8**).

For further differentiation between the Arab and other Afro-Asiatic groups, I looked at the third principal component, which explained only 6% of variance. **Figure 3.9** shows clustering patterns along the second and third PCs, which did not differ markedly from the previous PC1 and PC2 plot. However, in this last plot, the third PC (PC3) clearly showed the division of Arab (Gaalien and Shokriyah) and non-Arab (Beja, Ethiopians and Nubians) groups. This distinction between Arabs and non-Arabs was of particular interest, so I aimed to identify all the alleles that were associated with it. The PCSs of the second and third PCs are shown in **Figure 3.10**. As in the PCS plot in **Figure 3.6**, the previously identified clusters are also visible; however, in this plot, cluster 3 had more alleles than the plot of the first and second PCs (B*51:01, A*30:02, and A*03:02). The frequency distributions of the alleles of clusters 1 and 2 alleles are presented in **Figure 3.11** and that of cluster 3 in **Figure 3.12**. The foregoing findings point out to a high level of genetic substructure in the Sudanese population, particularly the substructure that distinguishes Arabs from the other groups in the Afro-Asiatic family.

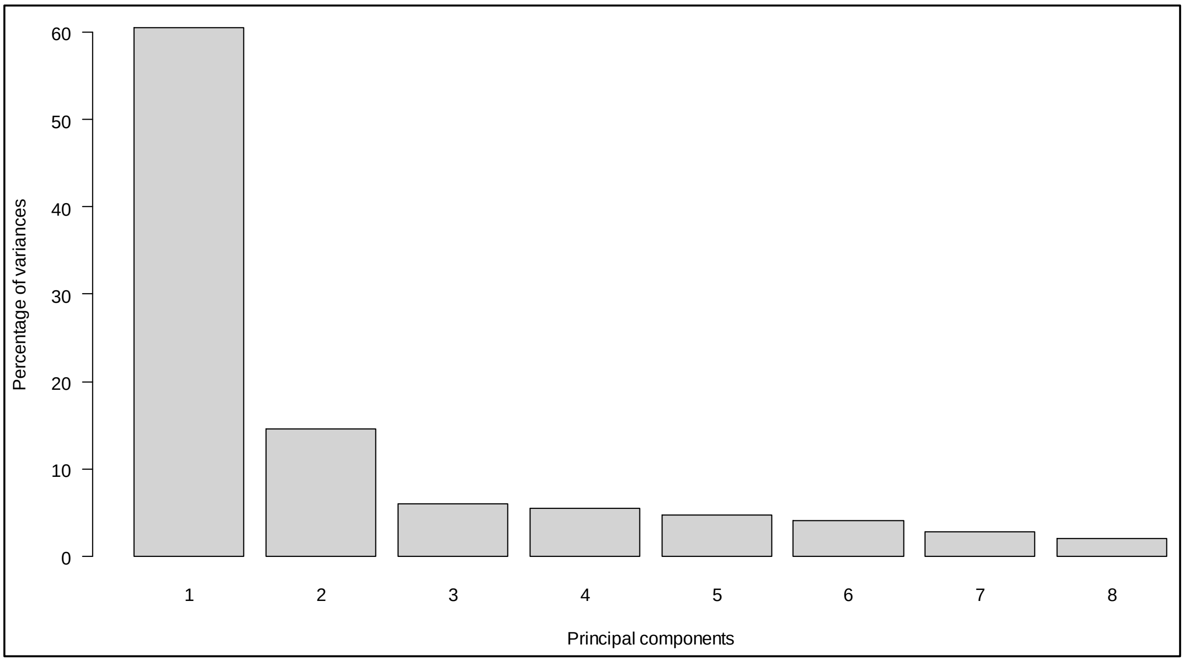

Figure 3.4: Percentage of variance explained by each principal component

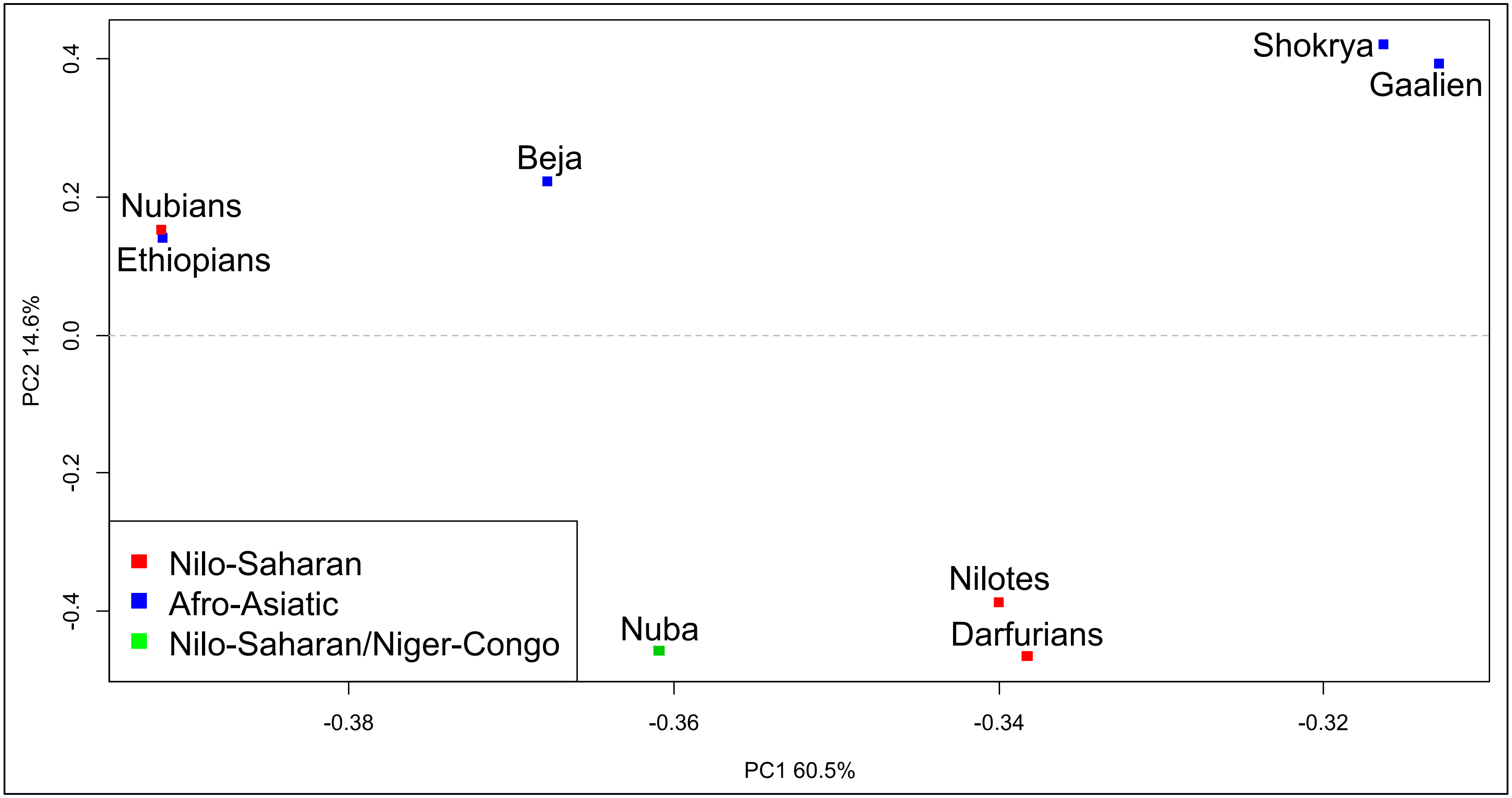

Figure 3.5: PCA analysis of class I HLA genes in eight East African populations

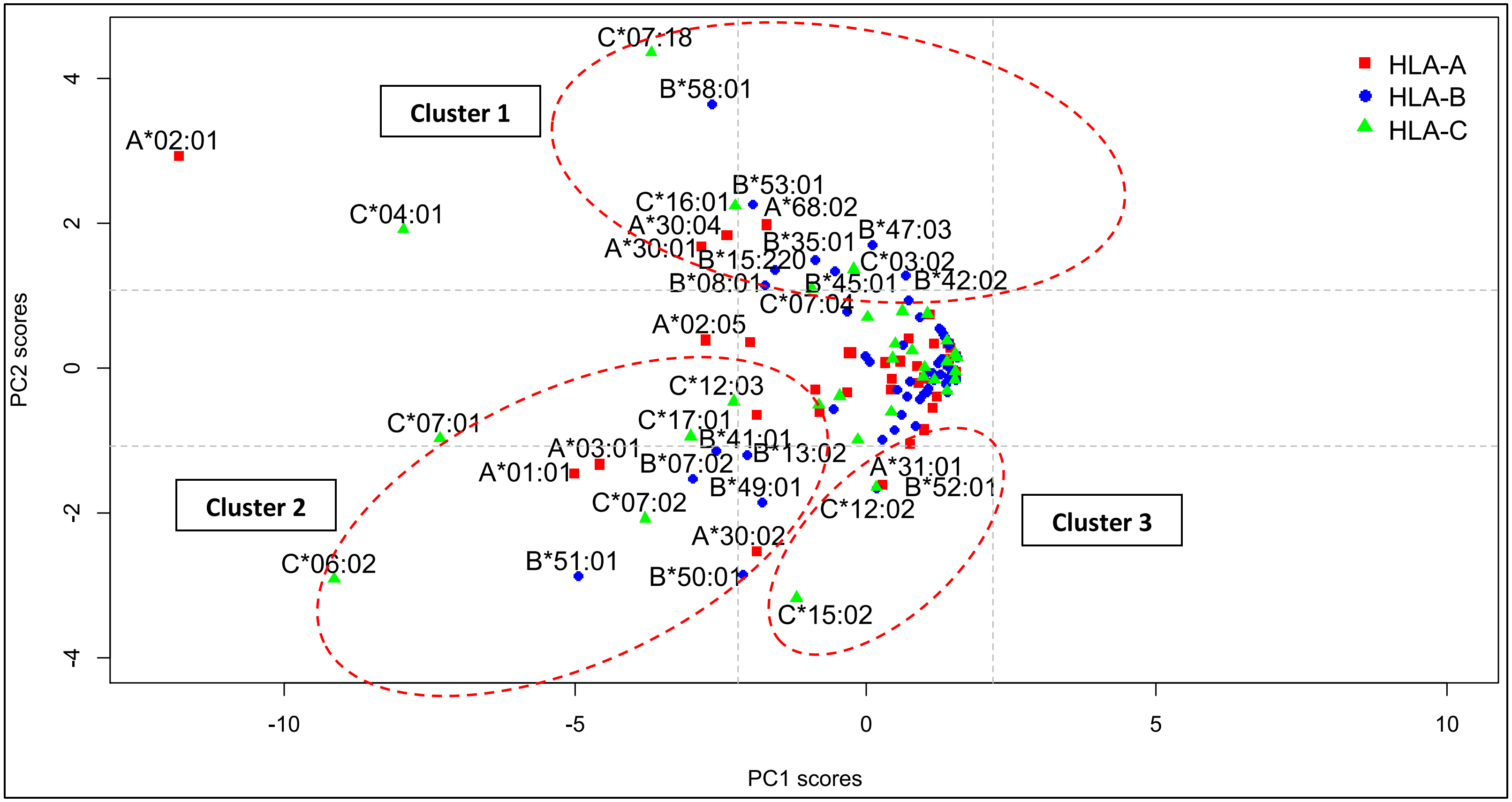

Figure 3.6: PC scores of the first and second principal components in class I HLA genes of eight East African groups

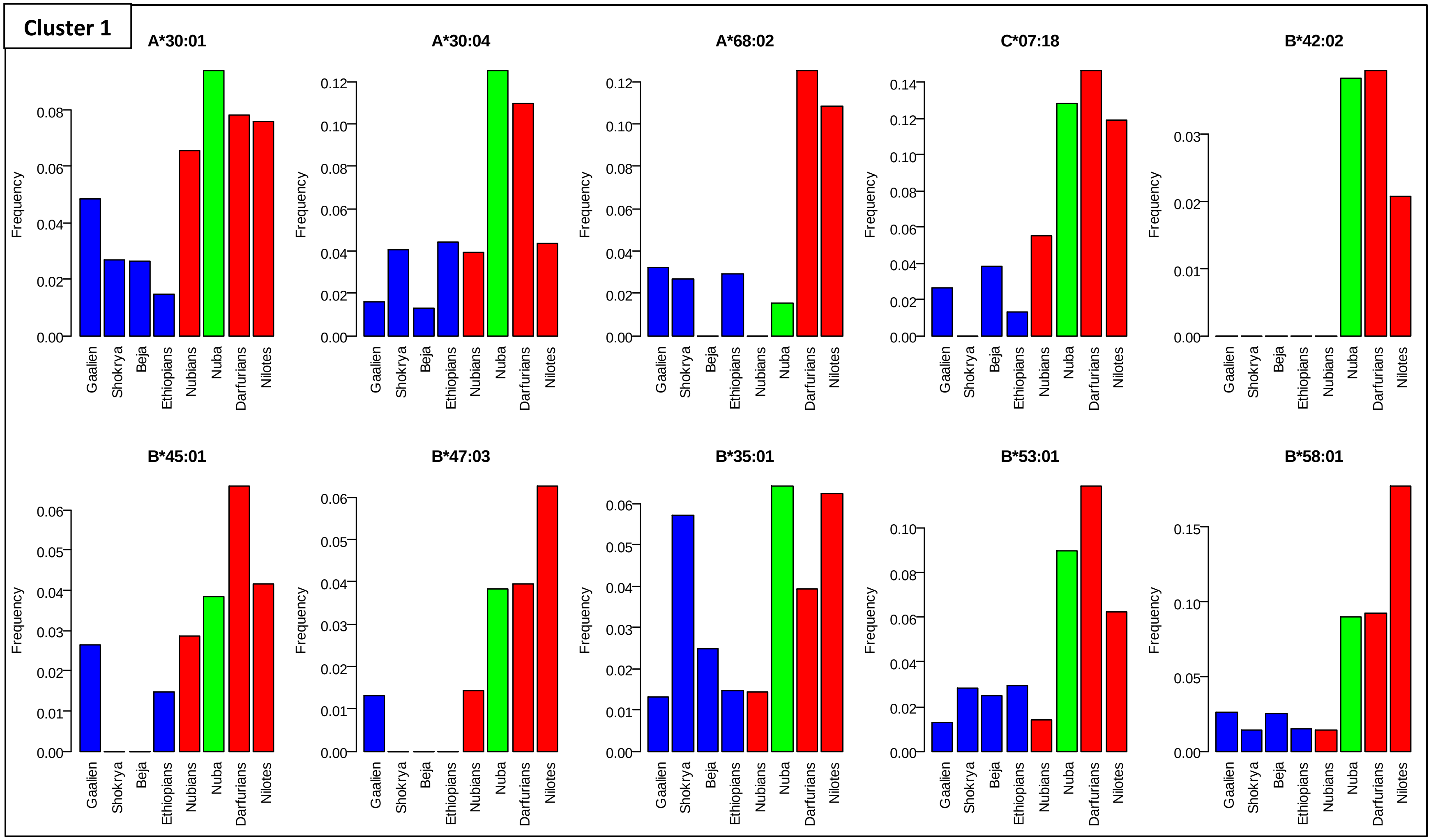

Figure 3.7: Frequency distribution of cluster 1 alleles in the PCS plot of Figure 3.6

Colors of the bars correspond to allele frequencies in the Nilo-Saharan (red), Afro-Asiatic (blue) and Nilo-Saharan/Niger-Congo (green) groups.

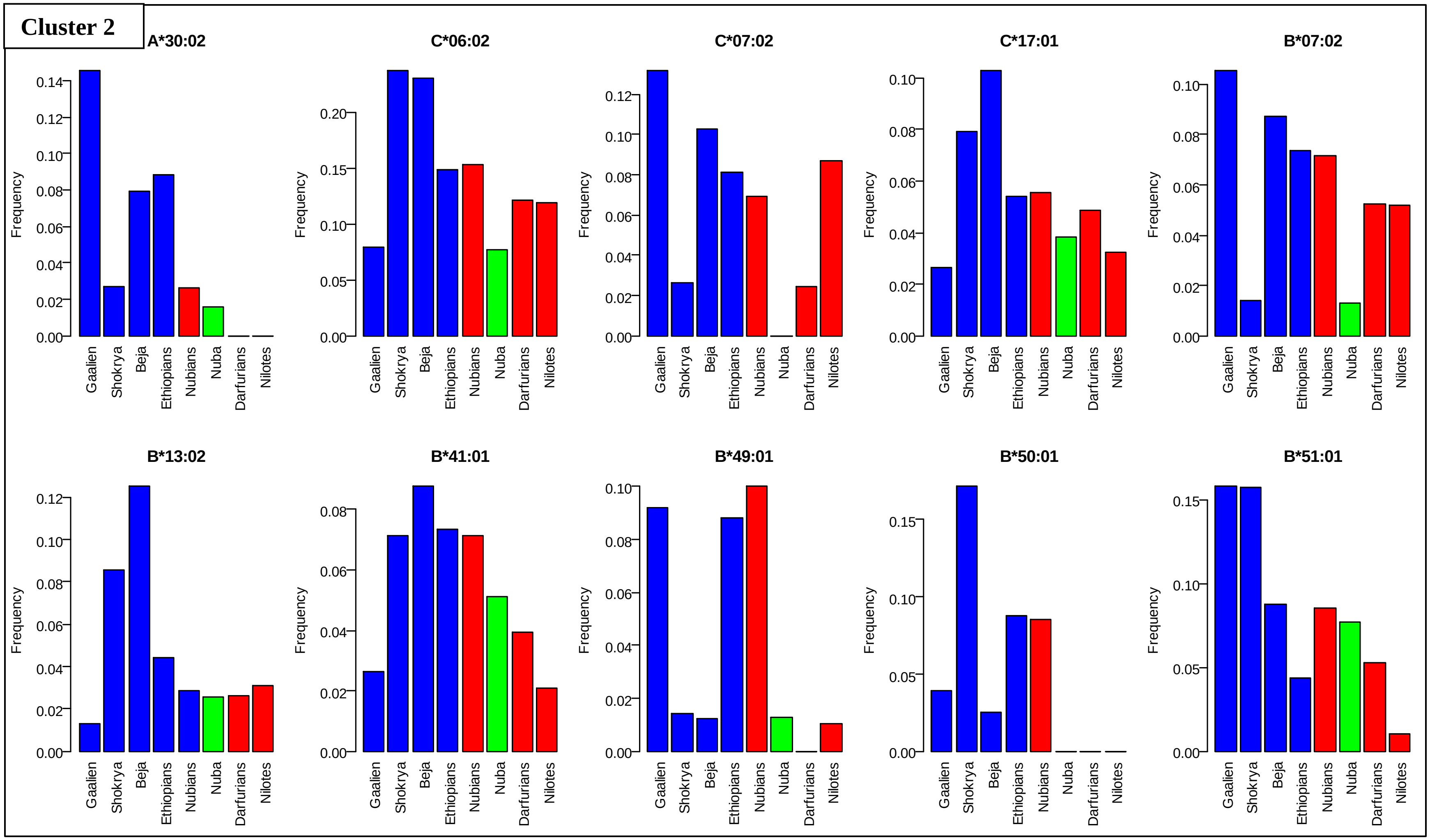

Figure 3.8: Frequency distribution of cluster 2 alleles in the PCS plot of Figure 3.6

Colors of the bars correspond to allele frequencies in the Nilo-Saharan (red), Afro-Asiatic (blue) and Nilo-Saharan/Niger-Congo (green) groups.

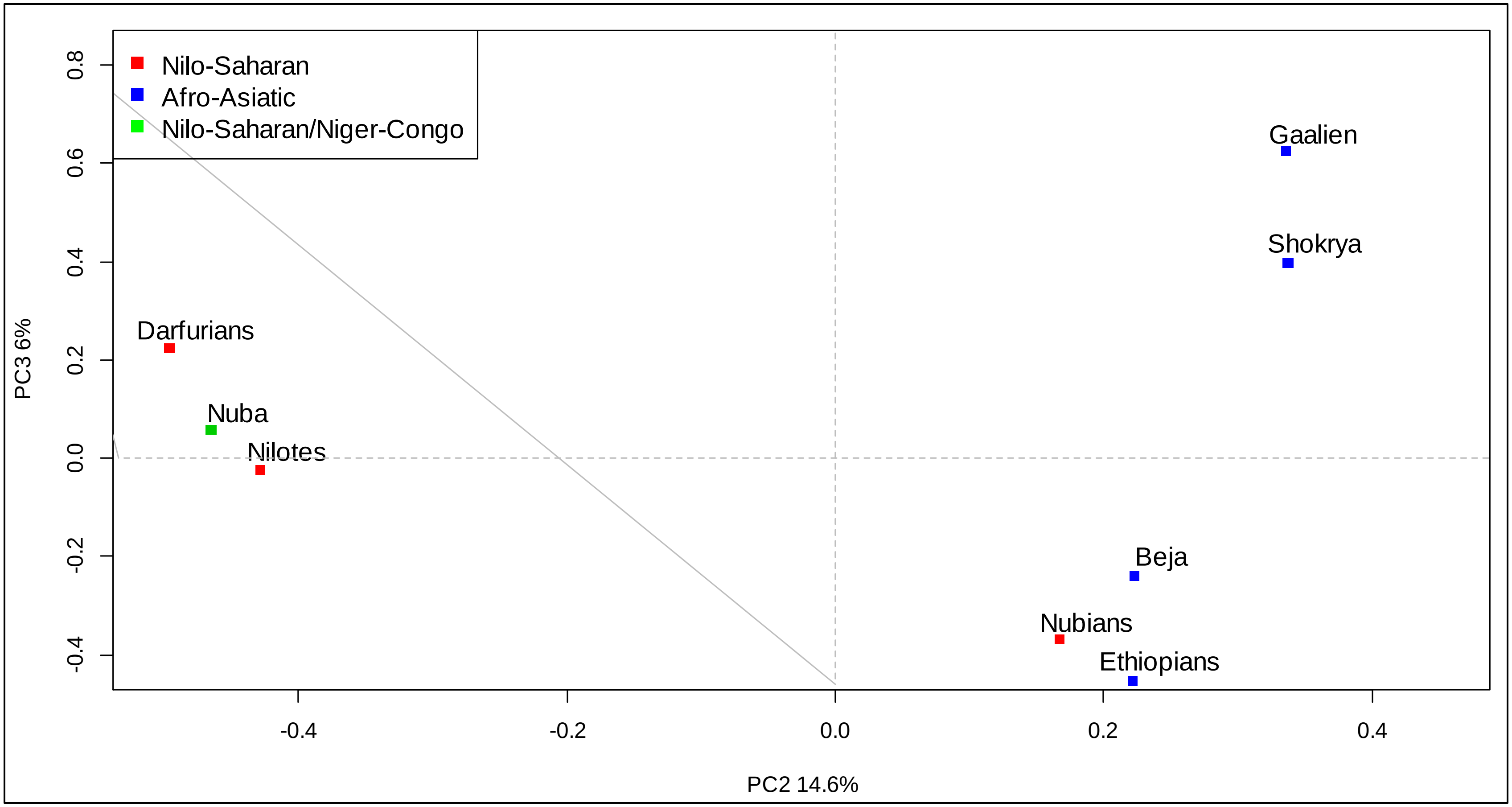

Figure 3.9: PCA of class I HLA genes in eight East African populations

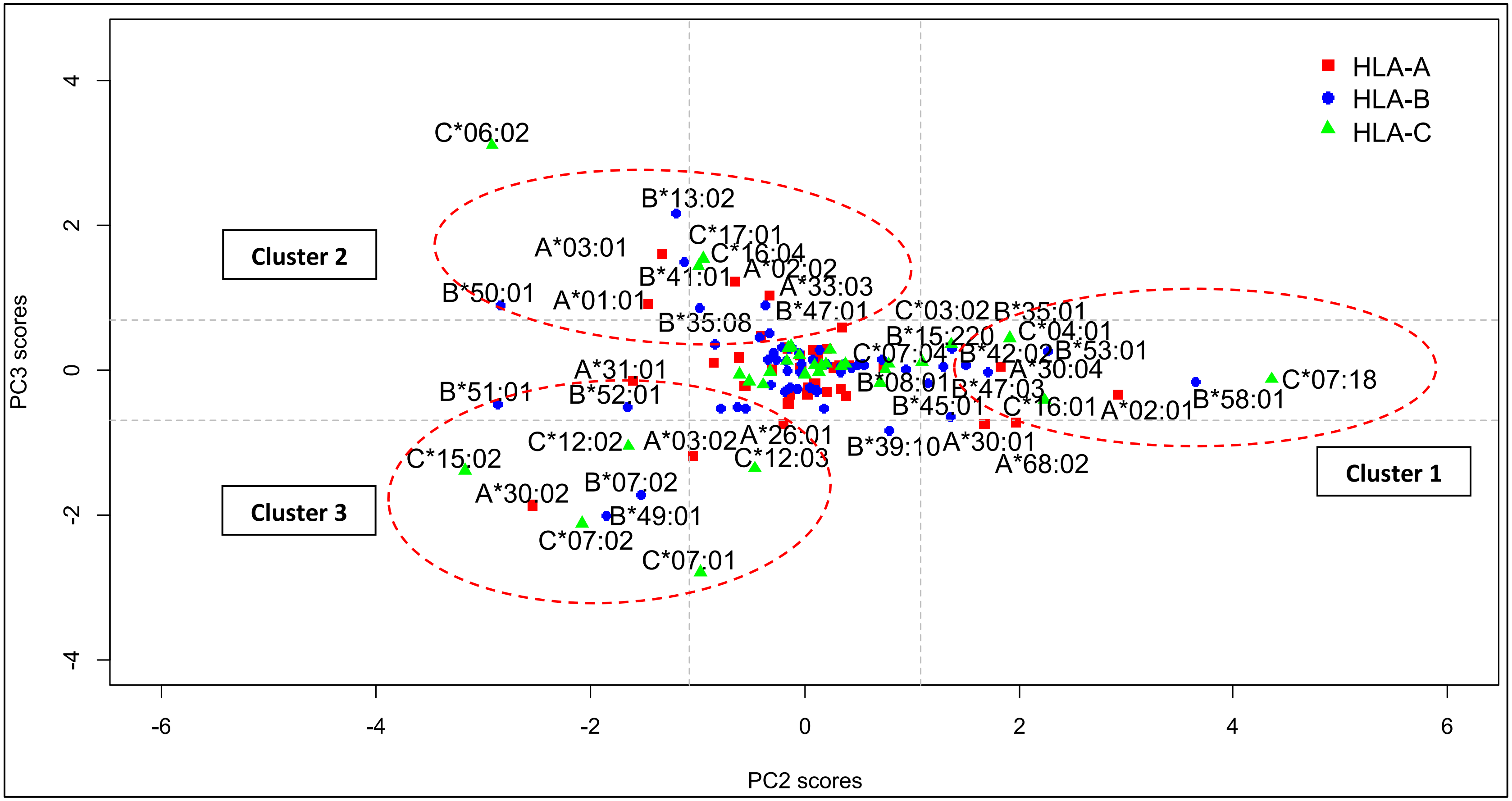

Figure 3.10: PC scores of the second and third PCs in class I HLA genes of eight East African groups

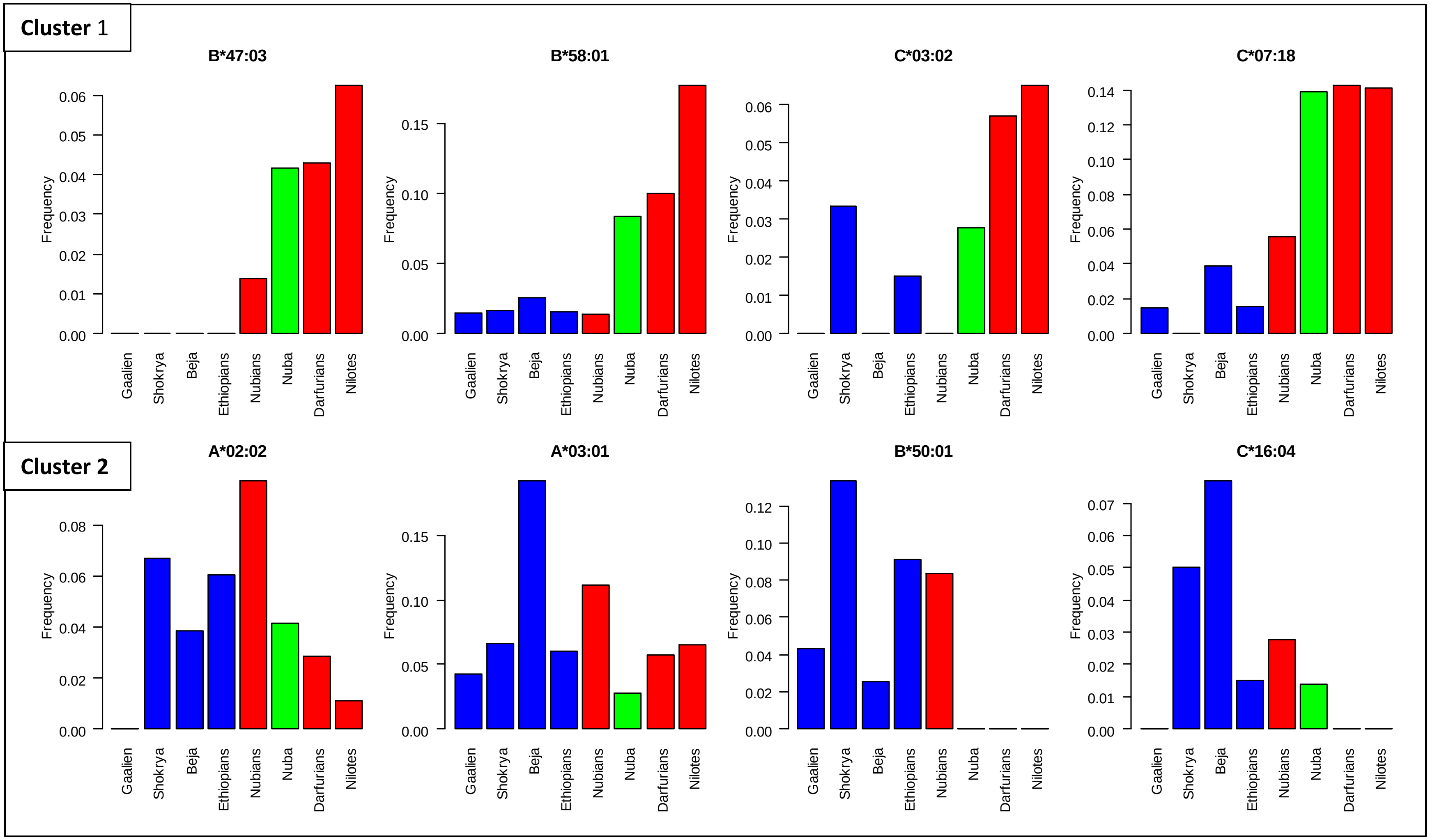

Figure 3.11: Frequency distribution of alleles from clusters 1 & 2 in the PCS plot of Figure 3.10

Colors of the bars correspond to allele frequencies in the Nilo-Saharan (red), Afro-Asiatic (blue) and Nilo-Saharan/Niger-Congo (green) groups.

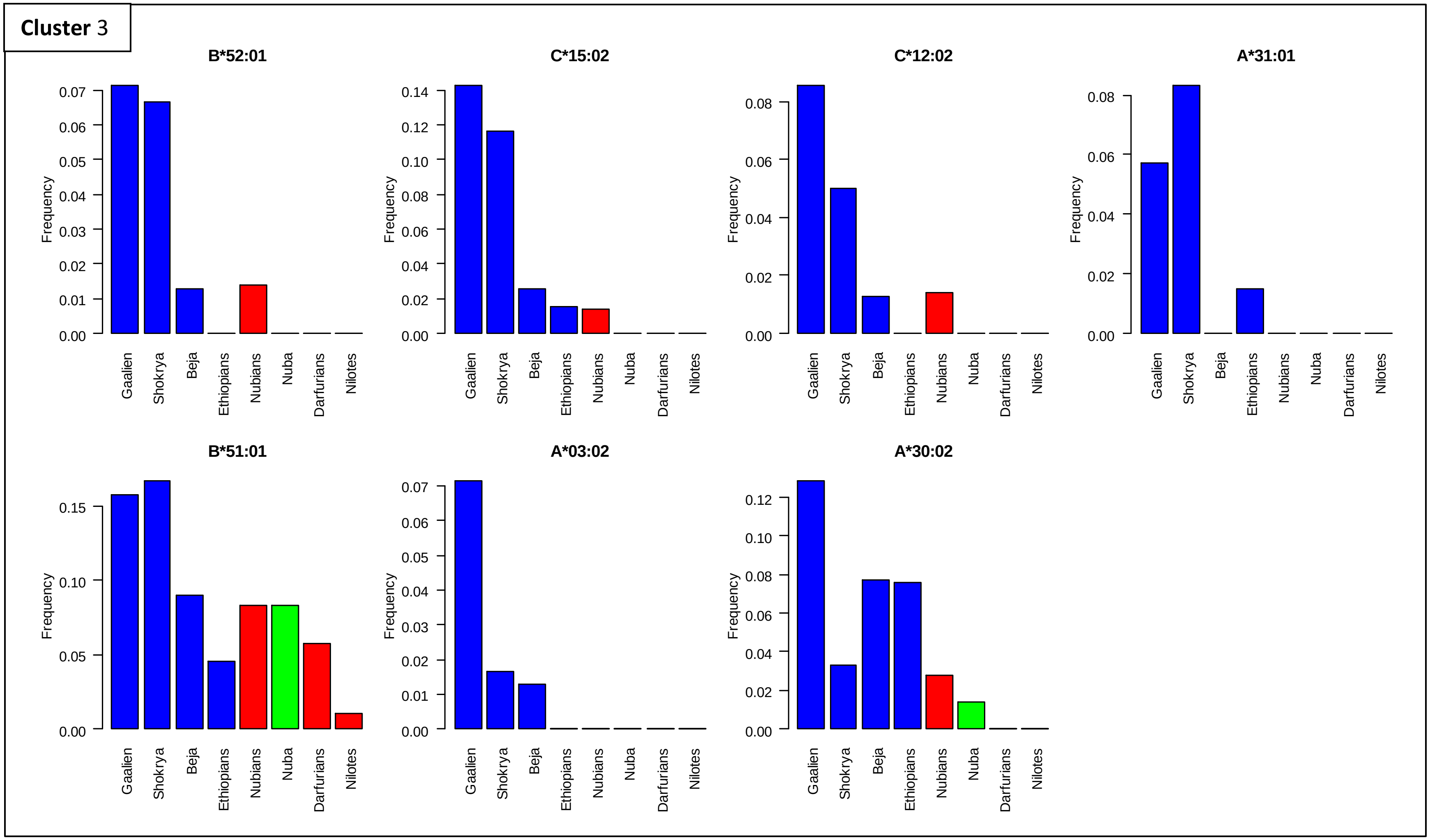

Figure 3.12: Frequency distribution of alleles from clusters 3 in the PCS plot of Figure 3.10

Colors of the bars correspond to allele frequencies in Nilo-Saharan (red), Afro-Asiatic (blue) and Nilo-Saharan/Niger-Congo (green) groups.

3.6 Finding common shared haplotypes

The previous PCA identified several alleles associated with population structure. Although the co-occurrence of these alleles in the same cluster may indicate their presence in the same haplotype, further investigation is needed to confirm this hypothesis. For this purpose, analysis based on shared haplotypes gives stronger evidence than just allele sharing because haplotypes represent single units (chromosomes) carried within populations. **Table 3.8** shows the most common three-locus HLA class I haplotypes among all populations. There was no single haplotype shared by more than three groups and most haplotypes were shared by only two groups. For some of these haplotypes, I found that their allelic compositions correlate with the previous PCA clustering patterns. In other words, alleles within the same haplotype are also found in the same cluster in PCS plots. For example, haplotypes H1, H4, H6, H18, H21, and H22 all had at least two alleles in cluster 1 (**Figure 3.10**). These haplotypes were characterized by high frequency in the Nilo-Saharan/Niger-Congo groups. Likewise, the haplotypes H10, H13, H14, H16, and H19 had alleles in cluster 3 and were mainly Arab haplotypes. On the other hand, fewer haplotypes had alleles residing in cluster 2 (H2 and H12). I also found that several haplotypes had alleles shared between more than one cluster (e.g., H3, H5, H7, and H17).

To follow the previously identified cluster 3 alleles in PCS plots, I focused on the most common haplotypes in the Arab groups. **Table 3.9** shows the most common haplotypes in Arabs that were also part of cluster 3. Interestingly, two of these haplotypes had the same alleles in *HLA-B* and *HLA-C* (H13 and H14; B*52:02 and C*12:02), but different alleles in *HLA-A*. The last finding of shared alleles suggested that an *HLA-B-HLA-C* haplotype might be maintained in the Arab groups; therefore, I estimated two-locus haplotypes in Gaalien and Shokrya groups. The most common haplotypes in these two groups are shown in **Table 3.10**. As I anticipated, one of the two most common haplotypes was composed of B*52:02 and C*12:02 alleles, with frequencies of 7.1% and 5% in the Gaalien and Shokrya groups respectively. Additionally, the haplotype B*51:01-C*15:02 was the most common, with frequencies of 11.4% and 11.7% in Gaalien and Shokrya respectively. Moreover, other haplotypes showed high frequencies in only one group (e.g., the B*50:01-C*06:02 haplotype in Shokrya and B*49:01-C*07:01 in Gaalien).

Table 3.8: The most common HLA class I haplotypes among the study populations*

ID	Haplotype			Arab		Beja	Ethiopians	Nubians	Nuba	Darfurians	Nilotes	Cluster
	A	B	C	Gaalien	Shokrya							
H1	A*01:01	B*08:01	C*07:04	0.00	0.00	0.00	0.00	0.00	1.39	0.00	5.44	1
H2	A*01:01	B*13:02	C*06:02	0.00	7.58	0.00	1.52	0.00	0.00	0.00	0.00	2
H3	A*02:01	B*13:02	C*06:02	0.00	0.00	6.41	0.00	0.00	0.00	0.00	0.00	-
H4	A*02:01	B*15:220	C*04:01	0.00	0.00	0.00	3.03	0.00	4.17	2.86	0.00	1
H5	A*02:01	B*41:01	C*07:01	0.00	0.00	0.00	0.00	0.00	0.00	4.29	0.00	-
H6	A*02:01	B*45:01	C*16:01	0.00	0.00	0.00	0.00	0.00	0.00	4.29	1.09	1
H7	A*02:01	B*51:01	C*16:01	1.43	0.00	1.28	0.00	0.00	4.17	0.00	0.00	-
H8	A*02:02	B*15:16	C*novel-3	0.00	0.00	0.00	0.00	4.17	0.00	0.00	0.00	-
H9	A*02:02	B*57:03	C*07:01	0.00	0.00	0.00	4.55	0.00	0.00	0.00	0.00	-
H10	A*02:05	B*07:02	C*07:02	4.29	0.00	1.28	0.00	0.00	0.00	0.00	0.00	3
H11	A*02:05	B*50:01	C*06:02	0.00	4.55	0.00	6.06	2.78	0.00	0.00	0.00	-
H12	A*03:01	B*47:01	C*06:02	0.00	0.00	7.69	0.00	0.00	0.00	0.00	0.00	-
H13	A*03:02	B*52:01	C*12:02	5.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3
H14	A*11:01	B*52:01	C*12:02	1.43	4.55	0.00	0.00	0.00	0.00	0.00	0.00	3
H15	A*30:01	B*49:01	C*07:01	0.00	0.00	0.00	1.52	6.94	0.00	0.00	0.00	-
H16	A*30:02	B*49:01	C*07:01	5.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3
H17	A*30:04	B*39:10	C*07:01	0.00	0.00	0.00	0.00	0.00	4.17	0.00	0.00	-
H18	A*30:04	B*53:01	C*04:01	0.00	0.00	0.00	0.00	0.00	0.00	4.29	0.00	1
H19	A*31:01	B*51:01	C*15:02	5.71	3.03	0.00	1.52	0.00	0.00	0.00	0.00	3
H20	A*33:03	B*50:01	C*06:02	0.00	6.06	0.00	0.00	0.00	0.00	0.00	0.00	-
H21	A*68:02	B*15:220	C*04:01	0.00	0.00	0.00	0.00	0.00	0.00	4.29	0.00	1
H22	A*68:02	B*47:03	C*07:18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.44	1

*(≥4% in at least one group)

Table 3.9: Common HLA class I haplotypes in the Arab groups*

ID	Haplotype			Arab		Beja	Ethiopians	Nubians	Nuba	Darfurians	Nilotes
	A	B	C	Gaalien	Shokrya						
H2	A*01:01	B*13:02	C*06:02	0.00	7.58	0.00	1.52	0.00	0.00	0.00	0.00
H10	A*02:05	B*07:02	C*07:02	4.29	0.00	1.28	0.00	0.00	0.00	0.00	0.00
H11	A*02:05	B*50:01	C*06:02	0.00	4.55	0.00	6.06	2.78	0.00	0.00	0.00
H13	A*03:02	B*52:01	C*12:02	5.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H14	A*11:01	B*52:01	C*12:02	1.43	4.55	0.00	0.00	0.00	0.00	0.00	0.00
H16	A*30:02	B*49:01	C*07:01	5.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H19	A*31:01	B*51:01	C*15:02	5.71	3.03	0.00	1.52	0.00	0.00	0.00	0.00
H20	A*33:03	B*50:01	C*06:02	0.00	6.06	0.00	0.00	0.00	0.00	0.00	0.00

*≥4% in at least one group, **ID**s are the same as in **Table 3.6**

Table 3.10: Most common two-locus haplotypes (*HLA-B* & *HLA-C*) in Arab groups*

Haplotype		**Arab**	
B	**C**	**Gaalien**	**Shokrya**
B*51:01	C*15:02	11.4	11.7
B*52:01	C*12:02	7.1	5.0
B*50:01	C*06:02	2.9	11.7
B*49:01	C*07:01	10.0	1.7
B*07:02	C*07:02	10.0	1.7
B*41:01	C*17:01	1.4	8.3
B*14:02	C*08:02	4.3	1.7
B*38:01	C*12:03	4.3	1.7
B*35:08	C*04:01	1.4	6.7
B*13:02	C*06:02	0.0	10.0
B*39:10	C*12:03	4.3	0.0

*≥4% in at least one group

3.7 Linkage disequilibrium between cluster 3 alleles

I previously found that high-frequency two-locus haplotypes are common among Arab groups, which suggested the maintenance of linkage disequilibrium (LD) between *HLA-B* and *HLA-C* loci. To confirm this, I estimated LD between allele pairs in cluster 3 haplotypes. **Figures 3.13** and **3.14** present pairwise LD coefficients (*D*`) of these haplotypes. In all pairwise comparisons between the *HLA-B* and *HLA-C* alleles, I observed high LD scores; in fact, for many pairs, LD values were 1, indicating complete linkage between these loci. On the other hand, the strength of LD between *HLA-A* and the other two loci was weaker. Furthermore, in the Gaalien group, three haplotypes showed strong LD between their alleles (H10, H13, and H19); however, in the Shokrya group, only H13 had a high LD score. Focusing on the most common two-locus haplotypes in **Table 3.9**, I found complete LD between B*52:01 and C*12:02 alleles. Furthermore, the allelic pair B*51:01-C*15:02 also had a high LD score, although it was lower than that of the former pair.

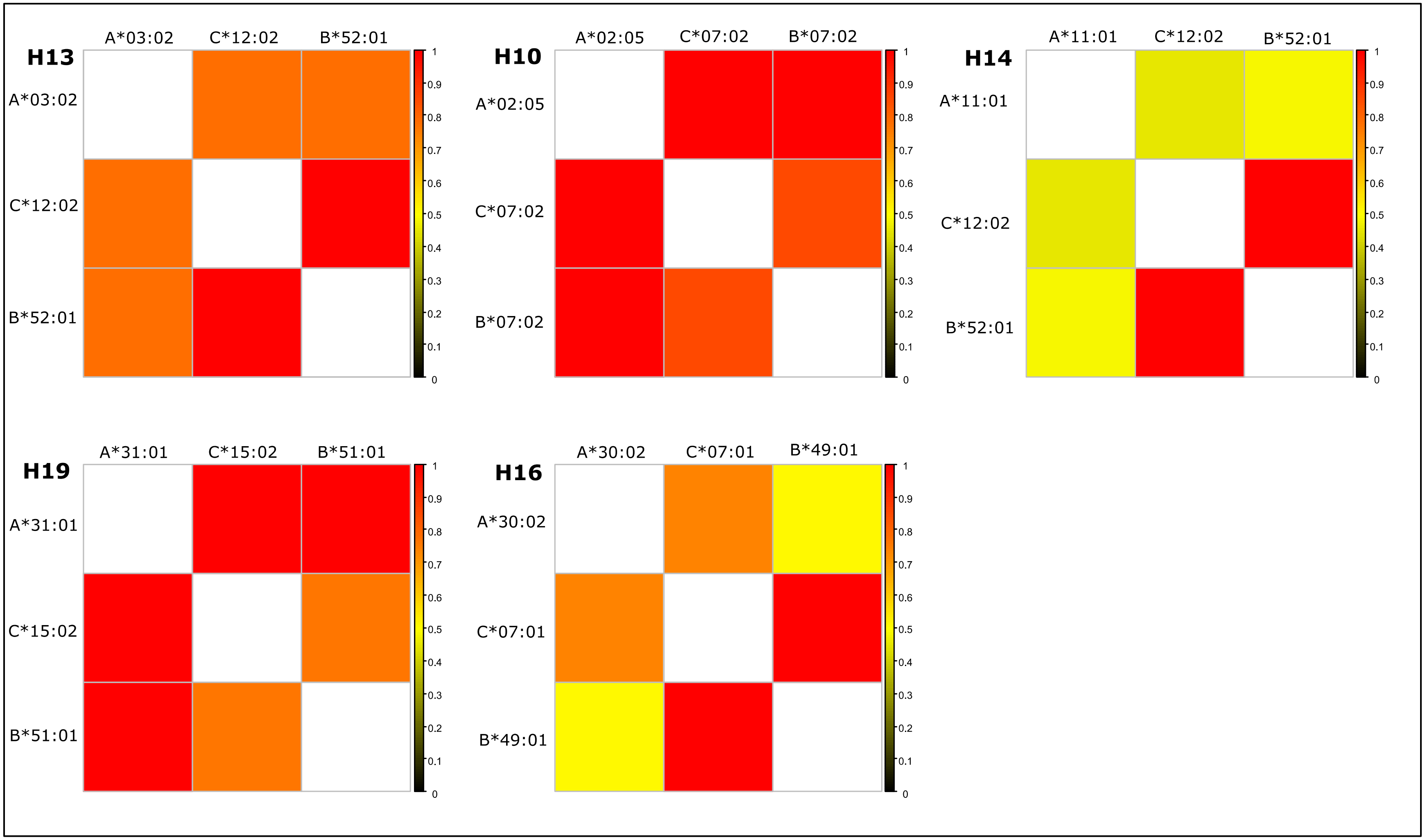

Figure 3.13: Pairwise linkage disequilibrium (*D*`) between HLA alleles of common haplotypes in the Gaalien group

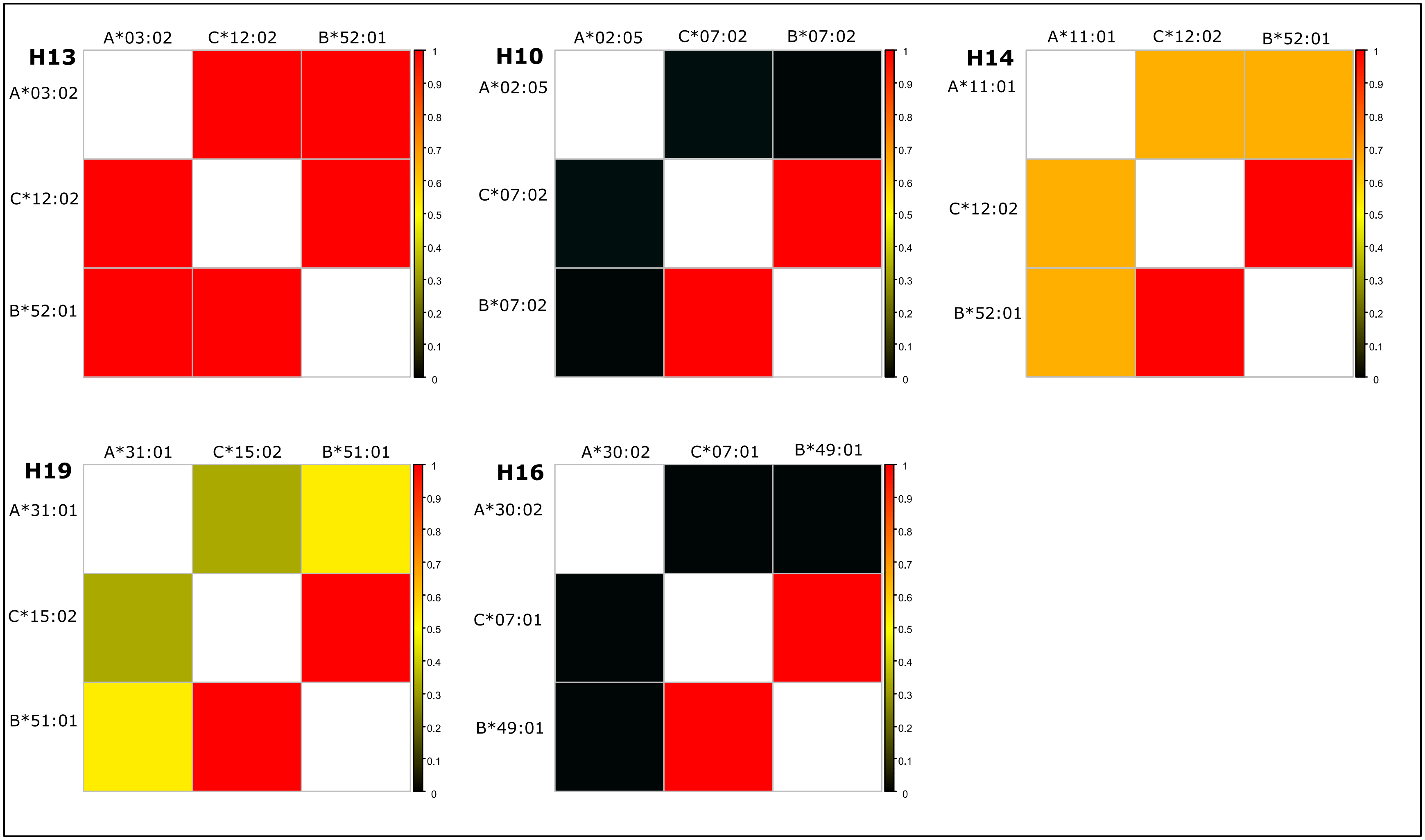

Figure 3.14: Pairwise linkage disequilibrium (*D*`) between HLA alleles of common haplotypes in the Shokrya group

3.8 Searching for the ancestral haplotypes of Arab groups

In the previous analyses, starting from the PCA to the last LD patterns, my findings suggest that Arab groups have shared components that distinguish them from other Afro-Asiatic groups. In particular, the strongly linked and most common Arab haplotypes seem to be informative for tracing the ancestry of these groups. Bearing this in mind, I searched the AFND database for the two-locus haplotypes B*51:01-C*15:02 and B*52:01-C*12:02. **Table 3.11** shows the prevalence of these haplotypes in various populations worldwide. As shown in this table, in both haplotypes, unsurprisingly, there was no single population from Sub-Saharan Africa and the only African population was from Tunisia, which linguistically belongs to the Afro-Asiatic family. Furthermore, all of the other populations were from Asia, Europe, or North America, suggesting that these haplotype might be non-African haplotypes. Interestingly, the haplotype B*51:01-C*15:02 has a frequency of 4.7% among the Saudi population.

I then searched the AFND for the extended three-locus haplotypes I previously found in the Arab groups (i.e., A*11:01-C*12:02- B*52:01 and A*03:02-C*12:02- B*52:01). The distributions of these haplotypes in various populations are shown in **Table 3.12.** Like in the case of two-locus haplotypes, there was no single Sub-Saharan African population with these haplotypes and all of the populations were mainly non-Africans. In particular, the haplotype A*11:01-C*12:02-B*52:01 is widely spread among Asians and Europeans.

Table 3.11: Frequency of B*51:01-C*15:02 and B*52:01-C*12:02 haplotypes in various populations*

B*51:01-C*15:02		B*52:01-C*12:02	
Population	**Frequency (%)**	**Population**	**Frequency (%)**
USA North American Native	7.0	Japan Central	10.5
Mexico Oaxaca Mixtec	3.0	USA Asian pop 2	3.3
Tunisia	2.5	Myanmar China	2.7
Saudi Arabia¶	4.7	South Korea pop 3	2.4
USA Hispanic	2.4	Tunisia	2.0
USA Hispanic pop 2	2.2	USA Hispanic	1.9
Italy pop 5	2.1	USA NMDP Korean	1.8
Mexico City Mestizo pop 2	1.9	Myanmar Chin	1.8
USA Caucasian pop 2	1.9	Myanmar Kayah	1.8
Mexico Oaxaca Zapotec	1.5	South Africa Caucasians	1.1
Iran Baloch	2.2	Myanmar Rakhine	1.0
Japan Central	1.1	England North West	1.0
Ireland Northern	1.1		
Myanmar Rakhine	1.0		
USA Alaska Yupik	1.0		

*Only haplotypes with a frequency of ≥1% are shown here. ¶From Hajeer *et al*. (2013).

Table 3.12: Frequency of A*03:02-C*12:02-B*52:01 and A*11:01-C*12:02-B*52:01 haplotypes in various populations

A*03:02-C*12:02-B*52:01			A*11:01-C*12:02-B*52:01		
Population	Frequency (%)	sample size	Population	Frequency (%)	sample size
USA Asian pop 2	0.04	1,772	Myanmar Chin	2.73	55
Germany Turkey min	0.03	4,856	USA NMDP South Asian Indian	0.83	185,391
			USA NMDP Southeast Asian	0.61	27,978
			USA NMDP Middle Eastern or North Coast of Africa	0.57	70,890
			Germany DKMS - Turkey minority	0.56	4,856
			Germany DKMS - Greece minority	0.39	1,894
			USA Asian pop 2	0.38	1,772
			Poland DKMS	0.37	20,653
			Brazil Vale do Ribeira Quilombos	0.7	144
			Germany DKMS - Croatia minority	0.56	2,057
			USA Hispanic pop 2	0.33	1,999
			USA NMDP South Asian Indian	0.54	185,391
			USA NMDP Mexican or Chicano	0.25	261,235
			Germany DKMS - Romania minority	0.44	1,234
			Germany DKMS - Italy minority	0.21	1,159

To understand the population structure in a wider context, I integrated the east African data and 12 additional data sets from different populations deposited in the AFND database (a total of 20 data sets). PCA analysis of the integrated data revealed that, among Sudanese groups, Gaalien and Shokrya (Arab groups) were the closest to all non-African populations (**Figure 3.14**). Interestingly, the closest population to the Sudanese Arabs was Arabs from the Middle East (two data sets named Saudi_Guraiat_Hail and Saudi_pop_5). Generally, all Sudanese groups and Ethiopians were located between non-Africans and Sub-Saharan Africans. Moreover, the Nilo-Saharan/Niger-Congo Sudanese groups (Darfurians, Nuba, and Nilotes) clustered close to Sub-Saharan Africans. Similarly, the Afro-Asiatic groups were seen as one cluster; however the only exception was the Nubian group, which is consistently found close to the Afro-Asiatic groups. In the same context, two groups in the AFND data sets (Sudan_mixed and Tunisia_mixed) with no specific ethnic affiliation clustered close to the Afro-Asiatic family.

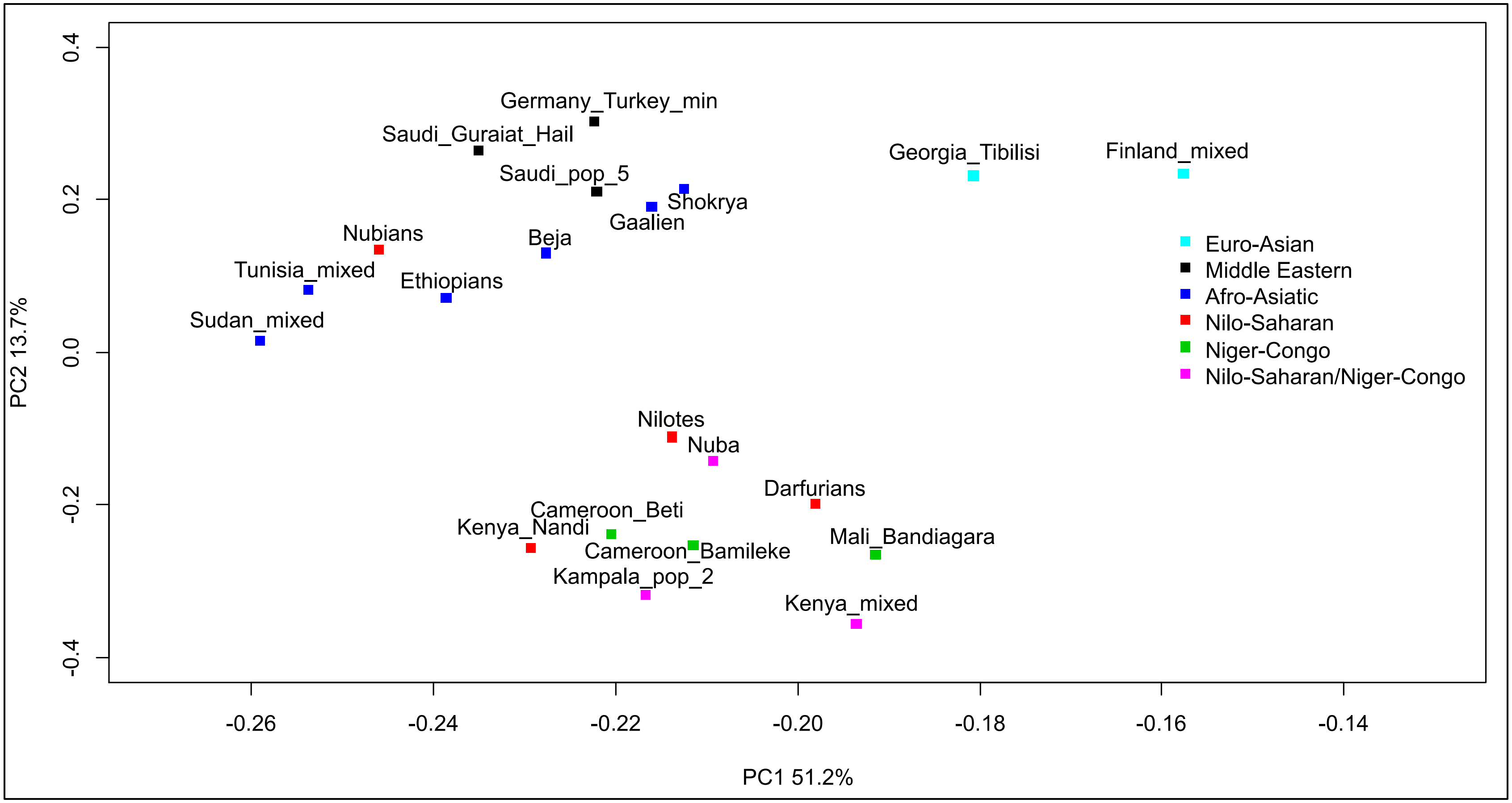

Figure 3.15: PCA of class I HLA genes in populations from Africa, the Middle East, and Europe

Chapter Four

DISCUSSION

East Africa is an interesting place to study genetic diversity in humans. As one of the candidates for where modern humans potentially originated and evolved (Tishkoff and Williams 2002), East Africa is characterized by high levels of linguistic, cultural, and genetic diversity (Tishkoff *et al*. 2009). Here, using samples from three East African countries (Sudan, South Sudan, and Ethiopia), representing eight ethnic groups, I studied the diversity of HLA class I genes. The several analyses performed indicated that diverse populations are living in East Africa and those populations have been affected by their neighboring populations, particularly from West Asia and the Middle East.

4.1 HLA class I diversity of East African groups

The several measures of diversity that I examined in this study indicate a high level of genetic diversity in the Sudanese population. Such diversity is evident from the large number of observed alleles per locus (**Table 3.1**), the high heterozygosity values (**Table 3.1**), and the substantial population differentiation as measured by the F_{st} statistic (**Table 3.2**). In terms of allele numbers and based on the analysis of 55 African populations, the east African groups have comparable number of alleles to the other Sub-Saharan African populations (Solberg *et al*. 2008). Furthermore, a large number of alleles and a high level of heterozygosity in class I genes could be an indication of balancing selection acting on these loci.

The F_{st} statistic showed that the study populations are differentiated at the level of macro-families (i.e., between Afro-Asiatic and Nilo-Saharan/Niger-Congo), suggesting limited gene flow between the different linguistic families.

Furthermore, test of selective neutrality based on Ewens-Watterson homozygosity test showed that, in the all populations with significant Fnd values (negative Fnd), homozygosity deviated from expectations under neutrality assumption, suggesting that balancing selection is acting on these loci. This observation is consistent with the findings from several other populations (Buhler *et al*. 2011).

In this study, I also found new HLA alleles in *HLA-B* and *HLA-C* genes, which were not previously reported (**Table 3.4**). The populations in which the new alleles were found belong to the Nilo-Saharan family (Nubians and Darfurians), suggesting that more alleles might be found in this family. It is possible that the genetics of the Nilo-Saharan family is not well studied because Dobon *et al*. (2015) found a new Nilo-Saharan component that was not identified in the large-scale study of Tishkoff *et al*. (2009).

4.2 Genetic structure of East African groups

Sudan is a country with very diverse populations, as indicated from the cultural and linguistic diversity present there. Dobon *et al.* (2015) recently studied the genetic structure of Sudanese populations using genome wide SNP data and found two main substructures define the diversity of eight East African groups. The two identified substructures were related to a north-east/south-west division and correlated with linguistic classification of the studied groups (Afro-Asiatic on one side and Nilo-Sahara/Niger-Congo on the other). The same study also revealed a Middle Eastern genetic component in the Afro-Asiatic groups from East Africa. In this study, I used the same set of samples studied by Dobon *et al*. (2015) and genotyped the highly polymorphic HLA class I genes. I aimed to expand upon the previous study by finding further population substructures within their identified clusters. The polymorphic HLA class I genes are very useful to study the population structure and have been proven effective in the admixed Japanese population by Nakaoka *et al*. (2013).

Of the several approaches used to study genetic diversity, PCA is a useful and powerful tool to delineate the population structure. I employed this method to dissect the structure of the study populations, which revealed three distinct clusters of HLA class I alleles (**Figure 3.6**). Two of the identified clusters define substructures that correlate with the linguistic affiliations of the study groups, an observation previously seen in other data types such as SNP data (Dobon *et al*. 2015). The first cluster (cluster 1) is associated with the Nilo-Saharan/Niger-Congo groups and some of the defining alleles in this cluster are common across Sub-Saharan African populations (i.e., A*30:04, B*42:04, and B*47:03) (**Figure 3.7**) (Cao *et al*. 2004). The Nilo-Saharan family is the main linguistic family in Sudan as most of the languages in the country belong to this macro-linguistic family (Greenberg 1963). The Niger-Congo (also called Niger-Kordofanian), however, is geographically restricted to southern Sudan, with a few groups in the Nuba Mountains. Three groups in this study were related to cluster 1: Darfurians in western Sudan, Nuba in southern Sudan, and Nilotes from South Sudan (the country). Although these groups are geographically and currently linguistically separate (at the subfamily level), they exhibit a shared component, which suggests their past genetic relatedness. Hassan *et al.* (2008) found that the Y-chromosome haplogroups A-M13 and B-M60 are common among the three groups. Furthermore, using the Mantel test, the same authors suggested that Nilo-Saharan groups have low levels of gene flow from other groups/families. The second cluster (cluster 2) is defined by alleles that are common in the Afro-Asiatic groups, in addition to Nubians (Nilo-Saharan) (**Figure 3.8**). The Afro-Asiatic family, geographically, spans the Sahara, and North and East Africa. Four groups in this study belong to the Afro-Asiatic family. Two of these four groups speak the Arabic language, so were considered as Arabs and they have settled in the central and eastern parts of Sudan. The other two groups are Beja, living by the Red Sea, and Ethiopians from Ethiopia. Only Beja and Ethiopians, out of the four, were in cluster 2. The grouping of Beja and Ethiopians into one cluster is consistent with previous findings, which showed a shared common Y-chromosome haplogroup (J1) (Hassan *et al*. 2008). Interestingly, this shared haplogroup is known to be of Euro-Asian origin (Giacomo *et al*. 2004), suggesting that both groups were subjected to gene flow from non-African populations. The second point of interest in cluster 2 was the clustering of the Nilo-Saharan group Nubians close to the Afro-Asiatic groups Ethiopians and Beja. The deviation of Nubians from their linguistic family (i.e., Nilo-Saharan) is not unexpected, as other studies have shown similar clustering patterns (Hassan *et al*. 2008; Dobon *et al*. 2015). Tishkoff *et al*. (2009) observed that, in the Nilo-Saharan family, geography has a better correlation with genetic diversity than linguistic class, possibly due to admixture with other groups. The last possibility is supported by the fact that Nubians occupy the north side entry point to Sudan, on several migration routes and the site occupations that came from Egypt. The third cluster (cluster 3) was defined by alleles predominant in the two Arab groups (**Figure 3.12**). In all previous analyses using different data types (Hassan *et al*. 2008; Dobon *et al*. 2015), Afro-Asiatic groups were clustered as single group; however, my analysis revealed a substructure within this family. The alleles that define this third cluster were almost unique to the Arab groups, so I thought that these alleles are informative for tracing their ancestry.

Although I identified a substructure within the Afro-Asiatic family, the interpretation of this finding is not straight forward because SNP data from Dobon *et al*. (2015) did not show similar clustering patterns. Several possibilities may explain this discrepancy: First, Dobon *et al*. performed PCA using very low number of markers to differentiate between genetically distant populations like Sub-Saharan, East African and Middle Eastern groups. Because the maximum variance in PCA is explained by the first and second principal components, Dobon's 1st two PCs differentiated between Nilo-Saharan and Afro-Asiatic groups. Therefore, the minor variations within the Afro-Asiatic family were not clear in the first PCs and it could be identified if they looked further into the other PCs. Second, it possible that the substructure identified based on the HLA data is not associated with a substructure at the genomic level because the ancestry-informative alleles were not shared by all individuals in the Arab groups; although these alleles have high frequency. In such case, my identified substructure could be due significant gene flow from other populations as mentioned earlier.

4.3 Searching for the ancestry-informative haplotypes of Arab groups

To follow the interesting cluster 3 alleles, I constructed haplotypes based on class I genes, which revealed many of the cluster 3 alleles were found in the same haplotype (**Table 3.9 & Figure 3.10**). Furthermore, the most common haplotypes among Arab groups included those found in cluster 3 (**H13**, **H14** and **H19**) (**Table 3.9**). Focusing on the most common haplotypes in cluster 3, I found two of the haplotypes (**H13** and **H14**) have the same allele in the *HLA-B* and *HLA-C* genes, but different alleles in *HLA-A*, which suggested LD decay between HLA-A and the other two loci. To confirm the last idea of LD decay, I estimated two-locus haplotypes between *HLA-B* and *HLA-C*, which showed that one of the most common two-locus haplotypes included the two alleles I that found in the **H13** and **H14** haplotypes (i.e., B*52:01 and C*12:02) (**Table 3.10**). I then estimated LD between allele pairs of the most common three-locus haplotypes among Gaalien and Shokrya (Arab groups). I found strong LD between allele pairs of *HLA-B* and *HLA-C* and to a lesser extent between

HLA-A and the other two B and C loci (**Figure 3.13 & Figure 3.14**). The strong observed LD confirmed the previous idea that *HLA-B* and *HLA-C* are tightly linked in these groups.

To investigate further the relationship between Sudanese populations in a wider context, I integrated data from AFND representing populations from Sub-Saharan Africa, the Middle East and Europe. PCA using the integrated data was consistent with the clustering patterns in the first PCA of east African samples (**Figure 3.15**). Moreover in the PCA of the combined data sets, I found cluster 1 groups (Darfurians, Nuba, and Nilotes) clustered very close to the other Sub-Saharan Africans. I also found the Afro-Asiatic groups Ethiopians and Beja in addition to Nubians clustered close to each other, which was in line with the first PCA. Intriguingly, the two Arab groups of Sudan (Gaalien and Shokrya) were the closest groups to Middle Eastern populations including Saudis and Turks.

Taken together, the previous analyses suggest the non-African origin of the haplotypes/alleles associated with the substructure defined by cluster 3. Historically, Sudan witnessed several waves of migration and occupation from neighboring countries such as Saudi Arabia and Egypt, which may have contributed to the current peopling of Sudan (Metz 1991). People who speak Arabic language in Sudan claim their Arab ancestry; however, such claims need further confirmation as many other Sudanese groups make the same claims. Interestingly, I found the alleles that comprise the most common two-locus haplotypes among Sudanese Arabs have a very high frequency in the Saudi population. Hajeer *et al*. (2013) found that the frequencies of B*50:01 and C*15:02 alleles among the Saudi populations were 15.8% and 8.7%, respectively. Among their interesting findings of relevance to this study are the high haplotypes frequencies of B*50:01-C*06:02 (12.9%), B*51:01-C*15:02 (4.7%), and B*07:02-C*07:02 (4%), which also have high frequencies among the Arab groups, as shown in **Table 3.10**. The proximity of East Africa to the Arabian Peninsula facilitates demographic movement and therefore gene flow between populations in these two regions. Several studies have provided evidence for gene flow from the Arabia to East Africa and *vice versa* (Gandini *et al*. 2016; Busby *et al*. 2016). Furthermore, the global distribution of the two-locus haplotypes (B*51:01-C*15:02 and B*52:01-C*12:02) in **Table 3.11** clearly show these are not common to Sub-Saharan African as all of the populations are from Asia, Europe of North America, except for Tunisia from North Africa. In the same line, the distributions of three-locus haplotypes (A*03:02-C*12:02-B*52:01 and A*11:01-C*12:02-B*52:01) were geographically restricted to Asia and Europe (**Table 3.12**). Of particular interest is the two-locus haplotype C*12:02-B*52:01, which is very common in Japanese as well as in the Arabs from Sudan. Extending this haplotype to three loci (i.e., A*11:01-C*12:02-B*52:01 or A*03:02-C*12:02-B*52:01) was enough to show the difference between those populations; in Japanese the haplotype C*12:02-B*52:01 is linked to A*24:02, while in the Sudanese Arab it is linked to either A*11:01 or A*03:02.

Although the direction of gene flow proposed in this study is in only one direction (i.e., from West Asia/Middle East to East Africa), It is also possible that the identified alleles/haplotypes originated in East Africa then brought to Asia in the early migration of modern humans. However, the last idea is challenged by the absence of these ancestry-informative haplotypes in the other Afro-Asiatic groups such as Beja and Ethiopians because old haplotypes are expected to be common among the whole family and not restricted to only one or two groups.

Conclusions

In this study, I investigated the genetic diversity of HLA class I genes in eight East African populations.

I first established allele frequency distributions among each group. Using the established distributions, I constructed haplotypes between the different HLA genes. I report the identified alleles and haplotypes as a map of HLA diversity in Sudan. Because this study features Sudanese groups from a wide geographical area, the identified alleles and haplotypes might be of interest for many people in the fields that are relevant to HLA. For example, the fact that ethnic groups usually inhabit a specific location, it is possible to predict HLA allele and haplotype frequencies in that area. For this reason, health policymakers and transplantation donor programs may set their policies and priorities according to the wide map of allele frequency distribution established in this study.

Using the allele frequency data, I employed the PCA method to dissect further the structure of Sudanese groups and Ethiopian samples. The patterns of clustering that I found are consistent with previous data; however, I identified a substructure within the Afro-Asiatic family that has not been identified before from SNP data. The substructure was defined by Arab groups who may have migrated to Sudan or experienced significant gene flow in the past, probably from the West Asia. This finding is valuable for anthropological studies seeking to study populations' history and demography.

References

- Ahmadloo, S., Nakaoka, H., Hayano, T., Hosomichi, K., You, H., Utsuno, E., Sangai, T., Nishimura, M., Matsushita, K., Hata, A., Nomura F., & Inoue, I. (2017). Rapid and cost effective high-throughput sequencing for identification of germline mutations of BRCA1 and BRCA2. J Hum Genet, in press.
- Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. Genome Res, 11(6), 1005-1017.
- Blackwell, J. M., Jamieson, S. E., & Burgner, D. (2009). HLA and infectious diseases. Clin Microbiol Rev, 22(2), 370-385.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.
- Buhler S., & Sanchez-Mazas A. (2011). HLA DNA Sequence Variation among Human Populations: Molecular Signatures of Demographic and Selective Events. PLoS ONE 6(2): e14643.
- Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., Amenga-Etego, L. N., Enimil, A., Apinjoh, T., Ndila, C. M., Manjurano, A., Nyirongo, V., Doumba, O., Rockett, K. A., Kwiatkowski, D. K., Spencer, C. C. A., & Malaria Genomic Epidemiology Network (2016). Admixture into and within sub-Saharan Africa. ELife 5, e15266.
- Cao, K., Moormann, A. M., Lyke, K. E., Masaberg, C., Sumba, O. P., Doumbo, O. K., Koech, D., Lancaster, A., Nelson, M., Meyer, D., Single, R., Hartzman, R. J., Plowe, C. V., Kazura, J., Mann, D. L., Sztein, M. B., Thomson, G., & Fernández-Viña. M. A. (2004). Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens, 63(4), 293-325.

- Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K., & O'Brien, S. J. (1999). HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. Science, 283(5408), 1748-1752.
- Dafalla, A. M., McCloskey, D. J., Alemam, A. A., Ibrahim, A. A., Babikir, A. M., Gasmelseed, N., El Imam, M., Mohamedani, A. A., & Magzoub, M. M. (2011). HLA polymorphism in Sudanese renal donors. Saudi J Kidney Dis Transpl, 22(4), 834-840.
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 43(5), 491-498.
- Di Giacomo, F., Luca, F., Popa, L. O., Akar, N., Anagnou, N., Banyko, J., Brdicka, R., Barbujani, G., Papola, F., Ciavarella, G., Cucci, F., Di Stasi, L., Gavrila, L., Kerimova, M. G., Kovatchev, D., Kozlov, A. I., Loutradis, A., Mandarino, V., Mammi, C., Michalodimitrakis, E. N., Paoli, G., Pappa, K. I., Pedicini, G., Terrenato, L., Tofanelli, S., Malaspina, P., & Novelletto, A. (2004). Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet, 115(5), 357-371.
- Dobon, B., Hassan, H. Y., Laayouni, H., Luisi, P., Ricaño-Ponce, I., Zhernakova, A., Wijmenga, C., Tahir, H., Comas, D., Netea, M. G., & Bertranpetit, J. (2015). The genetics of East African populations: a Nilo-Saharan component in the African genetic landscape. Sci Rep, 5, 9996.
- Elhassan, N., Gebremeskel, E. I., Elnour, M. A., Isabirye, D., Okello, J., Hussien, A., Kwiatksowski, D., Hirbo, J., Tishkoff, S., & Ibrahim, M. E. (2014). The episode of genetic drift defining the migration of humans out of Africa is derived from a large east African population size. PLoS One, 9(5), e97674.

- Ewens, W. J., (1972). The sampling theory of selectively neutral alleles. Theor Pop Biol, 3(1), 87-112.
- Excoffier, L., & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol, 12(5), 921-927.
- Fernandez Vina, M. A., Hollenbach, J. A., Lyke, K. E., Sztein, M. B., Maiers, M., Klitz, W., Cano, P., Mack, S., Single, R., Brautbar, C., Israel, S., Raimondi, E., Khoriaty, E., Inati, A., Andreani, M., Testi, M., Moraes, M. E., Thomson, G., Stastny, P., & Cao, K. (2012). 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, 367(1590), 820-829.
- Flajnik, M. F., & Kasahara, M. (2001). Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. Immunity, 15(3), 351-362.
- Flajnik, M. F., Ohta, Y., Namikawa-Yamada, C., & Nonaka, M. (1999). Insight into the primordial MHC from studies in ectothermic vertebrates. Immunol Rev, 167, 59-67.
- Gandini, F., Achilli, A., Pala, M., Bodner, M., Brandini, S., Huber, G., Egyed, B., Ferretti, L., Gómez-Carballa, A., Salas, A., Scozzari, R., Cruciani, F., Coppa, A., Parson, W., Semino, O., Soares, P., Torroni, A., Richards, M. B., & Olivieri, A. (2016). Mapping human dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes. Sci Rep, 6, 25472.
- Gaudieri, S., Dawkins, R. L., Habara, K., Kulski, J. K., & Gojobori, T. (2000). SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. Genome Res, 10(10), 1579-1586.
- González-Galarza, F. F., Takeshita, L. Y., Santos, E. J., Kempson, F., Maia, M. H., da Silva, A. L., Teles e Silva, A. L., Ghattaoraya, G. S., Alfirevic, A., Jones, A. R., & Middleton, D. (2015). Allele frequency net 2015 update: new features for HLA epitopes,

KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res, 43(Database issue), D784-788.

- Gough, S. C., & Simmonds, M. J. (2007). The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. Curr Genomics, 8(7), 453-465.
- Greenberg, J. H. (1963). The languages of Africa. International Journal of American Linguistics, Bloomington, Indiana University press.
- Guo, S. W., & Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics, 48(2), 361-372.
- Hajeer, A. H., Al Balwi, M. A., Aytül Uyar, F., Alhaidan, Y., Alabdulrahman, A., Al Abdulkareem, I., & Al Jumah, M. (2013). HLA-A, -B, -C, -DRB1 and -DQB1 allele and haplotype frequencies in Saudis using next generation sequencing technique. Tissue Antigens, 82(4), 252-258.
- Hassan, H. Y., Underhill, P. A., Cavalli-Sforza, L. L., & Ibrahim, M. E. (2008). Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history. Am J Phys Anthropol, 137(3), 316-323.
- Hernández-Frederick, C. J., Giani, A. S., Cereb, N., Sauter, J., Silva-González, R., Pingel, J., Schmidt, A. H., Ehninger, G., & Yang, S. Y. (2014). Identification of 2127 new HLA class I alleles in potential stem cell donors from Germany, the United States and Poland. Tissue Antigens, 83(3), 184-189.
- Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., Povey, S., Talbot, C. C. Jr., Wright, M. W., Wain, H. M., Trowsdale, J., Ziegler, A., & Beck, S. (2004). Gene map of the extended human MHC. Nat Rev Genet, 5(12), 889-899.
- Hosomichi, K., Jinam, T. A., Mitsunaga, S., Nakaoka, H., & Inoue, I. (2013). Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics, 14, 355.

- Hughes, A. L., & Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature, 335(6186), 167-170.
- Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., *et al*., (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet, 41(6), 657-665.
- Kawashima, M., Ohashi, J., Nishida, N., & Tokunaga, K. (2012). Evolutionary analysis of classical HLA class I and II genes suggests that recent positive selection acted on DPB1*04:01 in Japanese population. PLoS One, 7(10), e46806.
- Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res, 12(4), 656-664.
- Kochi, Y., Suzuki, A., Yamada, R., & Yamamoto, K. (2010). Ethnogenetic heterogeneity of rheumatoid arthritis-implications for pathogenesis. Nat Rev Rheumatol, 6(5), 290-295.
- Lancaster, A. K., Single, R. M., Solberg, O. D., Nelson, M. P., & Thomson, G. (2007). PyPop update--a software pipeline for large-scale multilocus population genomics. Tissue Antigens, 69 Suppl 1, 192-197.
- Lau, Q., Yasukochi, Y., & Satta, Y. (2015). A limit to the divergent allele advantage model supported by variable pathogen recognition across HLA-DRB1 allele lineages. Tissue Antigens, 86(5), 343-352.
- Lewis, M. P., Gary, F. S., & Charles, D. F., (2016). Ethnologue: Languages of the World. (19th ed.) Dallas, Texas: from SIL International http://www.ethnologue.com, Retrieved 01/12/2016.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754-1760.

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079.
- Magzoub, M. M., Stephens, H. A., Sachs, J. A., Biro, P. A., Cutbush, S., Wu, Z., & Bottazzo, G. F. (1992). HLA-DP polymorphism in Sudanese controls and patients with insulin-dependent diabetes mellitus. Tissue Antigens, 40(2), 64-68.
- Marsh, S. G., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., Fernández-Viña, M., Geraghty, D. E., Holdsworth, R., Hurley, C. K., Lau, M., Lee, K. W., Mach, B., Maiers, M., Mayr, W. R., Müller, C. R., Parham. P., Petersdorf, E. W., Sasazuki, T., Strominger, J. L., Svejgaard, A., Terasaki, P. I., Tiercy, J. M., & Trowsdale, J. (2010). Nomenclature for factors of the HLA system, 2010. Tissue Antigens, 75(4), 291-455.
- McGuire, W., Knight, J. C., Hill, A. V., Allsopp, C. E., Greenwood, B. M., & Kwiatkowski, D. (1999). Severe malarial anemia and cerebral malaria are associated with different tumor necrosis factor promoter alleles. J Infect Dis, 179(1), 287-290.
- Metz, H. C., Library of Congress. Federal Research Division, & Thomas Leiper Kane Collection (1992). Sudan: A Country Study. (4th ed.), Washington D.C., Federal Research Division, Library of Congress.
- Nakaoka, H., Mitsunaga, S., Hosomichi, K., Shyh-Yuh, L., Sawamoto, T., Fujiwara, T., Tsutsui, N., Suematsu, K., Shinagawa, A., Inoko, H., & Inoue, I. (2013). Detection of ancestry informative HLA alleles confirms the admixed origins of Japanese population. PLoS One, 8(4), e60793.
- Neefjes, J., Jongsma, M. L., Paul, P., & Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol, 11(12), 823-836.

- Nei, M., Gu, X., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A, 94(15), 7799-7806.
- Owen, J. A., Punt, J., Stranford, S. A., Jones, P. P., & Kuby, J. (2013). Kuby immunology (7th ed.). New York: W.H. Freeman.
- Peaper, D. R., & Cresswell, P. (2008). Regulation of MHC class I assembly and peptide binding. Annu Rev Cell Dev Biol, 24, 343-368.
- Penn, D. J., Damjanovich, K., & Potts, W. K. (2002). MHC heterozygosity confers a selective advantage against multiple-strain infections. Proc Natl Acad Sci U S A, 99(17), 11260-11264.
- Relethford, J. (2001). Genetics and the search for modern human origins. New York: Wiley-Liss.
- Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., & Marsh, S. G. (2015). The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res, 43(Database issue), D423-431.
- Robinson, J., Halliwell, J. A., McWilliam, H., Lopez, R., Parham, P., & Marsh, S. G. (2013). The IMGT/HLA database. Nucleic Acids Res, 41(Database issue), D1222-1227.
- Rudolph, M. G., Stanfield, R. L., & Wilson, I. A. (2006). How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol, 24, 419-466.
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol, 4(4), 406-425.
- Salamon, H., Klitz, W., Easteal, S., Gao, X., Erlich, H. A., Fernandez-Vina, M., Trachtenberg, E. A., McWeeney, S. K., Nelson, M. P., & Thomson, G. (1999). Evolution of HLA class II molecules: Allelic and amino acid site variability across populations. Genetics, 152(1), 393-400.

- Sanjeevi, C. B., Lybrand, T. P., DeWeese, C., Landin-Olsson, M., Kockum, I., Dahlquist, G., Dahlquist, G., Sundkvist, G., Stenger, D., & Lernmark, A. (1995). Polymorphic amino acid variations in HLA-DQ are associated with systematic physical property changes and occurrence of IDDM. Members of the Swedish Childhood Diabetes Study. Diabetes, 44(1), 125-131.
- Slatkin, M. (1994). An exact test for neutrality based on the Ewens sampling distribution. Genet Res, 64(1), 71-74.
- Solberg, O. D., Mack, S. J., Lancaster, A. K., Single, R. M., Tsai, Y., Sanchez-Mazas, A., & Thomson, G. (2008). Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum Immunol, 69(7), 443-464.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol, 30(12), 2725-2729.
- The MHC sequencing consortium. (1999). Complete sequence and gene map of a human major histocompatibility complex. Nature, 401(6756), 921-923.
- Thursz, M. R., Thomas, H. C., Greenwood, B. M., & Hill, A. V. (1997). Heterozygote advantage for HLA class-II type in hepatitis B virus infection. Nat Genet, 17(1), 11-12.
- Tishkoff, S. A., & Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet, 3(8), 611-621.
- Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L., & Oefner, P. J. (2000). Y chromosome sequence variation and the history of human populations. Nat Genet, 26(3), 358-361.
- Watterson, G. A., (1978). The homozygosity test of neutrality. Genetics, 88(2), 405-417.

- Zhao, L. P., Alshiekh, S., Zhao, M., Carlsson, A., Larsson, H. E., Forsander, G., Ivarsson, S. A., Ludvigsson, J., Kockum, I., Marcus, C., Persson, M., Samuelsson, U., Örtqvist, E., Pyo, C. W., Nelson, W. C., Geraghty, D. E., Lernmark, Å., & Better Diabetes Diagnosis (BDD) Study Group. (2016). Next-Generation Sequencing Reveals That HLA-DRB3, -DRB4, and -DRB5 May Be Associated With Islet Autoantibodies and Risk for Childhood Type 1 Diabetes. Diabetes, 65(3), 710-718.

Appendices

Appendix 1

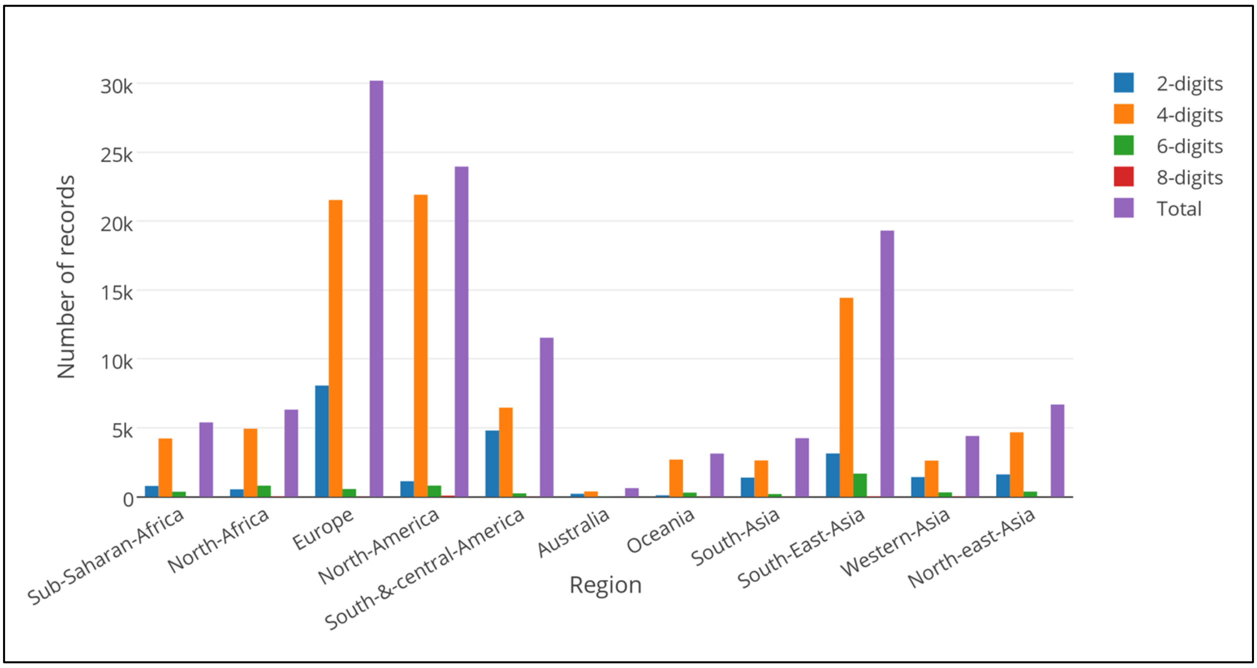

Figure A.1 Distribution of HLA data records per geographical region in the Allele Frequency Net Database

Appendix 2

Table A.1: Data sets downloaded from the AFND database

Country	Population	Sample size
Burkina_Faso	Burkina_Faso_Fulani	49
Burkina_Faso	Burkina_Faso_Mossi	53
Burkina_Faso	Burkina_Faso_Rimaibe	47
Cameroon	Cameroon_Baka_Pygmy	10
Cameroon	Cameroon_Bamileke	77
Cameroon	Cameroon_Beti	174
Cameroon	Cameroon_Sawa	13
CAR	CAR_Mbenzele_Pygmy	36
Ghana	Ghana_Ga-Adangbe	131
Kenya	Kenya_mixed	144
Kenya	Kenya_Luo	265
Kenya	Kenya_Nandi	240
Senegal	Senegal_Niokholo_Mandenka	165
South_Africa	SA_Black	200
South_Africa	SA_Caucasians	102
South_Africa	SA_Natal_Tamil	51
Uganda	Kampala_mixed	161
Uganda	Kampala_pop_2	175
Mali	Mali_Bandiagara	138
Morocco	Morocco_Coast_Chaouya	98
Morocco	Morocco_Metalsa_pop_2	73
Morocco	Morocco_Settat_Chaouya	98
Sudan	Sudan_mixed	200
Tunisia	Tunisia_mixed	100
Finland	Finland_mixed	91
Germany	Germany_Turkey_min	4856
Ireland	Ireland_North_mixed	1000
Saudi_Arabia	Saudi_Guraiat_Hail	213
Georgia	Georgia_Tibilisi	109
Iran	Iran_Baloch	100
Saudi Arabia	Guraiat and Hail	200