Development of an Analysis Pipeline for HLA Genotyping using Illumina Short Reads.

James Bird
474496

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, in fulfillment of the requirements for the degree of Master of Science.

Johannesburg
March, 2019

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Declaration

I, James Bird (474496), am a student registered for the degree of Master of Science in the academic year 2019. I hereby declare the following:

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

Signature  28th day May 2019

Abstract

Human leukocyte antigens are highly polymorphic loci located on chromosome six. This region is the most polymorphic region within the human genome, and as such, genotyping alleles in this region is problematic. Furthermore, the required resolution of genotyping is dependent on the application. For instance, organ transplants require two-digit resolution for kidney, and a minimum of four-digit resolution for bone marrow, while population disease related studies often require six-digit resolution. As specialized *HLA* genotyping tools have been developed which utilize NGS data, the aim of this study was to compare four *HLA* genotyping tools, namely - BWAkit, xHLA, Kourami and HISAT-Genotype, and to evaluate whether population-specific *HLA* variability would affect their accuracy. The accuracy of the tools were compared to Sanger sequenced *HLA* data, where exons 2 and 3 were sequenced for *HLA* class I. As exons 2 and 3 were available as a reference from the Sanger sequencing, an accurate allele call was determined on its similarity to the reference data. It was found that at the two- and four-digit resolution, xHLA was the most accurate, which was due to the inclusion of a nucleotide-to-protein alignment step in the algorithm. Kourami was the most accurate at the six-digit resolution due to the use of alternate loci, in the alignment step. To further identify possible error trends, the allele sequences produced by the tools were analyzed. It was found that the majority of errors occurred at heterozygous positions, where false homozygous positions were identified. It was also noted that, with the exception of HISAT-Genotype, each tool was most accurate at *HLA-B*, and least accurate at *HLA-C*. From evaluating *HLA* population-specific variability, it was found that the four super-populations tested - African, Asian, European and South American, did not significantly vary, in regards to *HLA* variability. It was, however, found that the different loci differed significantly from each other. Therefore, in conclusion, future improvements include varying the parameters when genotyping different loci. Currently, however, a consensus approach using xHLA and Kourami should be utilized.

Acknowledgments

I would like to first and foremost thank my supervisor, Dr Nikki Gentle, for her endless patience and support throughout this research.

Secondly, I would like to offer my gratitude to the National Research Foundation for funding during this project.

Last, but not least, I would like to thank my family and friends.

Structure and Outputs of Dissertation

Posters

1. *University of the Witwatersrand Post-graduate Symposium* (2017) Johannesburg, South Africa. Analysis of the Accuracy of NGS HLA data from the 1000 Genomes Project. Bird, J & Gentle N

2. *Molecular and Biosciences Research Thrust (MBRT)* (2017) Johannesburg, South Africa. Analysis of the Accuracy of NGS HLA data from the 1000 Genomes Project. Bird, J & Gentle N

Presentations

1. *South African Society for Bioinformatics (SASBi) and South African Genetics Society (SAGS) Conference* (2018) Golden Gate Nature Reserve, Free State. Evaluation of three HLA genotyping approaches: limitations and successes. Bird, J & Gentle N

2. *Molecular and Biosciences Research Thrust (MBRT)* (2018) Johannesburg, South Africa. Evaluation of three HLA genotyping approaches: limitations and successes. Bird, J & Gentle N

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	Analysis of variance
APC	Antigen Presenting Cell
BAM	Binary Alignment Map
β2M	*beta*-2-Microglobulin
CD	Cluster of Differentiation
cDNA	coding Deoxyribonucleic Acid
DNA	Deoxyribonucleic Acid
ER	Endoplasmic Reticulum
GATK	Genome Analysis Toolkit
GRC	Genome Reference Consortium
HGP	Human Genome Project
HLA	Human Leukocyte Antigen
HSD	Honest Significant Difference
IMGT	International Immunogenetics Information System
Indel	Insertion/deletion polymorphism
IPD	Immuno Polymorphism Database
KIR	Killer immunoglobulin-like receptor
li	Invariant Chain
LD	Linkage Disequilibrium
MHC	Major Histocompatibility Complex
MSA	Multiple Sequence Alignment
NGS	Next-generation Sequencing
NK	Natural Killer
PCA	Principal Component Analysis
PDBS	Pathogen-driven balancing selection
PLC	Peptide-loading Complex
RAM	Random Access Memory
SAM	Sequence Alignment Map
SBT	Sequence Based Typing
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
SSOP	Sequence-Specific Oligonucleotide Probe
TCR	T-cell receptor
TAP	Transporter Associated with Antigen Processing
SSP	Sequence-Specific Primer
WGS	Whole Genome Sequence
VCF	Variant Call Format

Chapter 1
Introduction

1.1 Human Leukocyte Antigen Gene

The Human Leukocyte Antigens (HLA) are encoded by a family of genes found within the *major histocompatibility complex* (*MHC*) on the short arm of chromosome six (Francke and Pellegrino, 1977). This is a highly polymorphic region, in which, to date, over 220 genes and 21 000 alleles have been identified (Robinson *et al.*, 2015). There are three classes of *HLA* genes, namely class I, class II and class III. The three classes are classified based on the structure and function of the encoded glycoproteins. HLA class I receptors are expressed on all nucleated cells and predominantly function to present intracellular peptide antigens to CD8 receptors on T-cells (Kristensen and Mossin, 1982). HLA class II receptors are expressed predominantly on the surface of antigen presenting cells (APCs), which includes B cells and dendritic cells, and function to present extracellular peptides to cells expressing CD4 receptors (Shiina *et al.*, 2004). HLA class III genes encode numerous pro-inflammatory cytokines, as well as molecules involved in lipid antigen presentation (Shiina *et al.*, 2004). *HLA* class I can be further divided into two groups, the classical and non-classical. The classical loci consists of *HLA-A*, *HLA-B* and *HLA-C*. The second group, the non-classical loci, include at least 15 genes, including three protein-coding and 12 pseudogenes (O'Callaghan and Bell, 1998). The non-classical loci are located near the classical *HLA* class I loci, and share a high degree of sequence similarity, but significantly reduced variability (O'Callaghan and Bell, 1998; Shiina *et al.*, 2004).

HLA class I molecules consist of a heavy (alpha) and a light (beta) chain, encoded by the *HLA* class I genes. These consists of seven or eight exons. Exon 1 encodes

the signal peptide and exons 2, 3 and 4 encode the alpha domains. Exons 4 to 8 encode the transmembrane domain and the cytoplasmic tail (Figure 1.1). The beta chain consists of β-2-microglobulin (β2M) that is encoded by the gene *B2M*, located on chromosome 15. This binds to the alpha chain in order to stabilize the peptide-binding domain of the HLA molecule in preparation for peptide loading (Figure1.1; Germain and Margulies 1993).

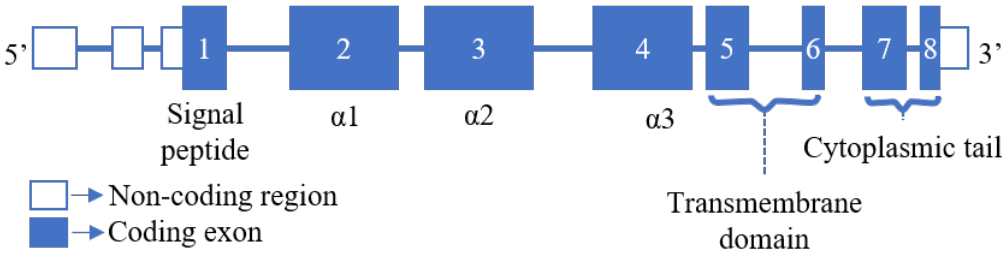


Figure 1.1: Graphical representation of the gene structure of the classical *HLA* class I genes. Coding exons are shown in blue. Non-coding regions are shown in white.

1.2 HLA Class I Protein

The signal peptide, which is encoded by exon 1, allows for the insertion of the HLA molecule into the endoplasmic reticulum (ER). Once the molecule enters the ER, the signal peptide is cleaved. The transmembrane domain is then inserted into the ER membrane, which anchors the HLA molecule (Saper *et al.*, 1991). Thereafter, HLA assembly occurs, and is initiated with the folding of the heavy chain. The heavy chain consists of the (now cleaved) signal peptide, the three alpha domains, the transmembrane domain and the cytoplasmic tail. The folding of the heavy chain is assisted by the chaperone molecule - calnexin, and an associated enzyme (ERp57). Calnexin functions by stabilizing the heavy chain, which prevents aggregation. ERp57 functions to catalyze the formation of disulfide bonds within the heavy chain. Thereafter, β2M binds to the heavy chain. This results in a conformational change in the molecule, which creates a peptide-binding groove. This forms the peptide-loading complex (PLC). Within the PLC, calnexin

dissociates and is replaced by calreticulin, a chaperone molecule that also binds ERp57 (Sadasivan *et al.*, 1996). Tapasin is the final molecule that binds to the heavy chain, ERp57 and calreticulin, to stabilize and prime the complex. Tapasin, furthermore, forms a binding site for Transporter associated with antigen processing (TAP), which allows for peptides to enter the ER lumen (Momburg and Tan, 2002).

1.3 *HLA* Function

HLA class I molecules present endogenous peptides derived from host, viruses that have infected the cell, and other intracellular pathogens. These peptides are typically eight to nine amino acids long (Rammensee *et al.*, 1993). Once the HLA class I molecules present the peptides extracellularly, they are able to form a complex with $CD8^+$ T-cells, via the T-cell receptor complex (TCR). During viral infection, the virus utilizes the host's cellular machinery to replicate. Through this process, viral proteins are synthesized. Within the host's cytosol, proteosomes are present, which process free floating proteins into smaller peptides. The peptides then move towards the ER. On the surface of the ER is TAP, which binds these peptides, and transports them into the ER lumen, where they can be bound by the PLC. Once a peptide is bound, the PLC dissociates. Calnexin, tapasin and ERp57 are retained in the ER. The antigen-HLA complex is transported through the cell secretory pathway to the cell surface. This allows the complex to present the peptides extracellularly to $CD8^+$ T-cells. It has been suggested that tapasin may play a role in selectively promoting the binding of high-affinity peptides into the peptide-binding domain by widening the peptide-binding groove (Wearsch and Cresswell, 2007). This decreases the binding affinity of peptides, which results in the disassociation of peptides with low binding affinity. High affinity peptides can then induce a conformational change, which results in further increased binding affinity.

As HLA can present peptides derived from either the host (”self”), or an external agent (”non-self”), this allows for the immune system to differentiate between healthy cells and infected cells. In the case of non-self peptide recognition, the

cells are targeted for apoptosis (Bjorkman and Parham, 1990). The $CD8^+$ T-cell-HLA-peptide complex is formed by the binding of the HLA-peptide complex to the TCR. In the case of molecules containing ”self” peptides, no immune response is initiated. In addition to HLA class I binding CD8 receptors, these molecules can also bind killer immunoglobulin-like receptors (KIRs), which are predominantly expressed on natural killer (NK) cells. This binding has been found to be specific to certain HLA motifs, with KIRs only recognizing *HLA* class I alleles presenting with Bw4, Bw6, C1 or C2 motifs (Ruggeri *et al.*, 1999; Gumperz *et al.*, 1997).

Furthermore, it has been found that HLA-C is the principle regulator of NK cell responses, due to the ability of HLA-C to bind to KIRs (Mandelboim *et al.*, 1997). This is vital, as the binding of HLA to KIR either results in inhibition or activation of an immune response, which is dependent on the specific HLA or KIR alleles (Mandelboim *et al.*, 1997). NK cell inhibitory receptors are found to bind to HLA-C proteins, which, in the absence of binding, results in the activation of the NK cell, which can occur through a non-self cell or in cases where the expression of *HLA-C* is inhibited (Figure 1.2; Ljunggren and Karre 1990). HLA class I also acts as an inhibitory natural killer (NK) cell ligand, which allows for targeted NK cell activation (Moretta *et al.*, 1996). In certain situations, such as cancer and certain pathogenic infections, *HLA* class I is down regulated (Vitale *et al.*, 1998). This prevents the NK cell inhibition by HLA-C and is termed ”missing self” (Figure 1.2). Once this occurs, the NK cell activation initiates lysis of the ”missing self” cell. An example of this is in HIV infections, where the viral protein Nef (Negative Regulatory Factor) downregulates HLA, in particular HLA-A and HLA-B. In laboratory HIV strains, the downregulation of HLA-C by Nef results in NK activation, however, in wild types, the HLA-C downregulation is mediated by Vpu (Viral protein unique), which simultaneously downregulates HLA-C, and prevents NK cell activation (Apps *et al.*, 2016).

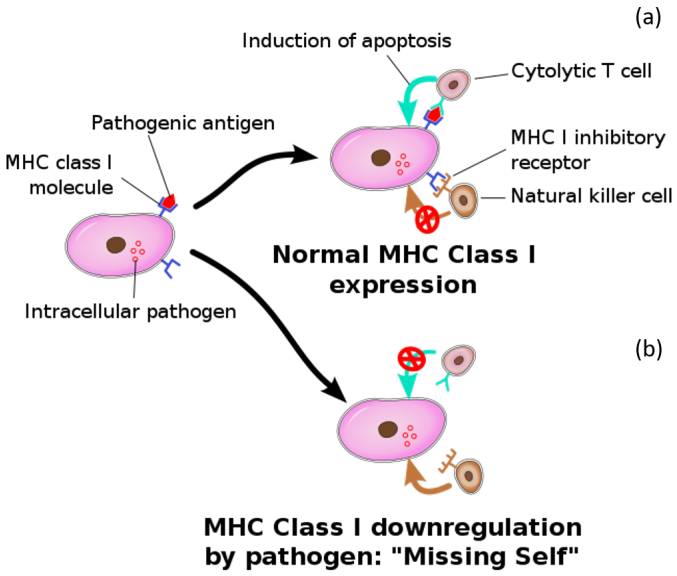


Figure 1.2: Diagram of HLA immune activation through (a) T-cell mediated recognition of pathogenic infection and (b) NK-mediated recognition of ”missing self”. Image obtained from https://en.wikipedia.org/wiki/Natural_killer_cell

1.4 HLA Supertypes

HLA molecules that bind to similar epitopes can be grouped into HLA supertypes based on the structure of their peptide-binding domains. Originally, only nine supertypes were identified (Sette and Sidney, 1998). However, these definitions have since been extended and now include seven HLA-A supertypes (A01, A01 and A03, A01 and A24, A02, A03, A24, and unclassified) and seven HLA-B supertypes (B07, B08, B27, B44, B58, B62 and unclassified; Sidney *et al.* 2008).

Within the peptide-binding domain, there are 57 accessible amino acid residues that are capable of binding peptides (1.3; Bjorkman *et al.* 1990). These residues largely dictate the peptide binding repertoire of the HLA molecule. The side-chains of the residues that interact with the antigen face towards the peptide-binding site, which is towards the interior of the molecule (Figure 1.3). Residues that interact with the TCR are located around the perimeter of the molecule (Figure 1.3). Because of this interaction, with both the peptide and the TCR (or KIR), the specific amino acid sequence of the HLA molecule affects which peptides are bound, and therefore, the

ability of the HLA molecule to initiate an immune response (van Deutekom and Kesmir, 2015).

The accessible amino acid residues can further be divided into six pockets, A to F (Figure 1.4; Saper *et al.* 1991. These pockets are separated based on the binding properties of the residues that make up the pocket. Pocket A binds to the N-terminus of the peptide and pocket F binds to the C-terminus (Figure 1.4). The remaining pockets bind the remainder of the antigen. Due to the selective binding of the pockets, pocket B and F are theorized to have a more significant role in antigen binding and presentation (Carreno *et al.*, 1993; Matsui *et al.*, 1993).

HLA supertypes are an important factor to consider in the context of disease prevention, as different supertypes have different levels of affinity for peptides, and therefore different tolerances to variation among peptides (Hendel *et al.*, 1999). By studying common epitopes that the supertypes bind, one can develop vaccines containing supermotifs (MacDonald *et al.*, 2000) that would provide resistance in a greater number of individuals.

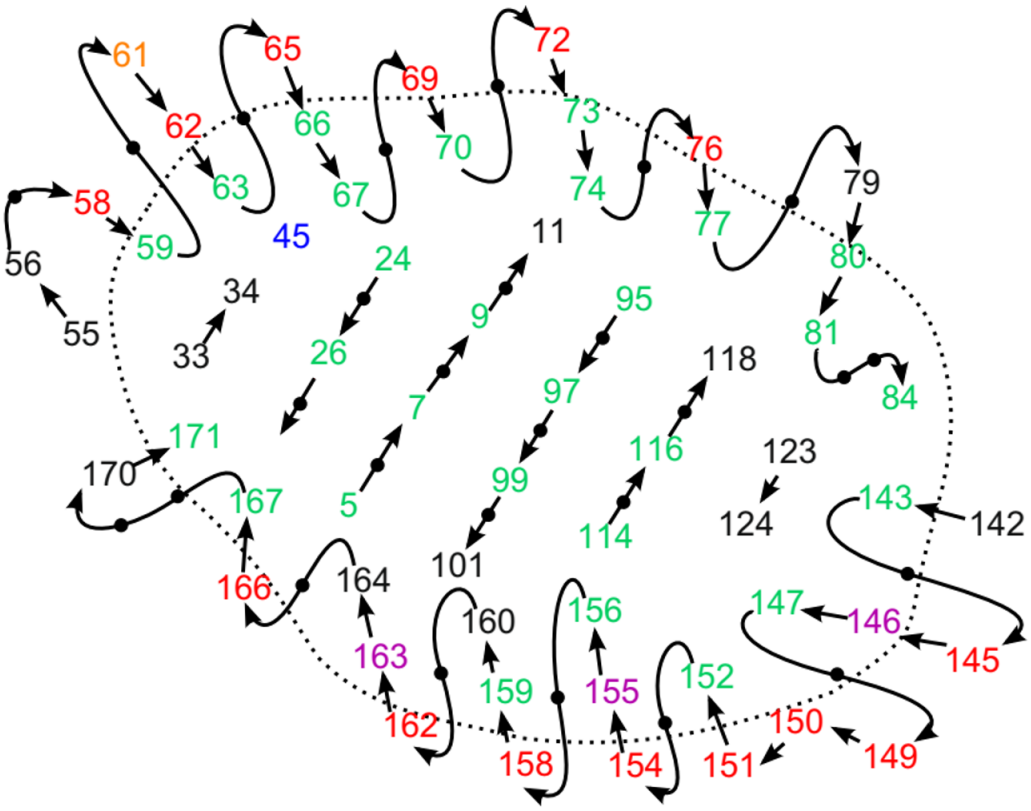


Figure 1.3: Schematic overview of the HLA antigen-recognition site. Numbers represent the positions of the amino acid residues. Numbers in green represent residues that interact with the antigen, those in red interact with the TCR and those in purple represent residues that interact with both the antigen-recognition site and TCR. Figure obtained from van Deutekom and Kesmir (2015)

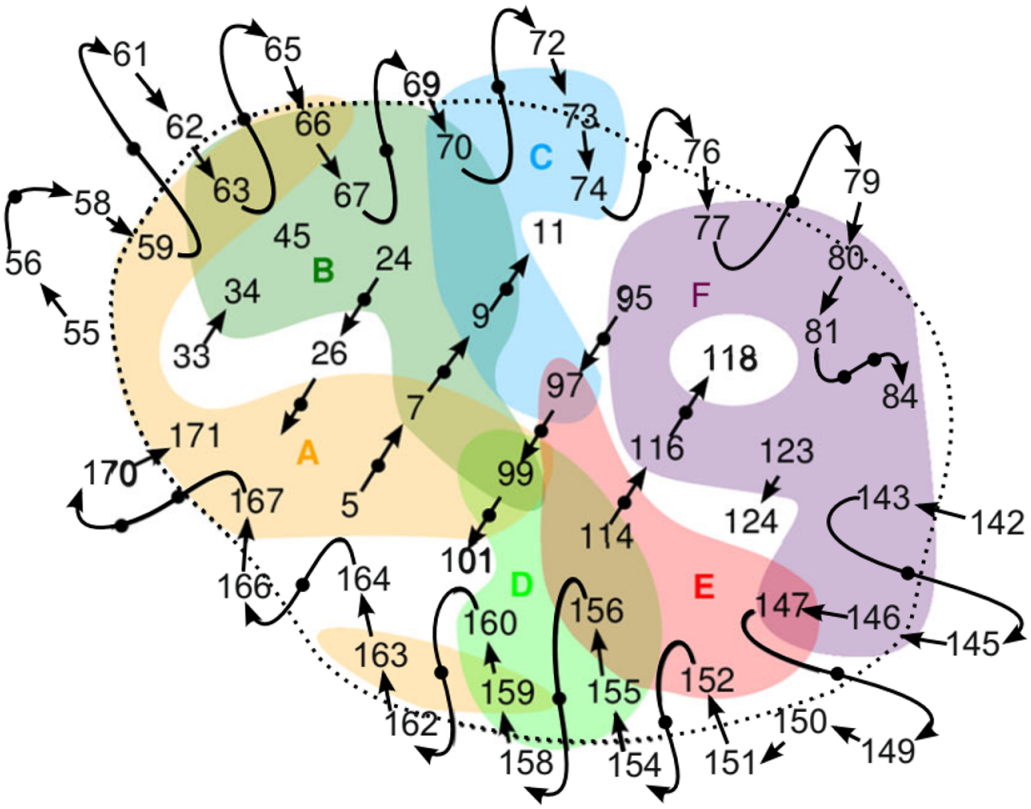


Figure 1.4: Locations of binding site pockets within the antigen-recognition site of *HLA* class I proteins. The residues encompassing binding pockets (A-F) are colored accordingly. Figure obtained from (van Deutekom and Kesmir, 2015)

1.5 *HLA* Nomenclature

Due to the high degree of variation within the *HLA* region, these genes have their own nomenclature system. This nomenclature relies upon a digit-based differentiation system (Table 1.1). For each of the three genes, allele groups are differentiated at the two-digit resolution, which separate the allele allotypes, or the peptide-binding repertoires. Non-synonymous mutations in the coding sequence are differentiated at the four-digit resolution, and as such, this resolution differentiates the amino acid sequence of the molecule. The six-digit resolution represents synonymous mutations within the protein coding sequence. The final two digits represent differences in non-coding sequences, such as the introns. *HLA* alleles may also include a suffix which denotes changes in gene expression. N – Null allele, L – Low expression, S – Secreted and Q – questionable (Table 1.1; Marsh *et al.* 2010, Robinson *et al.* 2013). With the development of new, sequence-based *HLA* genotyping methods, there has been a substantial increase in the amount of known *HLA* alleles. This has increased from under 1 000 known alleles in 1998, to over 21 000 alleles as of January 2019 - of which, ±15 000 alleles occur within *HLA* class I (Table 1.2).

Table 1.1: Explanation of *HLA* nomenclature system (Marsh *et al.*, 2010)

Identifier	Resolution	Explanation
HLA-A	Gene	Differentiates the different *HLA* genes - *HLA-A, -B, -C,* etc.
*HLA-A*01*	Two-digit	Differentiates *HLA* alleles based on the peptide binding repertoire of the encoded protein.
*HLA-A*01:01*	Four-digit	Differentiates *HLA* alleles based on the amino acid sequence encoded by the allele.
*HLA-A*01:01:01*	Six-digit	Differentiates *HLA* alleles based on the nucleotide sequence of the encoding gene region.
*HLA-A*01:01:01G*	Ambiguous	Groups alleles with identical nucleotide sequences across exon 2 and exon 3
*HLA-A*01:01:01:01*	Eight-digit	Differentiates *HLA* alleles based on the exonic and intronic nucleotide sequence.
*HLA-A*01:01:01:01N* *HLA-A*01:01:01:01L* *HLA-A*01:01:01:01S* *HLA-A*01:01:01:01Q*	Expression of encoded protein	N - Null allele L - Low S - Secreted Q - Questionable

Table 1.2: Total numbers of *HLA* Alleles reported by the IMGT/HLA database, as of January 2019

Gene	*HLA-A*	*HLA-B*	*HLA-C*	Total
Alleles	4846	5881	4654	15,381
Proteins	3286	4088	3070	10,444

Counts obtained from the IMGT/HLA database (v. 3.35; Robinson *et al.*2015)

1.6 *HLA* Evolution

As *HLA* is the most variable region within the human genome, much work has been done to try to explain the source of this variation. Currently, the leading hypothesis is pathogen-driven balancing selection (PDBS; Apanius *et al.* 1997). Balancing selection is the hypothesis that evolutionary pressures on a population level will favor heterozygosity within a given gene. Through the function of HLA in immune regulation, it is thought that a high degree of heterozygosity on a population level, allows for a broader range of antigens to be recognized, based on recognition of a broader peptide-binding repertoire, population-wide. The evidence for balancing selection within *HLA* is well supported, as most species that contain the *MHC*, possess a large number of alleles (Aguilar *et al.*, 2004; Bernatchez and Landry, 2003; Hughes and Nei, 1988).

Furthermore, the rate of non-synonymous substitutions is higher than synonymous substitutions within the *HLA* region (Hughes and Nei, 1988), which contrasts with most other regions within the human genome (The 1000 Genomes Consortium, 2015). The majority of variability within *HLA* class I occurs within exons 2 and 3 (which encode the peptide-binding domain). This variation suggests that the interaction between the antigen-recognition site and antigens, and by default, pathogens is the evolutionary force behind the balancing selection. Therefore, through this pathogen-driven pressure, it has been hypothesized that geographical regions with high pathogen diversity should have higher *HLA* diversity. Evidence for this was found by Prugnolle *et al.* (2005), who analyzed the *HLA* genetic diversity in 61 human populations, and compared this to the total number of intracellular human disease agents known to occur in each region that the populations originated from. The results suggested that PDBS is the reason for the high diversity observed within *HLA*. This has further been supported by dos Santos Francisco *et al.* (2015), who analyzed *HLA* supertype variation between populations, and observed a similar pattern.

1.7 HLA Population Structure

The *HLA* region is one of the most intensely studied regions within the human genome. Analysis of heterozygosity has shown that *HLA* variability differs greatly between populations separated by continents, whereas populations found within the same continent tend to be more similar to each other (Meyer *et al.*, 2018). This is due to both admixture within the same continent, and the fact that these populations are exposed to similar pathogens (Prugnolle *et al.*, 2005).

Furthermore, genetic variation within African populations is higher than in other populations (Tishkoff and Williams, 2002). African populations also display lower levels of linkage disequilibrium (LD; Reich *et al.* 2001) when compared to other populations. In contrast, previous research has found that African, Asian and European populations displayed similar amounts of high variation amongst *HLA*, with certain geographically isolated populations such as Taiwanese aboriginals and Oceanian populations displayed lower levels of variation (Buhler and Sanchez-Mazas, 2011). The East Africa origin theory, humans originated in Africa, migrated to Asia, Europe, and finally the Americas. It is, therefore, expected to see a *HLA* population variability in which *HLA* alleles are found to be unique to certain populations, with certain common alleles, which were maintained in all populations. This appears to have occurred, as certain two-digit alleles occur in similar frequencies, across numerous super-populations. Furthermore, these common alleles appear in similar frequencies, such as *HLA-A*02*, which is commonly found in African, Asian and European populations (Holmans, 2001). As certain HLA alleles are only located within certain populations (Gourraud *et al.*, 2014), the linkage between *HLA* allele frequencies and population location is important for determining global *HLA* population patterns. Due to these geographically-linked allele frequencies, many anthropological studies have used the HLA region, and as such require large cohorts, such as the 1000 Genomes Project.

As HLA molecules are involved in the recognition of potential pathogens, which can be highly variable, the genes encoding the molecules have also been found to

be highly variable. It has been found that the LD within the *MHC* region is very different from other regions of the genome (Miretti *et al.*, 2005). LD is the ”nonrandom association of alleles that arises when alleles occur together more often than is expected through independently segregating alleles” (Lewontin, 1964). LD can be influenced by numerous factors, such as mutations, non-random mating genetic drift and population structure. *HLA* has been affected by these factors, as well as pathogen-driven balancing selection (PDBS), which has been found to explain the degree of polymorphisms within the *HLA* region (Prugnolle *et al.*, 2005). One of the non-classical *HLA* loci - *HLA-G*, has been found to contain alleles in LD with alleles within *HLA-A* (Ober *et al.*, 1997). This is thought to originate from pressures due to maternal-fetal interface, which, according to Ober *et al.* (1997) suggests co-functioning of *HLA-A* and *HLA-G* during gestation.

1.8 *HLA* Class I Associations

1.8.1 Autoimmune

Discrimination between self and non-self is an important part of immunity. When this immune response is disregulated, this can result in autoimmune diseases. Due to the role of HLA in this process, numerous associations between then genes and autoimmune disorders have been identified (Table 1.3). One of the most well-studied autoimmune-HLA associations is between HLA-B and Ankylosing Spondylitis (AS). AS is an arthritic disease that most commonly presents as long term inflammation of the spine. At present, the cause of AS is not known, however the disease has been found to be strongly associated with specific *HLA-B* alleles (Table 1.3), with *HLA-B*27* found to be present in the majority of individuals with AS. A potential explanation for this may be that the specific amino acid sequence of HLA-B*27 can affect calnexin and calreticulin binding, which can result in a mis-folded molecule (Colbert, 2000). This could affect recognition of ”self”, thereby, initiating an immune response. In contrast, this association is not seen for HLA-B*14:02, which differs by two amino acids, and has a faster folding rate than HLA-B*27:05 (Merino *et al.*, 2008). The association between HLA-B*27 and AS is not limited to a selected population, however, the predominant *HLA-B*27* allele

(at the four-digit resolution) differs across populations. Further associations between HLA and autoimmune diseases including type I diabetes, multiple sclerosis and Graves' Disease (Table 1.3).

Table 1.3: Associations between autoimmune diseases and specific *HLA* class I alleles

Disease	*HLA* Allele	Association	Population	Reference
Ankylosing Spondylitis	*B*27:05*	Risk	African	Hill *et al.* (1991b)
	*B*27:02*	Risk	European	D'amato *et al.* (1995)
	*B*27:05*	Risk	South American	Sampaio-Barros *et al.* (2001) Martínez *et al.* (1999)
	*B*27:04*	Risk	Asian	Shih *et al.* (2001)
Diabetes Type I	*B*39*,	Risk	European	Nejentsev *et al.* (2007)
	*B*18*	Risk	European	
	*A*24*,	Risk	European	
	*A*01*,	Protective	European	
	*A*11*,	Protective	European	
	*A*31*	Protective	European	
	*B*42*	Risk	African	(Omar *et al.*, 1984)
Multiple Sclerosis	C*05	Risk	European	(Yeo *et al.*, 2007)
	*C*15*	Risk	European	(Fogdell-Hahn *et al.*, 2000)
	*C*01*	Protective	European	
	*A*02:01*	Risk	European	
Graves' Disease	*C*07*	Risk	European	(Simmonds *et al.*, 2007)
	*C*03*	Risk	European	
	*C*16*	Protective	European	
	*B*08*	Protective	European	
	*B*44*	Protective	European	
	*B*46*	Risk	Asian	(Li *et al.*, 2013)

1.8.2 Infectious Diseases

In the context of infectious diseases, the association between specific *HLA* class I alleles and HIV-1 has been well documented. The association between *HLA-B*57* and HIV-1 disease progression is particularly well described. In Caucasians *HLA-B*57:01*, and in Africans *HLA-B*57:03*, have been found to be associated with delayed disease progression (Table 1.4). These two alleles are very similar, differing by only two nucleotides, which are located within exons 2 and 3. In contrast, *HLA-B*58:02* is associated with rapid disease progression, whereas

*HLA-B*58:01*, which again differs by only two nucleotides, is associated with delayed disease progression (Klepiela *et al.*, 2004).

Further studies have found an association with another retrovirus, Hepatitis B virus, with numerous alleles associated with differing degrees of disease severity (Table 1.4). Furthermore, these associations have largely been found to be limited to certain populations. This is observed with *HLA-A*02*, which is associated with chronic hepatitis B infection in the Asian population, whereas it is reported to be associated with viral clearance in a European population (Popov *et al.*, 2005). Other alleles associated with differential disease progression include *HLA-B*53* in African populations, which is associated with protection from severe malaria infections (Hill *et al.*, 1991a). This allele is commonly found in West African populations (González-Galarza *et al.*, 2015), and is rare in other populations.

Table 1.4: Known associations between specific*HLA* class I alleles and infectious disease

Disease	*HLA* Allele	Association	Population	Reference
HIV	*B*57:01*	Protective	European	Goulder *et al.* (1996)
	*B*57:03*	Protective	African	Goulder *et al.* (1996)
	*B*13:02*	Protective	African	Fellay *et al.* (2009)
	*B*35:01*	Risk	European	The International HIV Controllers (2011)
	*B*58:02*	Risk	African	Klepiela *et al.* (2004)
	*B*58:01*	Protective	European	Klepiela *et al.* (2004)
Hepatitis B	*A*02:06*	Risk	Asian	Wu *et al.* (2004)
	*B*35*	Risk	Asian	Wu *et al.* (2004)
	*A*02*	Risk	Asian	Zeniya *et al.* (1993)
	*B*08*	Risk	European	Thio *et al.* (2003)
	*A*26*	Protective	Asian	Zeniya *et al.* (1993)
	*A*03:01*	Protective	European	Thio *et al.* (2003)
Malaria	*B*53*	Protective	African	Hill *et al.* (1991a)

1.8.3 Pharmacogenetics

Pharmacogenetics is the study of the association between genetic variants and drug responses. Numerous drugs have been found to cause adverse reactions in individuals, and association studies have found numerous links between specific *HLA* alleles and these adverse reactions. One example of this is Abacavir, an

antiretroviral treatment drug, which has been found to cause a drug hypersensitive response in African individuals with *B*57:01* (Table 1.5). In individuals with *HLA-B*57:03*, an allele that differs by two nucleotides, there are no reported adverse reactions (Lucas *et al.*, 2015). Nevirapine is another antiretroviral drug used in the treatment of HIV infections, and can be used in conjunction with Abacavir (Jose Casas *et al.*, 2015). It has been found to cause hepatotoxicity, or liver damage, in African individuals with *HLA-B*58:01* (Table 1.5). Both *HLA-B*58:01* and *HLA-B*57:01* are commonly found in sub-saharan African populations (González-Galarza *et al.*, 2015), and therefore, *HLA* genotyping is important to prevent adverse drug reactions.

Another example of an adverse *HLA*-associated drug reaction, is that caused by Allopurinol, a drug used to treat gout. In Asian and European populations, it has been found to be associated with adverse reactions in the skin and mucous membranes. This association has been found to be with *HLA-B*58:01* (Table 1.5). In other populations, such as Africans and South Americans, this adverse reaction is found to be extremely rare, despite these alleles being common in these populations (Jung *et al.*, 2018), indicating that this response may possibly be population dependent.

Lastly, Levamisole was an anti-parasitic medication used for the treatment of hookworms (Table 1.5). It is still currently used in cattle, but is no longer prescribed to humans in many countries due to causing agranulocytosis, which is the depletion of white blood cells. This was found to largely affect South American individuals with *HLA-B*27* (Phillips *et al.*, 2013), which is one of the alleles found to occur most frequently within the South American population, with five percent of individuals in certain areas possessing the allele (González-Galarza *et al.*, 2015).

While this is a short summary of adverse reactions to drugs associated with *HLA*, it demonstrates the importance of pharmacogenetics and understanding the underlying mechanisms. It further demonstrates the importance of studying pharmacogenetics in individual populations .

Table 1.5: Pharmacogenetics of *HLA*-associated drug induced symptoms

Disease	*HLA* Allele	Population	Reference
Abacavir	*B*57:01*	European	Mallal *et al.* (2002)
Allopurinol	*B*58:01*	Asian European	Chan and Tan (1989) Lonjou *et al.* (2008)
Levamisole	*B*27*	South American	Diez (1990)
Nevirapine	*B*58:01*	African	Phillips *et al.* (2013)

1.9 1000 Genomes Project

The 1000 Genomes Project aimed to create a reference dataset that captured the variation present in the human genome. To accomplish this, four phases were used. The first data set, obtained from the pilot phase, consists of low coverage, high coverage and exon-targeted sequencing (The 1000 Genomes Consortium, 2010). This was performed on 697 individuals from seven populations. This was performed as an initial part of a larger cohort study that aimed at identifying the vast majority of common variants within the human genome (The 1000 Genomes Consortium, 2012). After the pilot studies, phase 1 aimed at sequencing 2504 individuals. These individuals were selected as representative for five super-populations - African, East and South Asian, European and South American. The pilot phases and phase 1 data were generated through numerous sequencing tools, including Roche 454, SOLiD and Illumina (The 1000 Genomes Consortium, 2010, 2012).

Phase two expanded on the cohort, by sequencing a further 600 individuals, and was aimed at improving the methods utilized in phase one. Phase three, which was the final stage, consisted of sequence data for 2504 individuals (The 1000 Genomes Consortium, 2015). This data included additional African and South Asian samples, and utilized the improved methods generated from phase two (The 1000 Genomes Consortium, 2015). The sequence data obtained in phase 3 consists of Illumina low coverage and exon-targeted sequencing, and Illumina 250 bp paired-end whole genome data, with an average read depth of 30X. The high read depth data was generated through PCR-free sequencing for 30 individuals, and was performed to validate the low-coverage and exon-targeted sequencing. The read data was initially

aligned to GRCh37, utilizing the Genomes Analysis Toolkit (GATK) pipeline (Van der Auwera *et al.*, 2013), which included the use of BWA. For the low-coverage and exon-targeted sequencing, BWA-aln was utilized. For the high-coverage PCR-free data, BWA-mem was used to align the reads, in an alt-aware manner.

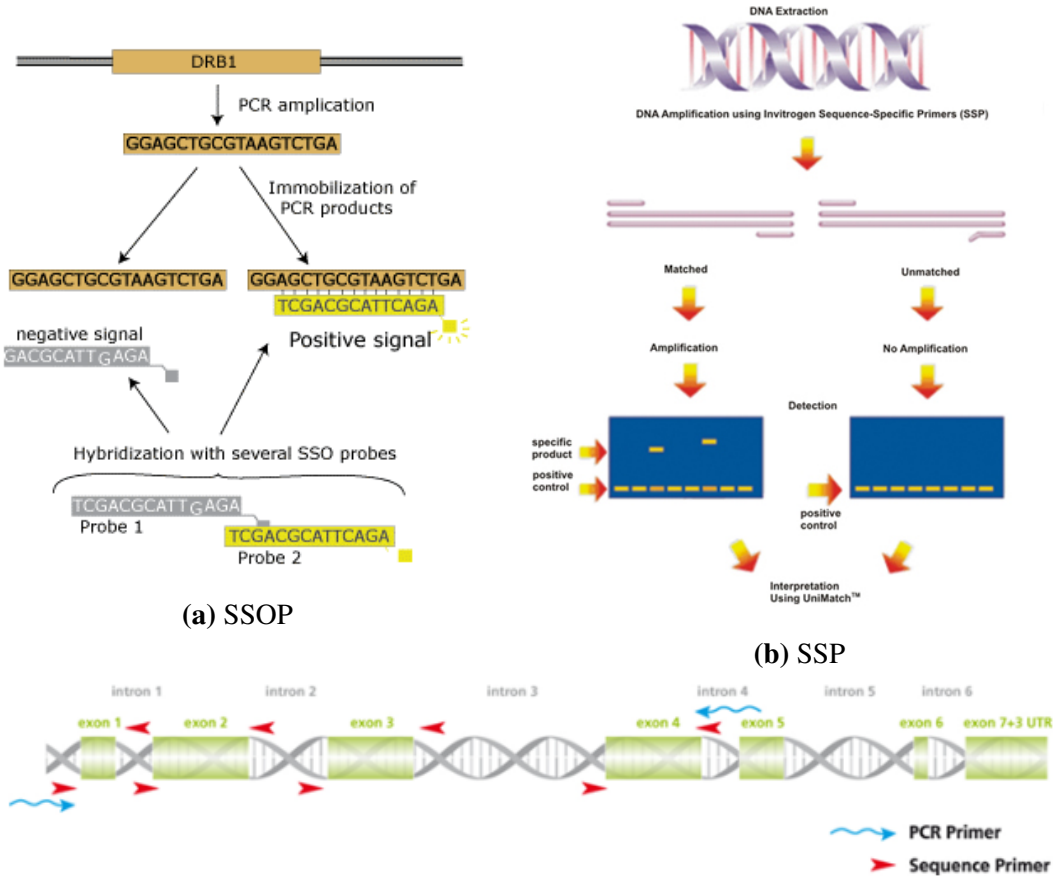
Throughout the project, the *HLA* region was found to be extremely variable, and as such, identifying variants was problematic. Once high-resolution *HLA* SBT genotypes were produced for 1267 individuals from the 1000 Genomes Project by Gourraud *et al.* (2014), further research was performed to identify the accuracy of identifying *HLA* variants from the NGS data for these individuals. It was found that 18.6 percent of the SNP genotype calls were incorrect (Brandt *et al.*, 2015), and that majority of the errors were due to a bias, in which the reference nucleotides were selected as opposed to the true variant present. In regards to individual loci, *HLA-B* was more susceptible to these errors than *HLA-A* or *HLA-C*, possibly due to the increased variation at this locus (Brandt *et al.*, 2015).

1.10 *HLA* Genotyping Methods

1.10.1 SBT, SSOP and SSP *HLA* Genotyping

Current laboratory-based *HLA* typing methods include sequence-specific oligonucleotide probes (SSOP), sequence based typing (SBT) and sequence-specific primers (SSP). SSP has been used to verify ambiguities identified through SBT and SSOP-typing (Erlich, 2012). Sequence-specific Oligonucleotide Primers (SSOP) is a method that utilizes probes to hybridize to specific sequences. This method requires the complimentary probe to be present for detection of an allele, and is therefore not suitable for novel allele detection (Figure 1.5a). Sequence-specific Primers (SSP) utilizes complimentary primers to bind to *HLA* allele sequences (Figure 1.5b). In the case of successful binding, amplification occurs that can be detected through electrophoresis on a gel. Similarly, to SSOP, this method is unable to genotype novel alleles. The gold standard for *HLA* typing is SBT, in which the alleles are typed through Sanger sequencing (Figure 1.5c; Santamaria *et al.* 1993) and the sequences are aligned to reference HLA alleles. It is common practice to only sequence exons 2 and 3 of

HLA class I alleles, which reduces the resolution of the genotyped allele to the ambiguous allele format. These ambiguities can arise from alleles differing from one another outside of the region that was sequenced, such as introns (for the eight-digit resolution), or exons encoding molecules outside of the peptide binding domain. Further ambiguities can arise from heterozygous allele combinations resulting in identical sequences to other heterozygous allele combinations (Figure 1.6). This results in two incorrect alleles from the incorrect phasing of exons 2 and 3.



(a) SSOP

(b) SSP

(c) SBT

Figure 1.5: Overview of three *HLA* genotyping methods.
(a) Sequence specific oligonucleotide probes (SSOP). Figure obtained from http://medweb4.unige.ch/immunologie/home/HSC/donor/HLA_typing/SSO.php.
(b) Sequence-specific Primers (SSP). Figure obtained from http://www.topdiag.com/top/0,27,ssp-hla-typing.html?sLang=en.
(c) Sequence-based Typing (SBT). Figure obtained from https://labproducts.caredx.com/products/sbt-resolver/hla-kits/hla-a/.

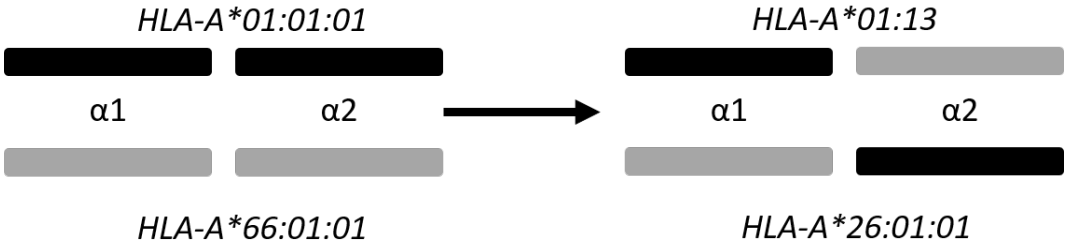


Figure 1.6: Schematic diagram of ambiguous typing combinations across exons 2 and 3

1.10.2 Next-Generation Sequencing *HLA* Genotyping

Next-generation sequencing (NGS) was the term introduced to differentiate more modern sequencing technologies from Sanger based sequencing techniques. NGS resulted in a massive increase in data generation due to the high-throughput of the techniques. The initial NGS technologies, which include Roche 454, Illumina and Ion Torrent, are termed second generation sequencing techniques (Table 1.6). This is due to the introduction of nanopore sequencing, which is third-generation sequencing.

Roche 454 introduced pyrosequencing. Pyrosequencing involves immobilizing DNA fragments onto beads. The Beads are emulsified, to allow for clonal amplification. Once the DNA is amplified, the beads are placed in a well. Pyrosequencing occurs through the introduction of one nucleotide, at a time, over the wells. When the nucleotide is bound, pyrophosphate is released. NGS involving Roche 454 involves pyrosequencing that begins with fragmenting DNA, followed by denaturation. The individual single stranded DNA (ssDNA) fragments are then bound to individual beads, whereby PCR is used to amplify the ssDNA fragments on each bead. The beads are then placed into fiber optic wells. Enzymes, which induced pyrosequencing, are then added to the wells that results in the release of inorganic pyrophosphate along with the release of photons during DNA synthesis. The release of photons is then measured when a nucleotide was added to the newly synthesized DNA chain (Margulies *et al.*, 2005). In comparison with alternate second-generation sequencing technologies, Roche 454 produced long read lengths.

While the reads produced by Roche 454 were beneficial for numerous applications, the chemistry behind the sequencing was costly, and error-prone. Furthermore, competing technologies, used in Ion Torrent, Life Technologies and Illumina were more cost effective. These technologies also had higher outputs, which resulted in increased read depths, and therefore, had a greater chance of error correction. As such, these technologies, and in particular, Illumina, have been preferentially used in numerous studies, with the 1000 Genomes Project using Illumina HiSeq 2500 in phase three to produce 250 bp paired-end reads (The 1000 Genomes Consortium, 2015). The third generation technologies, such as nanopore and single molecule real time sequencing, used by Oxford Nanopore and Pacific Biosciences, respectively, produce extremely long read lengths when compared to second generation technologies. Furthermore, the costs, in particular of nanopore sequencing, are extremely low. These technologies further do not require an amplification step, which reduces the stages at which sequencing errors can be produced.

Table 1.6: Comparison of NGS technologies. Adapted from Xuan *et al.* (2013); Carapito *et al.* (2016); Churko *et al.* (2013)

Company	Platform	Generation	Amplification	Sequencing	Read Length	Output	Cost/Mb
Ion Torrent	PGM 318 Ion Proton	2nd 2nd	Emulsion PCR	Ion semi-conductor	400 bp 200 bp	2 Gb 10 Gb	$0.38 $0.10
Roche 454	GS FLX Titanium XL+ GS Junior	2nd 2nd	Emulsion PCR	Pyrosequencing 40 Mb	1 kb 400 bp	700 Mb	$3.57 $20.00
Life Technologies	5500xl SOLiD	2nd	Emulsion PCR	Sequencing by ligation	75 bp	10 Gb	$0.15
Illumina	MiSeq HiSeq 2500	2nd 2nd	Bridge Amplification	Reversible dye termination	2x250 bp 2x100 bp	5.6 Gb 540 Gb	$0.18 $0.01
Pacific Bioscience	PacBio RS2	3rd	N/A	Single molecule real time sequencing	40 kb	260 Mb	$0.42
Oxford Nanopore	MinION	3rd	N/A	Nanopore sequencing	200 kb	30 Gb	$0.03

Gb - Gigabase
Mb - Megabase
kb - kilobase

1.10.3 Whole-Exome and Targeted Sequencing

Whole-Exome sequencing has previously been utilized for *HLA* genotyping (Szolek *et al.*, 2014; Liu *et al.*, 2013a). Within the human genome, the coding region of the genome forms the exons. This subset of the genome is approximately two percent of the whole genome, and therefore whole-exome sequencing is able to sequence the regions to a greater depth than whole-genome sequencing for a similar cost. The procedure for exome sequencing requires selective hybridization of DNA to oligonucleotide probes. Once the probes are bound, the remaining DNA consists of exons. The exons are amplified, through PCR, and then sequencing can occur, using a variety of second-generation sequencing platforms (Warr *et al.*, 2015).

Targeted sequencing is similar to WES, in that large amount of data can be generated, which results in increased read depth. Similarly to WES, targeted sequencing requires selective hybridization of DNA followed by sequencing through a variety of second-generation sequencing platforms (Harismendy *et al.*, 2009). Targeted sequencing is the most cost efficient method, however, it is subject to sequencing bias, in which one copy of an allele is disproportional sequenced. This effect has been shown to result in decreased accuracy (Harismendy *et al.*, 2009). Furthermore, when whole-exome and targeted sequencing approaches are used, regions outside of the targeted regions are excluded fro sequencing. This results in decreased ability to identify rare and novel alleles. Therefore, in large cohort studies, whole-genome sequencing is preferred, as the captured data is for a much wider region.

1.10.4 Imputation Methods

A further method, that is used in genome-wide association studies, is the use of microarrays. This technology uses probes, which are capable of binding certain nucleotide sequences, which produces a measurable response, such as fluorescence (Taub, Floyd *et al.*, 1983). Through this, SNP imputation occurs, which is the statistical inference of genotypes, based on the SNPs present, and a prior knowledge of haplotypes (Scheet and Stephens, 2006). Imputation has also been performed on data generated from NGS data, where the read data has been aligned,

and variants to the reference assembly are identified. Due to this, *HLA* imputation has also been introduced, however, due to the prior knowledge of LD, and the unique variation present in the *HLA* region, this has been found to be problematic (Karnes *et al.*, 2017). HIBAG (Zheng *et al.*, 2014), HLA*IMP (Dilthey *et al.*, 2011) and SNP2HLA (Jia *et al.*, 2013) are amongst the more widely used *HLA* imputation tools, and are capable of genotyping *HLA*, however, the accuracy beyond the two-digit resolution can be questionable, particularly in non-European populations (Karnes *et al.*, 2017).

1.10.5 Assembly Methods

An issue with *HLA* imputation is that the quality of the called SNVs can be called into question. A study comparing different SNV calling tools found a large discrepancy in the actual SNVs called (Liu *et al.*, 2013b). The researchers compared GATK - HaplotypeCaller (McKenna *et al.*, 2010), SAMtools mpileup (Li *et al.*, 2009a) and FreeBayes (Garrison and Marth, 2012). All three tools identified varying amounts of SNVs, and were prone to excluding rare SNVs (Liu *et al.*, 2013b). As *HLA* is so variable, many alleles exist in less than one percent of the population, and identifying alleles based on SNVs can be inaccurate. Researchers investigated LD patterns between *HLA* alleles, and found that while the more common alleles can be identified by tagSNPs, such as *HLA-C*07:02*, which occurs in numerous populations, other alleles such as *HLA-C*03:04* were not associated with any tagSNP (De Bakker *et al.*, 2006). Therefore tools which bypass the variant calling step and instead utilize aligned reads directly have been developed. This is beneficial for *HLA* genotyping, as data which may contain true variants is not discarded. Furthermore, many *HLA* alleles share certain variants, which can result in *HLA* ambiguity errors. As pairs of *HLA* alleles can have the same nucleotide sequence across exon 2 and exon 3, incorrect phasing of the exons can result in incorrect allele genotypes (Figure 1.6). Therefore utilizing the read data, correct phasing of the exons can occur. *HLA* assembly tools utilize a similar pipeline. Reads which map to the *HLA* region are aligned to coding *HLA* allele sequences, and thereafter reads spanning the intronic region are included, to phase the exons.

1.11 Reference Assemblies

The Human Genome Project (HGP) was initiated in 1990 with the aim of determining the nucleotide sequence of the human genome (The International Genome Sequencing Consortium and Human Genome Sequencing Consortium, 2004). The initial sequence was completed in 2003 and consisted of euchromatic regions of the human genome which accounts for approximately 92 percent of the human genome. Heterochromatic regions were excluded from the initial sequence due to the repetitive nature of the regions which resulted in the decreased sequencing accuracy (Schmutz *et al.*, 2004). Of the initial goals of the HGP, accuracy and completeness of the sequence was analyzed, and it was determined that 92 percent of the sequence had above a 99.99 percent accuracy. Furthermore, the initial sequence had 150 000 gaps which still required sequencing (The International Genome Sequencing Consortium and Human Genome Sequencing Consortium, 2004).

The human reference genome assembly was initially represented as a haploid linear consensus sequence, developed by the Human Genome Project (The International Genome Sequencing Consortium and Human Genome Sequencing Consortium, 2004). This linear haploid reference assembly is adequate for aligning most sequencing reads, however, in the presence of structural variation, or highly polymorphic regions, reads may not be accurately aligned. Furthermore, the initial sequence did not capture heterochromatic regions and possessed gaps in the sequence. The HGP released the 19th update to the reference sequence in 2009 - GRCh37. To better capture the variation which exists within the human population, GRCh37 included sequences which contained nine alternate sequences to the reference haploid sequence (Figure 1.7A). This allowed for regions with increased variation to better be demonstrated. These alternate sequences were termed alternate loci.

The 1000 Genomes Project initially utilized NCBI36, which was a reference sequence released in 2006. For phase 1 and phase 3, the more recent GRCh37 reference assembly was used. The specific assembly utilized by the 1000 Genomes

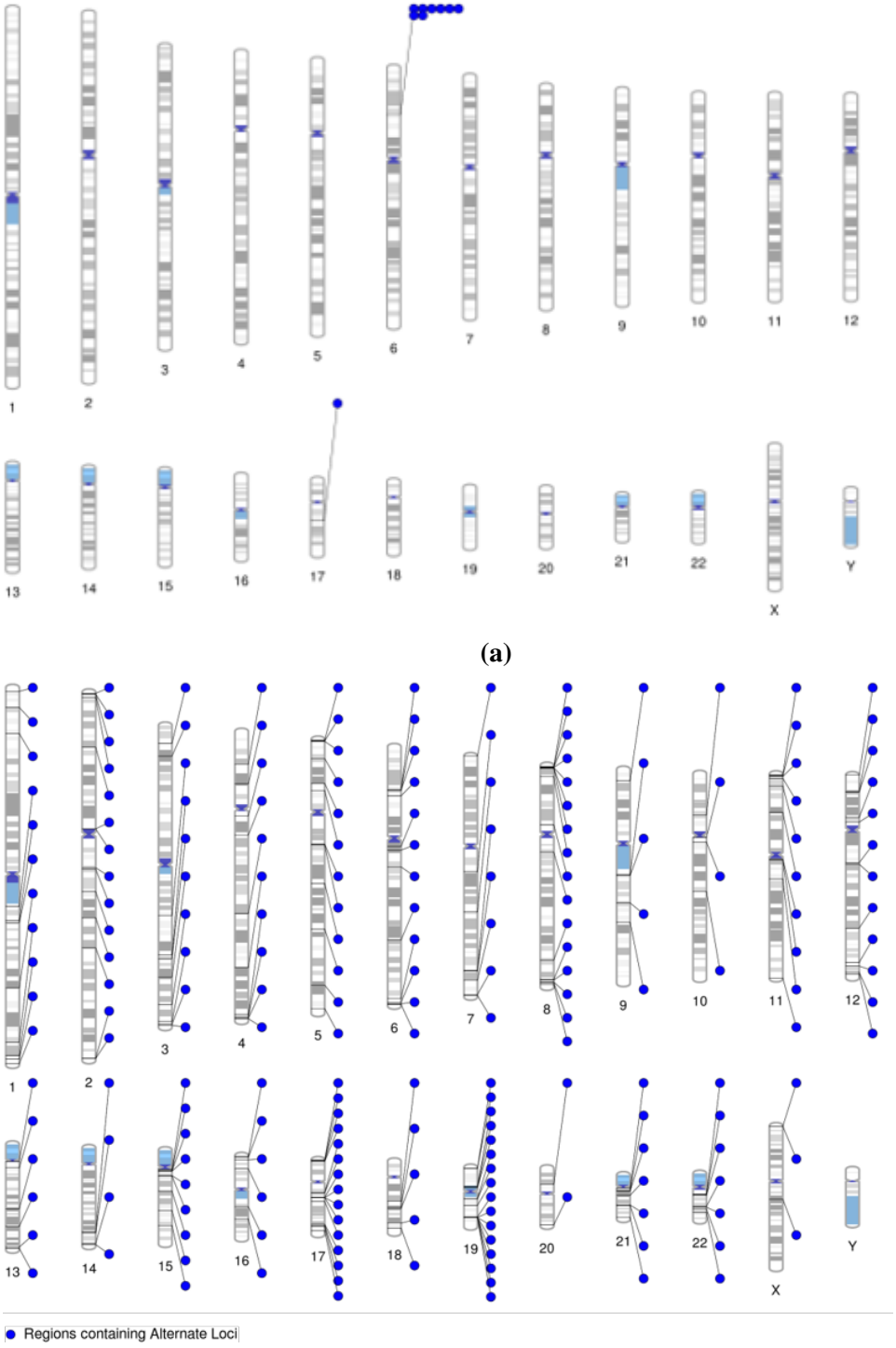
Project, however, did not utilize the nine alternate loci (The 1000 Genomes Consortium, 2012, 2015). This assembly did contain decoy sequences. These sequences are placed in the reference assembly, to which viral and other contaminant DNA sequences can align to. Furthermore, DNA from regions which have significant variation to the haploid reference sequence can also map to the decoy sequences.

The most recent genome assembly - GRCh38, was released in 2013. GRCh38 attempted to better represent the natural variation present in the human genome through the inclusion of 261 alternate loci spread across 178 regions (Figure1.7B). As GRCh38 only included seven alternate loci within the *HLA* region, *HLA* allele sequences obtained from the IMGT/HLA database (Robinson *et al.*, 2015) have been included to better capture the possible variation within the *HLA* region. To utilize the alternate loci, tools have to be ”alt-aware”, which is the ability of tools to connect the alternate loci to the haploid reference sequence as well as the DNA sequences that align to the alternate loci.

As knowledge about the structure of the variation within the human genome has improved, it has become apparent that a haploid reference sequence fails to capture the natural variation present in the human genome. However, individual human populations possess unique variations which cannot be captured in a reference sequence. Examples of variation which cannot be captured in a traditional genome assembly include large structural changes such as inversions, to large deletions and insertions.

Pangenomes are reference assemblies which include the entire gene set for a species. As population dependent variation is becoming increasingly apparent in the human genome, population-specific pangenomes have been an interesting research topic. Recently, a pangenome for a Danish population was created from 150 non-admixed individuals (Maretty *et al.*, 2017). However, the 150 individuals consisted of 50 trios (Mother, father and offspring) and therefore did not represent a heterogeneous cohort. Researchers have also attempted to create a pangenome utilizing an African cohort. It was found that the African pangenome contained

approximately 10 percent more DNA then GRCh38 (Sherman *et al.*, 2018), which suggests that further research into pangenomes is required.



(a)

(b)

Figure 1.7: Ideogram of reference assemblies (a) GRCh37, displaying nine alternate loci and (b) GRCh38, displaying 261 alternate loci. Figure produced using PhenoGram (Wolfe *et al.*, 2013)

1.12 Mapping and Alignment

Once read data has been generated through NGS, read mapping and alignment is performed. Mapping refers to the process in which the genomic location of reads are determined. This process is less computationally intensive then read alignment, and is useful for quick assignment of reads to genomic features such as genes without requiring exact alignment of the read to the reference assembly. Tools which utilize this include Kallisto (Bray *et al.*, 2016) and Salmon (Patro *et al.*, 2017) and are often used for RNA-seq data, where exact alignment is not always required.

Read alignment utilizes a FASTQ file, which contains the read identifier, the individual nucleotide calls, and the quality of each nucleotide as a Phred score. With alignment the nucleotide sequence is matched to the reference assembly. Read alignment involves comparing each individual nucleotide in a read to the reference assembly. Through the comparison, mapping quality score is assigned to the position to which the read aligns. The score is created from a composition of each matching nucleotide in relation to the nucleotide quality. In the case of variations to the reference assembly, deductions in the mapping quality occur. This decreases the chance of multimaps, which is when a read can align to numerous regions, as the highest mapping quality is indicative of the correct alignment.

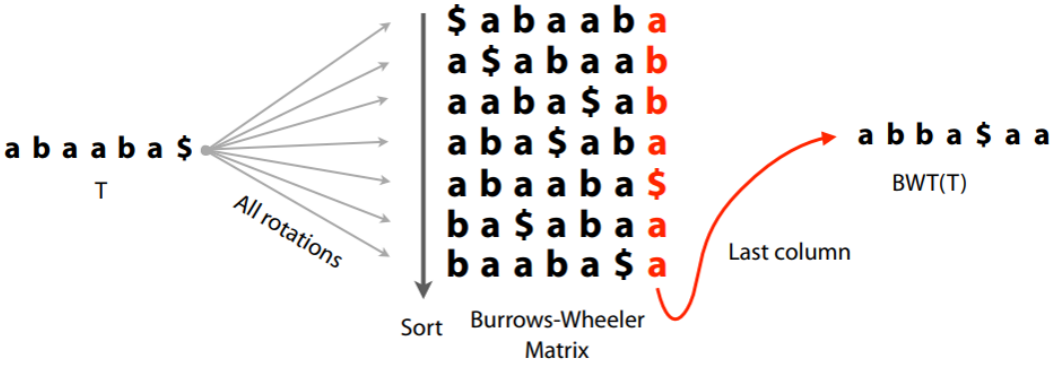
As modern reference assemblies include alternate loci, often reads will align to one, or more, of the alternate loci. In this situation, the alignment to the alternate loci is recorded as a secondary alignment, to allow for the read to be mapped to the linear haploid reference sequence. Once the reads have been aligned to a reference assembly, the output is a sequence alignment (SAM) file. To reduce the computational size of this file, binary compression is performed which results in a binary alignment (BAM) file.

1.12.1 Linear Alignment

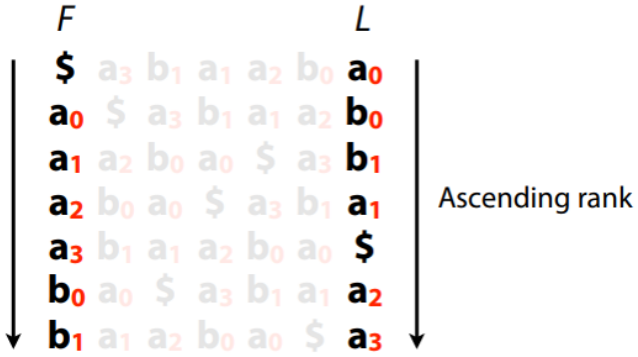
With the introduction of second-generation sequencing, large amounts of data has been generated. To align the data to a genome is a challenging process which requires a large amount of computational power. Early aligners relied upon

hashing techniques attempts to map data of variable size to data of a fixed size. Early algorithms utilizing the hash function, such as RMAP (Smith *et al.*, 2008) and MAQ (Li *et al.*, 2008) would utilize a hash function on the reads, and thereafter read through a reference sequence to identify areas to which the reads could map. Other algorithms, such as SOAPv1 (Smith *et al.*, 2008) and NovaAlign (http://www.novocraft.com/), would utilize the hash function on the genome, thereby creating a lookup (hash) table for reads to aligned to (Smith *et al.*, 2008). This approach requires a large amount of RAM, especially for large genomes. As the human genome has incorporated alternate loci, the size of the genome has increased, and therefore the need for algorithms with a decreased RAM footprint have been beneficial.

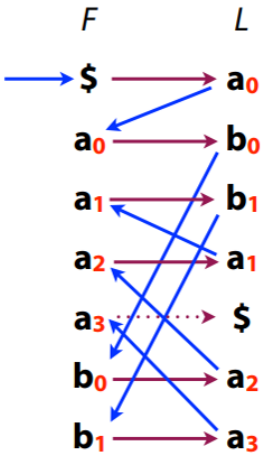
The Burrows Wheeler Transform (BWT) is a method of rearranging a character string originally created to aid in file compression (Burrows and Wheeler, 1994). The transform utilizes a front to back rotation of a string (Figure 1.8a). Thereafter, the occurrence of a character is counted and the value of each character is assigned to the character. The array of rotated strings is then lexicographically sorted (Figure 1.8a). Thereafter, all data within the array, with the exception of the last line can be discarded. This allows for compression, as similar characters are grouped together. This method has been found to have a good compression ratio which is important for large genomes. Utilizing the Burrows-Wheelers Transform has resulted in the size indexed reference assembly equal to ±1.2X the size of the genome. Other methods, which utilize suffix arrays, produce an index roughly ±5X the size of a reference assembly(Kurtz, 1999). Numerous algorithms have been developed which utilize the BWT, namely - BWA (Li *et al.*, 2009a), SOAPv2 (Li *et al.*, 2009b) and Bowtie (Langmead, 2010).



(a) Lexicographic sorting used in the BWT



(b) First and Last column in the BWT



(c) First-to-last column matching in the BWT

Figure 1.8: Example of the Burrows-Wheeler Transform. Obtained from http://www.cs.jhu.edu/~langmea/resources/lecture_notes/10_bwt_and_fm_index_v2.pdf

FM Index

For the purpose of read mapping and alignment, the BWT cannot be utilized directly. For alignment, a method of utilizing the BWT string was identified by Lippert (2005), which utilized a data structure named - Full-text in Minute space index (FM Index)(Ferragina *et al.*, 2004). The structure of a FM index consists of the BWT, which is the ”last” column (Figure 1.8b). From the BWT, the ”first” row is generated. Thereafter, by counting each unique character traversing down the ”last” column, the number of each character can be assigned (Figure 1.8b). This allows for when a read has been mapped to a row in the FM index, the genomic position of the row can be determined.

The FM index allows for read matching from a right-to-left direction to a reference genome. This is performed through a first-to-last matching method, where using the BWT, which is the last column, and reversing the process to create the first column, a string can be aligned. The first character from the right of a string is matched to the last column. The matching characters from the last row are identified in the first row. Thereafter the next character in the string is matched to the last column of the rows with the previously matched characters. This is repeated for the length of the query string, until the final matching row is identified (Figure 1.8c). This allows the the location in the genome, which the read maps to, to be determined.

1.12.2 Graph-based Alignment

Due to the structure of a linear reference assembly, a read which varies to the reference can be excluded from mapping and aligning. Through this, variation to the reference assembly may be under-reported. When this occurs, a bias is observed in which the reference assembly sequence is over-reported. This bias has previously been identified in the 1000 Genomes Project phase 1 data (Brandt *et al.*, 2015). A method to decrease this bias is through the use of genome graphs. In the linear reference assembly, natural variation is incorporated through the addition of alternate loci. In a graph genome, the genetic variation is reported as a pangenome, where the variation is incorporated as separate path connected to a reference sequence by nodes (Figure 1.9). This results in less reference strand bias, which

the under-reporting of data which differs to the reference sequence. This can occur as when mapping reads to the reference strand, the read must be similar enough to the reference sequence to not be excluded. Whilst the addition of alternate loci has decreased reference sequence bias, structural variations and variations not present in the reference assembly can result in the discarding of true variants.

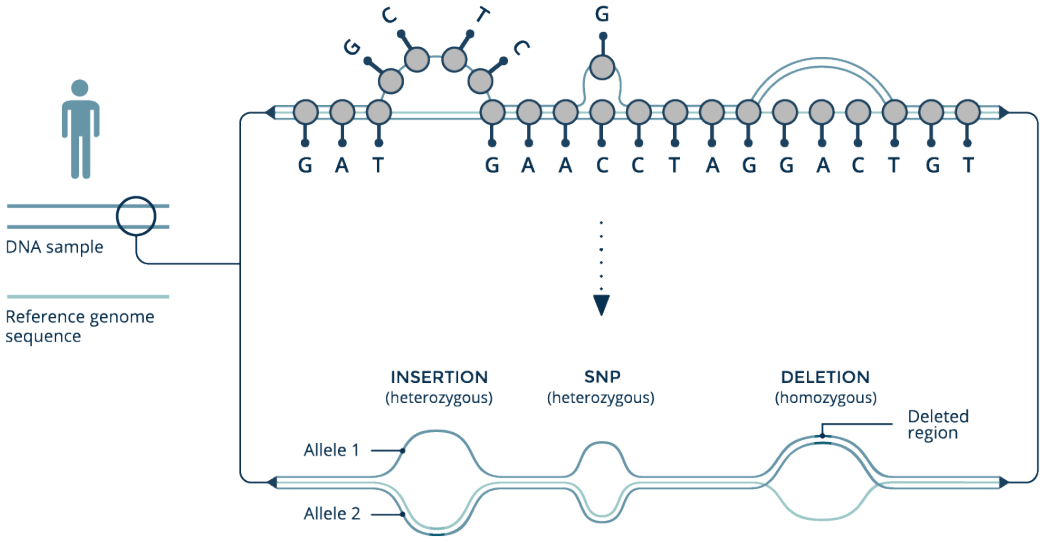


Figure 1.9: Example of a genome graph. Obtained from https://www.sevenbridges.com/graph/

de Bruijn Graph

A popular graph type is the de Bruijn graph. These graphs utilize a directed graph. A directed graph is made up of a set of nodes and a set of edges. A directed graph represents the connection between nodes that in unidirectional. In the example given (Figure 1.9), from left to right, A can only be followed by T, which can only be followed by C or G, which can only be followed by G. The use of de Bruijn graphs has largely been restricted to *de novo* genome assembly. Programs which utilize de Bruijn graphs include SOAP*denovo*2 (Luo *et al.*, 2012) and Velvet (Zerbino and Birney, 2008).

BWT FM Index Graph

More recently, another graph type has been utilized, which extended the BWT to a graph genome through the use of a hierarchical FM index. This involves utilizing the linear sequence of the reference genome. Thereafter, structural and nucleotide

variations are incorporated as alternate paths to form a directional genome graph (Kim *et al.*, 2015). Thereafter, the graph is converted to a lexicographically prefix-sorted graph, similarly to the BWT, which allows for quicker identification of plausible mapping sites. Through the use of the sorting, a FM Index is formed. This results in a first-to-last column, with outgoing edges in the first column and incoming edges in the last column. As the linear reference sequence is approximately 3 billion base pairs long, the FM index is partitioned into smaller indices which are referred to as local indices. This allows for local alignment, which is important for RNA sequence alignment across exon-intron boundaries as well as increased efficiency when compared to using a global FM index (Kim *et al.*, 2015). The use of the local indices with the correlating global FM index forms the hierarchical FM index (Kim *et al.*, 2015).

1.13 Difficulties with NGS *HLA* Genotyping

HLA typing from Next-generation sequencing (NGS) data was first performed in 2011 using Roche 454 generated reads (Erlich *et al.*, 2011), and was found to be capable of identifying *HLA* alleles to a four-digit resolution (Holcomb *et al.*, 2011). Roche 454 sequencing was a viable method, as it produced reads between 250 bp and 700 bp (Bentley *et al.*, 2009), which were longer than *HLA* exons. However, Roche 454 was expensive and prone to homo-polymer errors (Table 1.7), which are short repetitive regions of the same nucleotide. Furthermore, the limited throughput of Roche 454 (Approximately 700 Mb) was not suited to *HLA* sequencing (Carapito *et al.*, 2016). Another technology, used by Ion Torrent PGM, measures hydrogen ion release when a nucleotide in incorporated into the DNA strand. This technology has the benefits of short read times, as well as longer read lengths. While the reads used by this technology were capable being used to genotype *HLA*, this technology was also prone to homo-polymer errors, which can influence the accuracy (Barone *et al.*, 2015; Carapito *et al.*, 2016).

Illumina utilizes sequencing by synthesis, which measures the fluorescence of a fluorescent nucleotide when it is incorporated into the DNA strand (Turcatti *et al.*, 2008). This platform is capable of generating reads up to 300 bp long, and is

further capable of generating paired-end reads. Paired-end reads consist of a single fragment of DNA, which is incompletely sequenced from both ends. This results in a partial forward sequence paired with a partial reverse sequence, with a known insert size between the reads. This is beneficial when aligning reads to the *HLA* region, through reducing multi-maps, or the occurrence of a read which can align to separate regions of the genome. As the *HLA* region includes numerous loci derived from duplication events, the possibility of a read aligning to two loci is increased. Therefore, if one read, in a pair, aligns to two regions, the other paired-end can infer the correct mapping location. Illumina is furthermore beneficial to *HLA* sequencing, as the platform is not prone to homo-polymer errors, there is decreased costs, and there is increased accuracy when compared to competing platforms (Table 1.7).

Life Technologies 5500XL SOLiD generates short read lengths (70 bp) (Table 1.6). The short read length is not ideal for *HLA* genotyping, as phasing ambiguities are possible, which occurs when incorrect phasing of exons 2 and 3 occur (Figure 1.6). This would be further compounded by numerous *HLA* alleles within the IMGT/HLA database (Robinson *et al.*, 2015) which have only been sequenced across these regions, and therefore, the intronic region between the exons cannot be used to resolve the ambiguity. Furthermore, the G/C bias introduced by SOLiD sequencing would furthermore result in decreased accuracy in *HLA* genotyping (Carapito *et al.*, 2016).

The third-generation sequences introduced methods to generate substantially longer reads than previous technologies (Table 1.6). Pacific Bioscience RS2 is capable of generating reads up to 40 kb in length. This would result in *HLA* haplotypes without phase ambiguity errors. The high error rate of this technology requires numerous passes over the same immobilized DNA to prevent errors (Carapito *et al.*, 2016). Oxford Nanopore Technologies MinION is another third generation sequencing platform, which utilizes nanopore sequencing to generate long reads. This has the same benefit as Pacific Biosciences, in removing phase ambiguity of *HLA* genotypes. Nanopore sequencing is still in its infancy and requires improvement to the accuracy of the reads, as the high error rate is not

ideal for *HLA* genotyping. However, the reduced cost of this platform with the long reads which are generated, would make nanopore sequencing ideal for a hybrid sequencing approach, where the nanopore sequencing would be capable of inferring structural changes (Carapito *et al.*, 2016).

Table 1.7: Advantages and Disadvantages of different NGS technologies for *HLA* genotyping. Adapted from Xuan *et al.* (2013); Carapito *et al.* (2016); Churko *et al.* (2013)

Platform	Error prone type	Error rate	Pros	Cons
Ion Torrent PGM 318	Indels	1 %	Short sequencing times	Homo-polymer errors
Roche 454 GS FLX	Indels	0.1 %	Long reads High Accuracy	Homo-polymer errors Expensive
Life Technologies 5500Xl SOLiD	G/C bias	0.1 %	Low cost	Short read length
Illumina MiSeq	Substitutions	0.1 %	Paired-end Sequencing High Accuracy	Long sequencing times
Pacific PacBio RS2	Insertions	13 %	Long reads Affordable	Low accuracy Low throughput
Oxford Nanopore MinION	Indels	15 %	Long reads	Low accuracy Low throughput

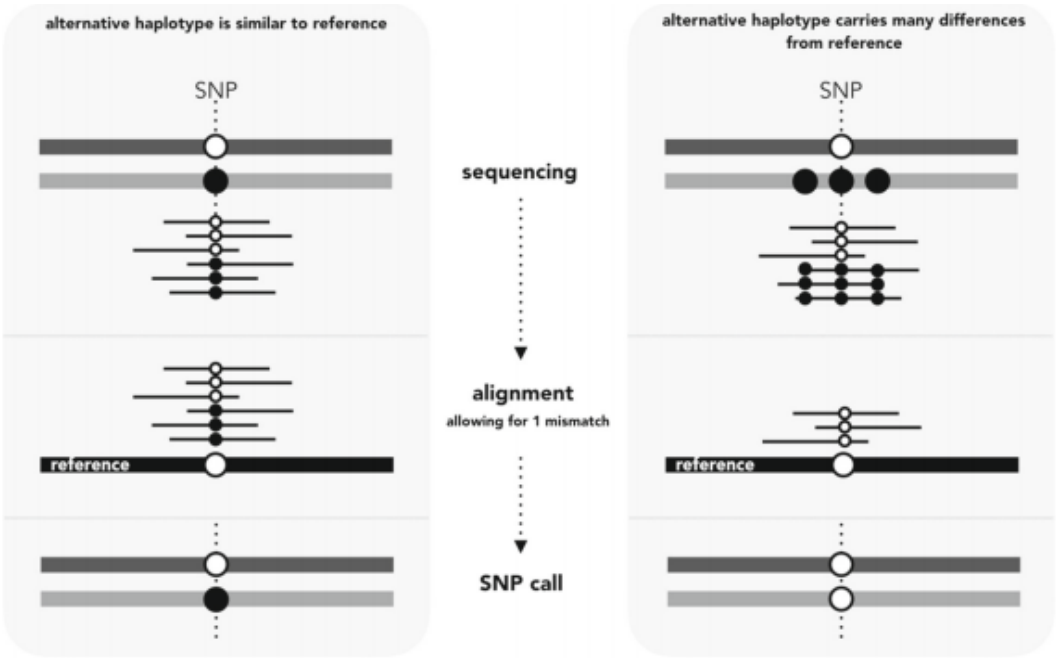


Figure 1.10: Effect of aligning of reads to a single reference genome compared to a reference assembly with alternate loci. Obtained from Meyer *et al.* (2018)

Both depth and breadth of coverage have been reported to have an effect on alignment and variant calling accuracy (Bauer *et al.*, 2016; Major *et al.*, 2013). Depth of coverage is the average number of informative reads per base and breadth of coverage is the number of bases of the reference genome (or alternate locus) covered by sequencing reads. It has previously been found that breadth of coverage has a far higher impact on the accuracy of variant calling then depth of coverage (Bauer *et al.*, 2016). Within the *HLA* region, this can have a greater impact on genotyping accuracy, as if the intron between two exons is insufficiently covered, phasing the two exons can be inaccurate.

HLA typing from NGS data presents a separate set of difficulties, largely due to the degree of polymorphisms present as well as the degree of sequence similarity between genes. During mapping, reads which differ significantly from the reference (or alternate locus) sequence are discarded due low quality sequence alignment between the reads and the reference genome (Figure 1.10). High sequence similarity provides opportunity for multimaps, again resulting in mapped data being discarded (Nielsen *et al.*, 2011; Treangen and Salzberg, 2012). GRCh38

has improved HLA mapping by including eight alternate loci, however due to the high variation found within this region, the inclusion of more alternate sequences is beneficial, which is why sequences obtained from the IMGT/HLA database (Robinson *et al.*, 2015) can be included.

1.14 Aim and Objectives

HLA plays a vital role in immune regulation (Bjorkman and Parham, 1990), and due to unique selective pressures (Prugnolle *et al.*, 2005; dos Santos Francisco *et al.*, 2015), has evolved to be the most variable region within the human genome (Robinson *et al.*, 2015). Through the variation within this region, and the function of *HLA*, different *HLA* alleles have been found to be associated with numerous diseases. As many of these associations have not been attributed to a single SNV within a *HLA* allele, genotyping individual *HLA* alleles is vital. Furthermore, many of these associations have been found to restricted to specific populations, for which further research may rely on larger, population-specific cohorts.

Through the introduction of NGS, massive amounts of data have been generated. From large whole genome sequencing studies, such as the 1000 Genomes Project, to the growth of precision medicine approaches which often sequences individual genomes, the need for fast and accurate *HLA* genotyping from NGS data has become apparent. With this, however, numerous approaches are available. The human reference genome has evolved from a haploid linear sequence to an assembly which includes alternate loci. Reference graphs have been developed to better represent variation within the human genome. Finally, knowledge of population-specific variation has improved, and with that, presented further difficulties.

While targeted sequencing can result in high quality NGS data, from which accurate *HLA* genotyping is possible (Major *et al.*, 2013), this is not always performed and often represents an additional step. Therefore, accurate *HLA* genotyping from existing NGS data is the aim of many tools. Furthermore, utilizing targeted sequencing, in which exons are sequenced, present difficulties in

HLA genotyping, as phasing between exons typically relies upon prior knowledge about linkage. Within the *HLA* region, LD differs from other regions in the human genome (Jain, 2011), and therefore, it is possible to resolve these phasing ambiguities through the use of WGS data due to the sequencing of intronic regions.

The aim of this study was to compare the different approaches used by four NGS *HLA* genotyping tools (BWAkit, xHLA, Kourami and HISAT-Genotype) that utilize WGS data, and the possible effects populations-specific and/or locus-specific variability may have on these approaches. The objectives of this study were therefore:

1. To *HLA* genotype 12 individuals for whom both high-resolution *HLA* SBT genotype and high coverage (30X) WGS data were available, using four NGS *HLA* genotyping tools: BWAkit, xHLA, Kourami and HISAT-Genotype.

2. To evaluate the accuracy of these four tools (BWAkit, xHLA, Kourami and HISAT-Genotype), by comparing the genotypes obtained to corresponding high-resolution *HLA* SBT genotyping data for 12 individuals.

3. To evaluate *HLA* variability across the three classical *HLA* class I loci (*HLA-A*, *HLA-B* and *HLA-C*), in 1267 individuals, from four super-populations, using high-resolution *HLA* SBT genotyping data.

4. To evaluate the effects that population specific and/or locus-specific variability may potentially have on the accuracy of the four *HLA* genotyping tools evaluated.

Chapter 2
Methods and Materials

2.1 Cohort Selection and Data Acquisition

Twelve individuals were selected in order to evaluate four *HLA* genotyping tools that utilize short-read NGS data (Table 2.1). These 12 individuals were selected due to the availability of high-coverage (30X) 250 bp Illumina paired-end WGS data from the 1000 Genomes Project, as well as corresponding high-resolution SBT *HLA* genotyping data (Gourraud *et al.*, 2014). The SBT data were obtained for the classical class I loci, *HLA-A*, *HLA-B* and *HLA-C*. SBT genotyping, which is considered the ”gold standard” for *HLA* genotyping (Shankarkumar *et al.*, 2008), was performed by gene-specific PCR amplification of exons 2 and 3 of each class I gene, followed by Sanger sequencing (Gourraud *et al.*, 2014). Gourraud *et al.* (2014) compared these sequences to sequences obtained from the IMGT/HLA database (v. 2.26; Robinson *et al.* 2015). For the purposes of this study, the allele data was converted into ambiguous allele codes using the IMGT/HLA ambiguous allele combinations database (v. 3.28; Robinson *et al.* 2015). The list of alleles representing these ambiguous combinations have been provided for these 12 individuals in Supplementary Tables B3-B5.

The high-coverage WGS data for these individuals were obtained for the *HLA* class I region on chromosome six (GRCh37 - Chr6: 2 987 500 - 3 236 000), decoy sequences, and non-aligned reads, in the form of a BAM file (The 1000 Genomes Consortium, 2015). This region was selected, as it contained the other *HLA* class I loci, to which classical class I reads could have aligned (Table A3). Decoy sequences are sequences included within a reference assembly, to which non-origin reads (such as reads originating from the Epstein-Barr virus, as well as regions which have not been completely incorporated into the assembly) align

(The 1000 Genomes Consortium, 2015). Reads aligning to decoy sequences, as well as non-aligned (unmapped) reads, were also included, as these may contain additional reads that might align to the *HLA* region in GRCh38. The data were obtained from the 1000 Genomes Project FTP server in BAM format (Table A2). Samtools (v. 1.5; Li *et al.* 2009a), which provides numerous tools for manipulating BAM files, such as subsetting, sorting, merging and indexing (Li *et al.*, 2009a), was used to download the data (Listing C1). From the GRCh37-aligned BAM files, read counts pertaining to the *HLA* region were obtained, again using SAMtools (v. 1.5).

Table 2.1: Description of the 12 individuals for whom both high-coverage (30X) WGS and high-resolution SBT *HLA* genotyping data were available

Sample	Population	Population code
HG00096	British in England and Scotland	GBR
HG00268	Finnish in Finland	FIN
HG00419	Southern Han Chinese	CHS
HG01051	Puerto Ricans from Puerto Rico	PUR
HG01112	Colombians from Medellin	CLM
NA18939	Japanese in Tokyo, Japan	JPT
NA19238	Yoruba in Ibadan, Nigeria	YRI
NA19239	Yoruba in Ibadan, Nigeria	YRI
NA19240	Yoruba in Ibadan, Nigeria	YRI
NA19625	Yoruba in Ibadan, Nigeria	YRI
NA19648	Mexican Ancestry, Los Angeles, USA	MXL
NA20502	Toscani in Italia	TSI

2.2 Selection of *HLA* Genotyping Tools

The accuracy of *HLA* imputation has previously been found to be problematic. This is due to the inaccuracy of individual SNP calls reported in the VCF file following WGS variant calling. In accordance with this, a previous study found that 18.6 percent of reported SNP calls within the *HLA* region, in individuals from

the 1000 Genomes Project, were found to be incorrect (Brandt *et al.*, 2015). As a consequence, *HLA* imputation is generally only accurate at a two-digit resolution. Therefore, to genotype *HLA* at higher resolutions requires a different approach. This study, therefore, aimed to compare different approaches that utilize read data, rather than SNP data, to genotype *HLA*.

In terms of read alignment methods, there are two widely used approaches, namely linear- and graph-based alignment methods. Linear-based alignment strategies have commonly been used, as the initial human reference sequences consisted of linear haploid sequences. With the inclusion of alternate loci in the GRCh37 and GRCh38 genome builds, there was a need to develop alt-aware methods. This allowed more modern aligners to align reads to alternate loci, while maintaining their position in relation to the reference sequence coordinates. Alternatively, graph-based methods incorporate alternate loci into genome graphs, which provide numerous potential paths along which reads can be aligned. This is performed to better represent natural variation that would not be considered in the alignment process. This further decreases reference bias, as reads that vary from the reference sequence may be aligned to alternative paths within the graph (Paten *et al.*, 2017).

Table 2.2: Overview of the four *HLA* genotyping tools evaluated in this study

Tool	Version	Highest Resolution	Ambiguous Allele Output	Input	Method	IMGT/HLA DB version	Reference Assembly
BWAkit	v. 0.7.11	8-digits	No	FASTQ	Assembly	3.18	GRCh38DH[1]
HISAT-Genotype	v. 1.0.1b	6-digit	No	FASTQ	Alignment + Assembly	3.31	GRCh38DH, Consensus sequence obtained from MSA of IMGT/HLA database, SNPs and Indels from dbSNP
Kourami	v. 0.9.6	6-digit	Yes	BAM	Assembly	3.24	GRCh38DH or GRCh38[2]
xHLA	Released:04/10/2017	6-digit	No	BAM	Alignment	3.21	GRCh38

[1] GRCh38DH - GRCh38.p7 with alternate loci and IMGT/HLA database.

[2] GRCh38 - GRCh38.p7 reference sequence

To better compare and contrast the affect of these alignment approaches on *HLA* genotyping, *HLA* genotyping tools using both linear- and graph-based alignment strategies were identified and evaluated. BWAkit and xHLA, were chosen to represent the linear-based methods, while Kourami and HISAT-Genotype were chosen to represent graph-based methods. Furthermore, the two graph-based methods can be distinguished from each other in that Kourami incorporates alt-aware alignment, whereas HISAT-Genotype does not. Ultimately, these four tools were therefore selected because they incorporate distinct, but overlapping combinations of algorithmic strategies (Figure 2.1), making them well suited to evaluate the effectiveness of these individual approaches.

2.3 Data Preprocessing

As the BAM files were aligned to GRCh37, the reads needed to be reverted to FASTQ files, as the four tools required either FASTQ (BWAkit and HISAT-Genotype) or GRCh38-aligned BAM files (xHLA and Kourami) as input. The aligned BAM files were reverted to unmapped BAM files using RevertSam included in Picard (v. 2.17.6). This resulted in BAM files, where the positional information had been removed. During the data retrieval process, paired-end reads in which one of the pairs were located outside of the regions specified may have been lost, resulting in unpaired reads. Furthermore, reads that were aligned at locations on the boundaries of these regions may have been truncated. Therefore, these truncated or unpaired reads were removed using the ”sanitize” command within RevertSam (Listing C2). Reads were then shuffled, and sorted according to query name. This was done in order to prevent possible bias introduced by the positional order of the reads within the original BAM file. Finally, original quality scores were restored, as the pipeline initially used to align the reads (The 1000 Genomes Consortium, 2015) performed a quality control step, in which the qualities may have been adjusted.

The unmapped BAM files were then converted to paired-end FASTQ files using SamToFastq within Picard (v. 2.17.6)(Listing C3). The paired-end reads were split, based on the directionality of the read, into two files, with the suffix ”.1.fq”

(forward) and ".2.fq" (reverse) to differentiate the two files. The resulting FASTQ files were then either aligned to GRCh38, using BWA-mem (in the case of xHLA and Kourami), or provided as direct input for the *HLA* genotyping tools (BWAkit and HISAT-Genotype). For the linear alignment required by xHLA, the reads were aligned to the haploid GRCh38 reference sequence. Alternatively, for the alt-aware alignment required by Kourami, the reads were aligned to the GRCh38 reference assembly, alternate loci, and allele sequences from the IMGT/HLA database (v. 2.26) (Listing C4). Once the reads were aligned, the counts of reads aligning to the *HLA* regions were determined (Listing C5).

In order to analyze the depth of coverage across the classical *HLA* class I region, the read depth was also determined. This was performed through DepthOfCoverage, included in the Genome Analysis Toolkit (v. 3.7.0; Van der Auwera *et al.*2013). As alternate loci were included in the alignment strategy, calculation of read depth included reads aligned to alternate loci. To do this, the positional coordinates and region-specific annotations for the *HLA* loci were obtained using the Table Browser tool, available through the University of California, Santa Cruz Genome Browser (Karolchik, 2003). Thereafter, the depth of coverage was calculated, per locus, per sample (Listing C6).

2.4 *HLA* Genotyping of High-Coverage WGS Data

In this study, 12 individuals were selected in order to evaluate the performance of the four selected *HLA* genotyping tools: BWAkit, xHLA, Kourami and HISAT-Genotype. Genotyping was performed for the classical *HLA* class I loci, namely *HLA-A*, *HLA-B* and *HLA-C*. The methodology employed for each tool followed the best practices recommended by the tool manuals (Figure 2.1).

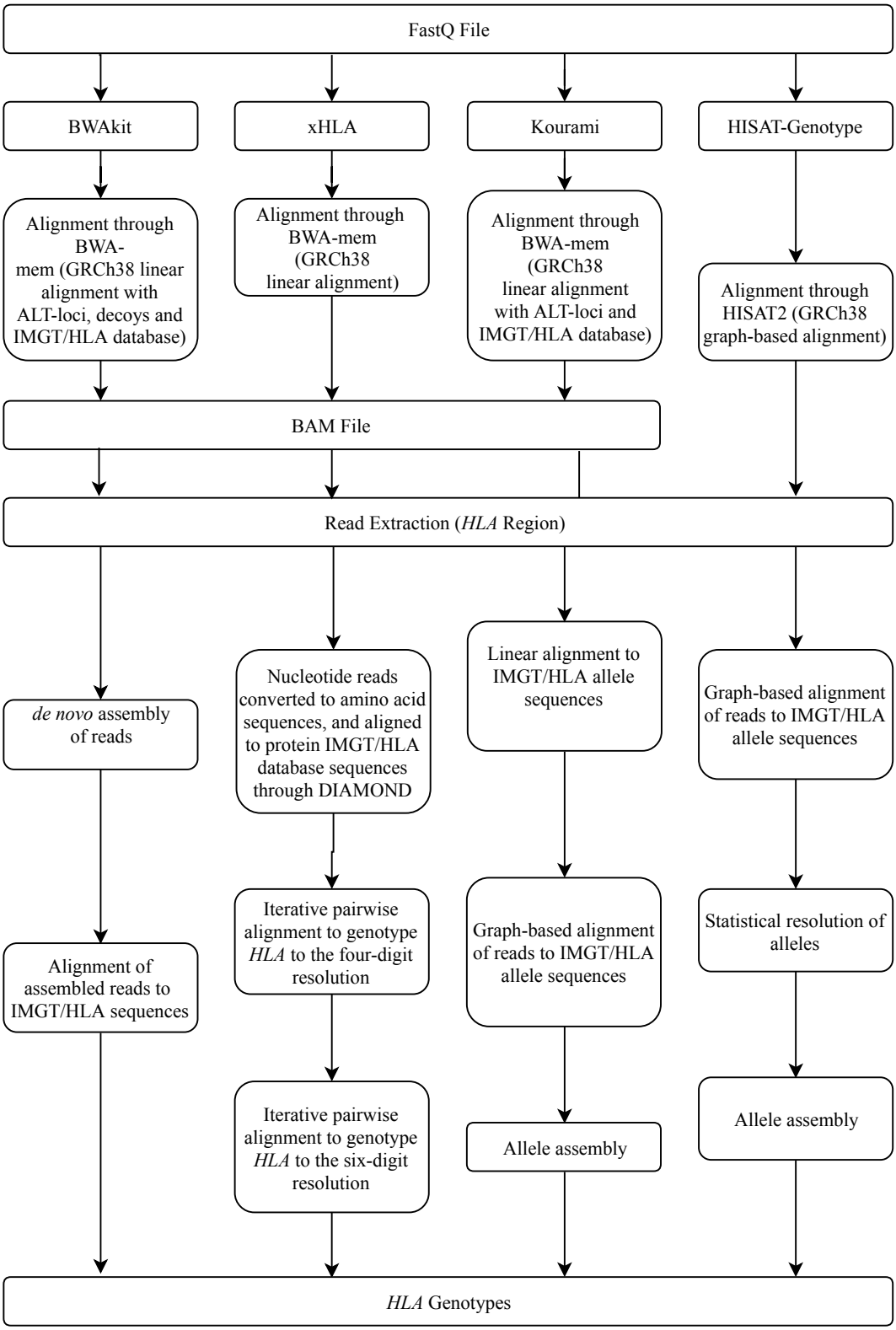


Figure 2.1: Flow diagram outlining the algorithmic steps utilized by BWAkit, xHLA, Kourami and HISAT-Genotype

2.4.1 BWAkit

BWAkit (v. 0.7.11) is an extension for BWA, which is able to utilize alternate loci to improve read mapping, and through this, has additional functionality to genotype the *HLA* region. BWA employs the BWT with a modified FM index. Firstly, as the standard use of the FM index with the BWT fails when an exact match cannot be determined, methods that allow for aligning reads that contain variants are required. BWA utilizes backtracking for inexact matching (Li *et al.*, 2009a). When the algorithm encounters a situation where an exact match cannot be produced, the algorithm reverses across the read, until a nucleotide with low sequencing quality (Phred score) is encountered. Thereafter, the nucleotide is ”substituted” and the process of first-to-last matching is continued. This step can incorporate single nucleotide substitutions, deletions and insertions.

Secondly, most technologies produce reads with higher quality scores towards the 5’ end of the read. In typical left-to-right matching, algorithms apply a threshold to the number of mismatches between a read and the FM index to prevent incorrect mapping and alignment of reads. As the read quality decreases along a read, increased variation can be expected due to sequencing errors. This can introduce bias, as reads that align along the 5’ end but not the 3’ are prevented from aligning. To resolve this, BWA utilizes a mirrored FM indexed, which is an index of the reverse reference assembly. This allows for string matching to be performed from a left-to-right and a right-to-left direction.

BWAkit includes BWA functionality in the script ”run-bwamem”, with additional commands to specify *HLA* genotyping (Listing C7). Through the script ”run-bwamem”, reads are aligned to a reference assembly (in this case, GRCh38.p7), which includes alternate loci, decoy sequences and *HLA* allele sequences obtained from the IMGT/HLA database (v. 3.24) using BWA-mem (v. 0.7.11) (Table 2.2). Following this, BWAkit performs post-alignment processing, in which the coordinates of reads that aligned to alternate loci or IMGT/HLA allele sequences are lifted over to the GRCh38 reference sequence, in order to maintain consistent positional information. Thereafter, reads that align to the *HLA* region are extracted, on a per-locus basis, and *de novo* assembly of the reads is performed.

The assembled reads are then aligned pairwise to exon sequences obtained from the IMGT/HLA database (v. 3.18; Robinson *et al.* 2015). *HLA* alleles to which the reads align are then subset and iterative alignment is performed, using increasing mismatch penalties until only two alleles per gene remain. The alleles with the highest reported alignment score are then assigned as the genotype. In order to facilitate comparisons across tools, reported genotypes were renamed using ambiguous allele codes (Supplementary tables B3-B5).

2.4.2 xHLA

xHLA takes as input a BAM file, which has been aligned to the linear haploid reference sequence only. Reads were aligned to GRCh38 in a non-alt-aware manner using BWA-mem (v. 0.7.11: Table 2.2). The BAM file was subsequently positionally sorted, and indexed. This was performed using SAMtools (v. 1.5; Li *et al.* 2009a). A fork of xHLA, which is implemented in the R language, xHLA_R (https://github.com/chrissyhroberts/xHLA_R), was used (Listing C8). Reads aligning to the *HLA* region (GRCh38.p7-chr6: 29 844 528 - 33 100 696) were extracted, and translated into amino acid sequences, before being aligned to amino acid sequences from the IMGT/HLA database (Robinson *et al.*, 2015) using DIAMOND aligner (v. 0.9.14.115; Buchfink *et al.* 2015).

Alleles to which reads align with 100 percent accuracy are retained, and used to form an alignment matrix, which is utilized in the four-digit and six-digit resolution *HLA* genotyping steps (Figure 2.2a). The first step of the four-digit genotyping step is then performed (Figure 2.2b), in which the reads are aligned to exons 2 and 3 of each potential allele in the solution set. This is done individually for each locus. The number of reads that align unambiguously to an allele are counted, to form a score for each allele. This forms the ”comp”. The ”comp” with the highest number of reads that align to it are retained in the solution set, and form the ”sol”. Thereafter, the subsequent allele used for comparison is referred to as the ”comp”. This process is repeated iteratively, resulting in a reduced solution set that consists of alleles with the highest number of aligned reads (Figure 2.2b).

Once each allele within the ”comp” has been compared to the ”sol”, genotyping to the four-digit resolution genotyping is performed (Figure 2.2c). This step is similar to the previous step, with respect to comparison of the ”comp” and ”sol”, however, pairwise-allele comparison is performed. In this step, reads are temporarily aligned to either the ”comp” allele or ”sol” allele, thereby preventing a read from contributing to the score of more than one allele at a time. This is performed iteratively, updating one allele at each run, until the highest scoring pair of alleles remains (Figure 2.2c). Once the best pair of candidate alleles per locus have been identified, a zygosity check is performed. If one allele in a pair has five times the number of aligned reads then the partnered allele, the heterozygous call is changed to a homozygous call (Xie *et al.*, 2017).

For six-digit resolution genotyping, xHLA repeats the iterative step (Figure 2.2b), using the four-digit genotyped alleles as a ”comp”. For this process, the full coding nucleotide sequence per allele is included in the alignment matrix (Figure 2.2d). Thereafter, the reads are aligned to each allele within the ”comp”. The allele with the highest number of aligned reads becomes the ”sol”, and this is repeated iteratively, with the allele with the highest number of aligned reads becoming the ”sol”, until each allele within the ”comp” has been compared. As a zygosity check was performed in the four-digit genotyping step, the two alleles with the highest number of aligned reads are considered the final six-digit resolution *HLA* genotype call (Xie *et al.*, 2017).

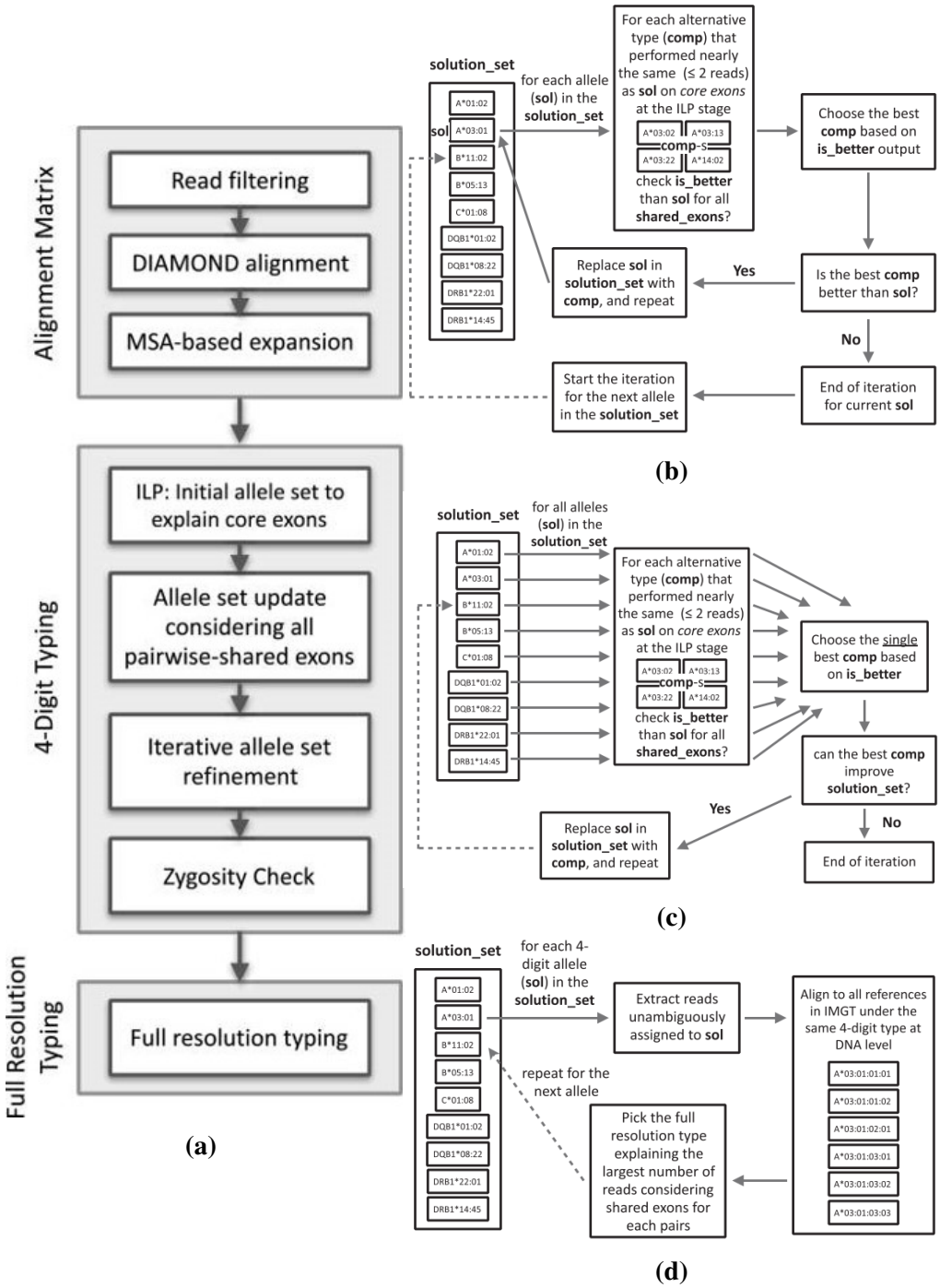


Figure 2.2: Schematic diagram of the xHLA genotyping algorithm depicting (a) the xHLA algorithm, (b) first round iterative allele set refinement, (c) second round iterative allele set refinement for four-digit resolution, and (d) *HLA* genotyping to the six-digit resolution. Figure obtained from Xie *et al.* (2017)

2.4.3 Kourami

Kourami (v. 0.9.6) takes as input a BAM file that has been produced using BWA-mem to align reads to GRCh38. This alignment can optionally include alternate loci and IMGT/HLA allele sequences. In this case, reads were aligned using BWA-mem (v. 0.7.11) to both the linear haploid GRCh38 reference assembly, and GRCh38 with alternate loci and *HLA* sequences obtained from the IMGT/HLA database (v. 3.24) (Table 2.2). The resulting BAM file was then positionally sorted and indexed using SAMtools (v. 1.5; Li *et al.* 2009a).

Reads that aligned to the *HLA* region (GRCh38.p7-chr6:29 723 340 - 33 129 113), as well as any IMGT/HLA allele sequences are extracted through the ”alignAndExtrac_hs38DH.sh” script (Listing C10); Figure 2.3b). Thereafter, the reads are aligned to the Kourami reference panel (Figure 2.3a). As numerous alleles within the IMGT/HLA database only have sequence information for exons 2 and 3, this panel consists of a MSA derived from both full length *HLA* genomic- (where available) and cDNA sequences obtained from the IMGT/HLA database. This MSA is then incorporated into a custom reference assembly, to which the extracted reads are aligned using BWA-mem. This results in an intermediate BAM file, containing reads with the information pertaining to the *HLA* loci to which the reads align.

Once this preprocessing has been completed, *HLA* genotyping is then performed by creating a gene-wise partial order graph (POG) (Figure 2.3d). The graph is constructed from the MSA obtained from the IPD database. Each sequence within the MSA is constructed as a chain of vertices with direct edges at each nucleotide (Figure 2.4b). The vertices are aligned into columns, with each column representing a nucleotide position. The redundant nucleotides are then grouped together (Figure 2.4c), and are collapsed to form a POG (Figure 2.4d).

The reads that align to the MSA, from the intermediate BAM file, are then projected onto the POG (Figure 2.3d). Each aligned read read provides a ”weight” to the POG. The weight in a POG refers to the score assigned to the node from one edge to another, as the read is aligned from left-to-right. If the majority of reads align

to one path through the POG, the score between the nodes on that path increases, which results in a lower alignment score for reads that do not align to that path. In the presence of a novel allele, variants that are not present in the original MSA can be included into the POG, through the modification of the POG due to the alignment projection (Lee and Kingsford, 2018).

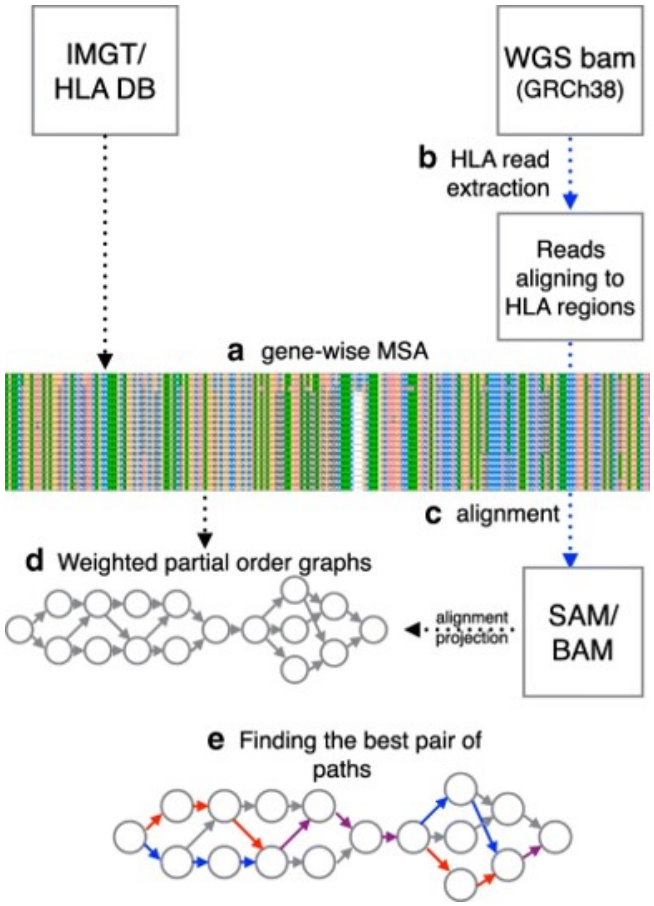


Figure 2.3: Schematic diagram outlining the Kourami genotyping algorithm. (a) A multiple sequence alignment per gene is obtained from the IMGT/HLA database. Reads aligning to the *HLA* region are extracted. (b) The extracted reads are aligned to the sequences within the MSA. (c) The alignments are exported and (d) projected onto a weighted partial order graph. (e) Haplotype assembly occurs by identifying the two best paths through the weighted partial order graph. Figure obtained from Lee and Kingsford (2018)

Once the weighted POG is formed, *HLA* assembly occurs by resolving the two best paths through the weighted POG (Figure 2.3e; Lee and Kingsford 2018). The phase of aligned reads is not taken into account when initially generating the weighted POG. In order to resolve the phase of individual reads, the phase of the

POG is resolved. This is performed using variants that occur on the same read or read-pair (local phasing). Next, contigs are constructed from aligned reads and using information from read-backed variants through the weighted POG (full-length phasing). This then provides phased paths through the POG (Lee and Kingsford, 2018).

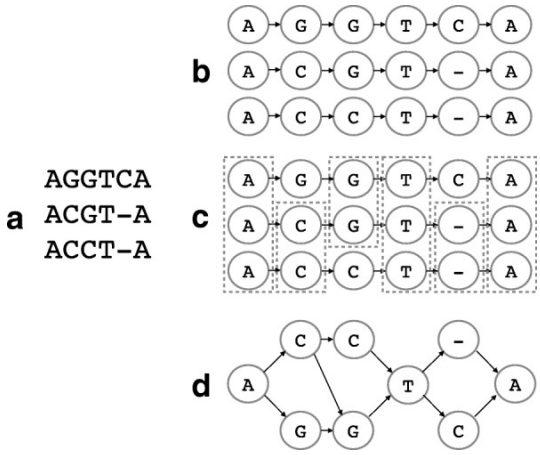


Figure 2.4: Schematic diagram describing the construction of a partial order graph for HLA assembly by Kourami. (a) The precompiled MSA is formed into (b) a chain of vertices connected by direct edges. (c) For each column, the redundant vertices are grouped together, and when collapsed, (d) form the partial order graph. Figure obtained from Lee and Kingsford (2017)

Once the phased paths through the weighted POG have been identified, a bubble graph is constructed from pairs of paths through the weighted POG (Lee and Kingsford, 2018). A bubble graph is a form of a graph, where the nodes are formed by homozygous regions, and the paths between nodes represent the variable positions (Figure 2.5). If more than two possible paths are present in the bubble graph, this may be due to sequencing errors or mis-alignments. Therefore, by pruning the discordant paths, the two best paths are retained. Thereafter, the paths are scored based on alignment scores, which take the sequencing quality of the reads into account. This is performed on all possible pairs of pathways, which can include the same path, as would occur in a homozygous individual. Once scoring is complete, the paths represent assembled alleles. The assembled alleles include all possible pairs of alleles explained by the bubble-graph.

To select the best set of assembled alleles, the phasing and alignment scores are combined, and the pair of paths with the highest score is extracted as two sequences. The sequences are then pairwise aligned to the sequences of the alleles obtained from the IMGT/HLA database, and the allele pair with the highest alignment score is assigned as the genotype. Genotypes are reported as ambiguous combinations of alleles (Lee and Kingsford, 2018).

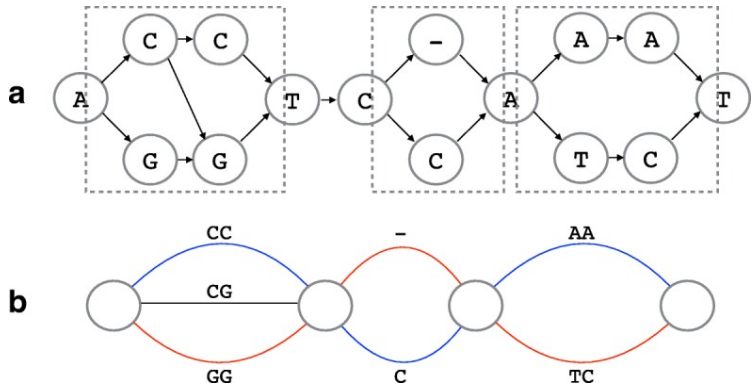


Figure 2.5: Diagram distinguishing between (a) a partial ordered graph and (b) a bubble graph. Figure obtained from Lee and Kingsford (2018)

2.4.4 HISAT-Genotype

HISAT-Genotype (v. 1.01b) utilizes a graph-based approach to *HLA* genotyping and takes as input either a single FASTQ file, or two FASTQ files in the case of paired-end reads. When aligning paired-end reads, pair information is not retained, rather reads are aligned individually and thereafter the alignments are combined. Read alignment is performed through HISAT2 (Kim *et al.*, 2015). The reference assembly used by HISAT2 additionally includes a collection of small variants and indels, obtained from dbSNP and *HLA* allele sequences from the IMGT/HLA database (Figure 2.6). The reason for the addition of small variants and indels to the assembly by HISAT2 is due to the primary reference sequence largely consisting of data from one individual (Green *et al.*, 2010), and a haploid linear reference does not therefore fully represent the genetic variation found within all human populations. From this, HISAT2 creates a sequence with alternate pathways to which the aligner can align reads.

From this, HISAT2 creates an FM index, similarly to BWA (Kim *et al.*, 2015). HISAT2 employs a hierarchical graph FM-index, which allows for the inclusion of

a subset of FM indices as hierarchical local indices. Included within the standard reference assembly are 48 000 local indices (Kim *et al.*, 2015), which increases to 55 000 within the HISAT-Genotype modified reference assembly (Kim *et al.*, 2018). The extra indices include additional *HLA* sequences, as well as SNPs and indels from dbSNP database. For each alignment, the read is compared to one of the 55 000 local indices (Figure 2.6). The local indices are used, as searching the global FM index is comparatively slower. This is because local indices fit into the CPU cache memory, and therefore do not completely rely upon RAM. Thereafter, once a read is aligned to a local index, the position of the local index on the global index is used to map the read back to the reference sequence (Kim *et al.*, 2015).

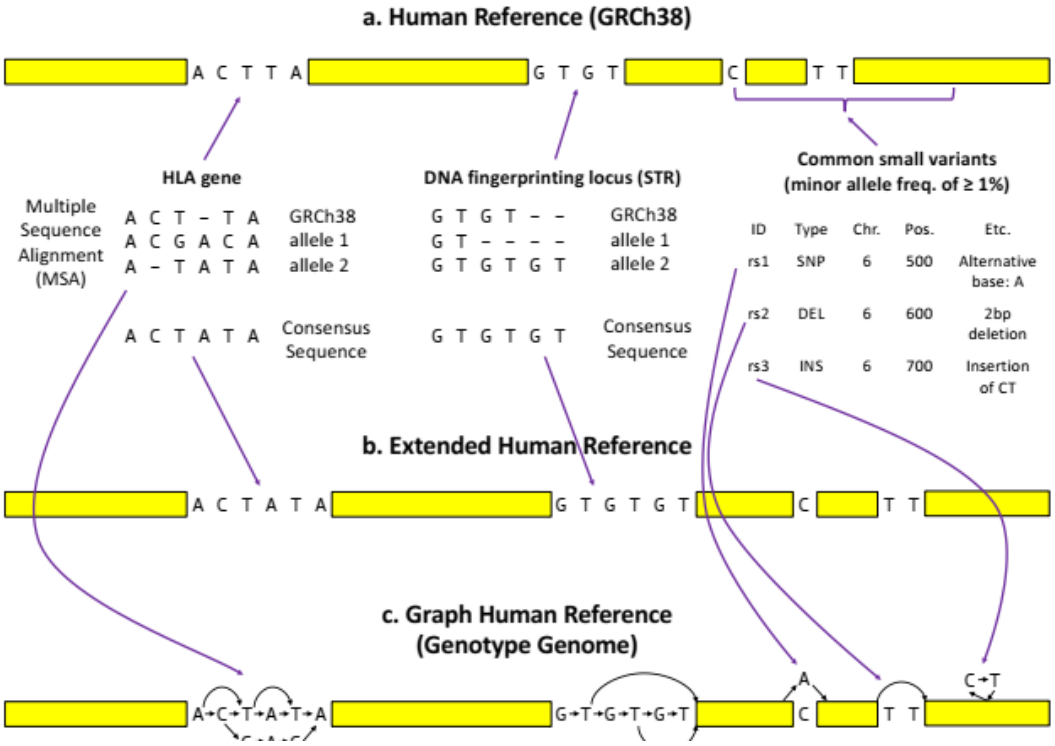


Figure 2.6: Schematic diagram depicting the construction of the HISAT-Genotype graph reference genome. (a) The linear haploid reference sequence, GRCh38, and (b) consensus sequence obtained from a multiple sequence alignment of known variants are incorporated, along with known variants, into (c) the graph human reference. Image obtained from Kim *et al.* (2018)

Read alignment and extraction are both performed through the ”hisatgenotype_extract_reads.py” script (Listing C9). For the genotyping step, HISAT-Genotype utilizes a graph representation of the reference assembly and the

genomic variants utilized in the read extraction step (Figure 2.6). HISAT-Genotype *HLA* genotyping is performed through two steps, namely alignment-based genotyping, and gene assembly (Figure 2.7). Initially, reads are aligned to exons 2 and 3 of *HLA* allele sequences obtained from the IMGT/HLA database (v. 3.31; Robinson *et al.* 2015). The alleles to which the reads align are then subset, and a second alignment occurs, this time to the genomic *HLA* sequences of the subset of alleles. As many *HLA* sequence data only consists of exons 2 and 3, the genomic sequences provide a possible path to which the reads aligning to incompletely sequences alleles can still align.

Once a set of candidate alleles has been identified, HISAT-Genotype applies an expectation-maximization model. The expectation step of the model estimates the number of reads that align to an allele, in the presence of another candidate allele. Similarly to Kourami, reads are unambiguously assigned to alleles (that is, a read cannot be aligned to both alleles being compared). This is performed iteratively. The number of reads that align to an allele, along with the alignment score, form the abundance measure, which the algorithm attempts to maximize. The results are then reported in decreasing order of abundance, at the six-digit resolution (Kim *et al.*, 2018).

The next step, *HLA* assembly, is then performed to determine the two alleles present or to genotype novel alleles. Assembly occurs though splitting reads into k-mers, which, through alignment on the graph, are assembled into an assembly graph. As sequencing and alignment errors can introduce multiple paths through the graph, the paths with fewer supporting reads than the threshold are pruned (Figure 2.7a). Phasing is resolved through the incorporation of paired-end information, as well as *de novo* alignment of reads, in which the overlap of reads provides information for contig formation. In the case of regions where the allele pairs are homozygous (Figure 2.7b), the reads are aligned to sequences from the IMGT/HLA database (Figure 2.7c), enabling the identification of downstream heterozygous regions (Kim *et al.*, 2018). Genotypes are reported at a six-digit resolution.

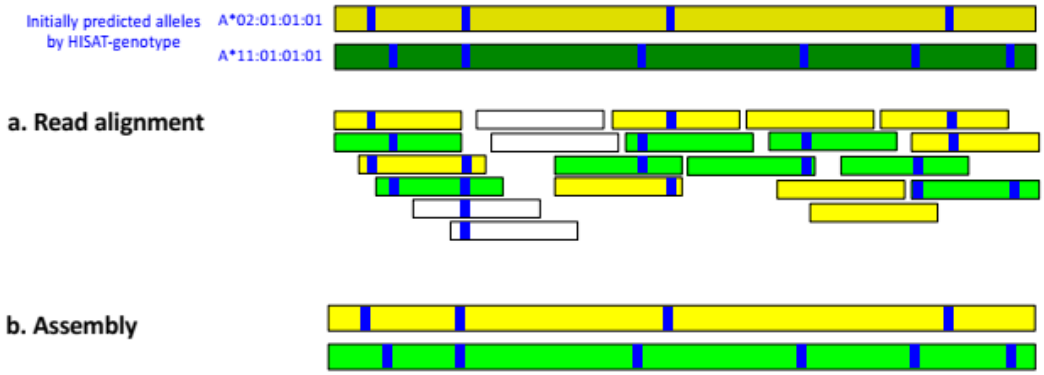


Figure 2.7: Schematic diagram depicting read alignment and assembly in HISAT-Genotype. (a) Reads are aligned to *HLA* alleles. In this example, yellow reads align to *A*02:01:01:01* and green reads align to *A*11:01:01:01*. Reads that align to both alleles are displayed in white. (b) Once two probable alleles are identified, allele assembly occurs through *de novo* alignment of the reads. Thereafter, the aligned reads are compared to the initial predicted alleles. Figure obtained from Kim *et al.* (2018)

2.5 Evaluation of *HLA* Genotyping Methods

In order to evaluate the four *HLA* genotyping tools, both the accuracy of the genotypes called (as compared to the high-resolution SBT genotypes), as well as the computational requirements of each tool were evaluated. This was performed, as the need to rapidly and accurately perform *HLA* genotyping is growing, and many researchers only have access to a standard desktop computer or laptop.

2.5.1 Computational Time and RAM Use

To compare the time required to align reads in either an alt-aware or non-alt-aware manner, the run time for each alignment protocol was measured. Thereafter, the computational time and peak RAM usage for the entire pipeline employed by each tool was measured. For Kourami, BWAkit and xHLA, this included the alignment and genotyping steps. For HISAT-Genotype, this included the read extraction, genotyping and assembly steps. This was performed using the GNU (Linux) ”/usr/bin/time” command as a prefix for the command line statement for each tool, which was evaluated using two, four, six, eight and ten threads. The run time for all 12 samples was recorded, and from this the average run time per sample was calculated.

Thereafter, the RAM usage throughout each component of the pipeline was evaluated, for each tool. This was performed to measure peak RAM use of each tool, as well as to observe RAM use over the duration of the individual tool pipelines. The peak RAM use was measured using the GNU (Linux) ”/usr/bin/time” command, as previously stated, to measure peak RAM use across the entire pipeline and all 12 samples. This was evaluated using two, four, six, eight and ten threads. To measure RAM use across the duration of the individual pipelines, One sample (NA19239) was selected to benchmark the tools, as each tool correctly genotyped this individual. Eight threads were selected, as this was suitable for all tools.This was performed using ”psrecord” (https://pypi.org/project/psrecord/), which measured the RAM use at 10 second intervals for the duration of the run.

2.5.2 Analysis of *HLA* Genotyping Accuracy

For each *HLA* genotyping tool, the accuracy of the genotyping results was analyzed in two ways. Firstly, the accuracy at which individual alleles were reported were analyzed at the two-, four- and six-digit resolution. To do this, the genotyped allele was compared to the corresponding high-resolution SBT genotyping call. Genotyping accuracy was evaluated on a binary scale where, at each resolution, if the genotypes were concordant with the reported SBT alleles, the genotype was recorded as correct. The accuracy of each tool was reported as a proportion, based on the number of correctly genotyped alleles compared to the total number of ambiguous SBT alleles present in the cohort, across all three *HLA* loci. This comparison was further performed for both the alt-aware and non-alt-aware strategies employed by Kourami.

Secondly, the accuracy of the genotyping results was analyzed based on the number of individual nucleotide positions that were incorrectly genotyped. To do this, a pairwise comparison of both the correct and incorrectly genotyped allele was performed for each individual and tool. This was performed by obtaining and aligning the cDNA sequences for all alleles. The sequences were then aligned using MUSCLE (Gap open penalty: - 400, gap extend: 0, UPGMB, minimum length of diagonal: 24). These parameters were selected, as the resulting alignment

was concordant with alignments from the IMGT/HLA database. The sequences were then trimmed to only include nucleotides from exons 2 and 3 (position 73 - 620). This was performed on a per sample, per locus, basis. The number of differences between the SBT genotype, and the incorrectly genotyped alleles, were counted separately based on whether the incorrect call was falsely classified as either heterozygous or homozygous. A false heterozygous call would result from a genotyping call, where an invariant position was incorrectly reported as variable. Alternatively, a false homozygous call would result at a variable position, where an invariant call was incorrectly reported. Thereafter, to visualize the differences between the incorrectly genotyped and SBT alleles, a dendogram was constructed using the neighbor-joining method. This was performed per locus, using MEGA (v. 7.17; Kumar *et al.* 2015). In each instance, the number of nucleotide differences were utilized as a model, and no test of phylogeny was performed.

2.6 Analysis of *HLA* Variability within the 1000 Genomes Project

2.6.1 Allele-level *HLA* Variability within the 1000 Genomes Project

A cohort of 1267 individuals (Gourraud *et al.*, 2014), from which the 12 individuals used for *HLA* genotyping had been selected, was used to evaluate how population-specific variability within the *HLA* region can potentially impact *HLA* genotyping accuracy. This was done in order to evaluate (1) how population specific *HLA* variability may impact *HLA* genotyping tools that rely on WGS data, and (2) whether biases inherent in these *HLA* genotyping methods may impact the accuracy of *HLA* genotyping results in specific populations. These individuals, derived from 14 populations, were divided into four super-populations based on geographical demography (Table 2.3). To determine the degree of allelic variability within each super-population, the number of ambiguous SBT alleles in each super-population were counted. Thereafter, the number of population-specific alleles were determined. This was performed on a per-locus basis.

2.6.2 Nucleotide-level *HLA* Variability within the 1000 Genomes Project

Thereafter, the nucleotide variability within and between super-populations was assessed and visualized. To do this, the cDNA sequences for each allele present in the cohort were obtained from the IMGT/HLA database. Once all the representative sequences were obtained and formatted into a FASTA file, the sequences were aligned using MUSCLE (Gap open penalty: - 400, gap extend: 0, UPGMB, minimum length of diagonal: 24). Thereafter, only nucleotide positions within exons 2 and 3 were included (73-620). This resulted in a MSA consisting of 546 bp, for each gene, for each super-population. The MSAs were then converted to VCF files, using MSA2VCF (Lindenbaum, 2015). Thereafter, the VCF files were imported into R using the R package SeqArray (v. 1.22.6; Zheng *et al.* 2017). From this, the number of variable positions, per locus, per super-population, were determined. Variable positions unique to each super-population were then determined. Due to the admixture in the South American population (Creanza *et al.*, 2015; Homburger *et al.*, 2015; Wang *et al.*, 2010), only positions unique to the African, Asian and European super-populations were reported. Additionally, to further visualize the variation within and between super-populations, a principal component analysis (PCA) was performed using the R package SNPRelate (Zheng *et al.*, 2012). Variants which were in complete LD (D'=1 and $r^2=\pm1$) were excluded. Finally, heatmaps of the intragenic distances, both between and within loci and super-populations, were created using superheat (v. 0.1.0 Barter and Yu 2017).

To compare whether the degree of *HLA* variability was significantly different between super-populations and individual *HLA* loci, intragenic distances were calculated by generating a pairwise-distance matrix based on pairwise differences between each allele and the GRCh38 reference sequence. Thereafter, the mean intragenic distance for each super-population was calculated and compared using an analysis of variation (ANOVA) test. This was performed both between super-populations, and between loci. ANOVA is useful for an omnibus statistical test, in which a significant result indicates that at least two of three (or more groups) differ significantly. An ANOVA test requires a dependent variable, as well

as two or more categorical independent variables. A Tukey Honest Significant Difference (HSD) test was used for *post hoc* adjustment of the ANOVA results, where appropriate. A significance threshold of $\alpha = 0.05$ was selected for both the ANOVA and Tukey HSD tests. All statistical analysis were performed within R (v. 3.5.1). As the Tukey HSD is a *post hoc* test, it was only performed when the ANOVA test was found to be significant.

Table 2.3: The number of individuals included in each population and super-population, from the 1267 individuals from the 1000 Genomes Project for whom high-resolution SBT *HLA* genotyping data were available. Data were obtained from Gourraud *et al.* (2014)

Super-population	Population[1]	Number of Individuals
African	ASW	90
	LWK	90
	YRI	90
	Total	272
Asian	CHB + JPT	181
	CHD	90
	CHS	100
	Total	376
European	CEPH	111
	FIN	100
	GBR	96
	TSI	90
	Total	396
South American	CLM	70
	MXL	89
	PUR	70
	Total	229
Grand total		1267

[1] ASW - African American. LWK - Luhya in Webuye, Kenya, YRI - Yoruban in Nigeria. CHB - Han Chinese in Beijing, China. JPT - Japanese in Tokyo, Japan. CHD - Chinese in Metropolitan Denver. CHS - Southern Han Chinese. CEPH - Utah residents with Northern and Western European Ancestry. FIN - Finnish in Finland. GBR - British in England and Scotland. TSI - Toscani in Italy. CLM - Colombians from Medellin, Colombia. MXL - Mexican Ancestry from Los Angeles - California, USA. PUR - Puerto Rican, Puerto Rico.

Chapter 3
Results

3.1 Data Preprocessing

Data from 12 individuals were obtained in BAM format in order to evaluate four *HLA* genotyping tools that utilize short-read NGS data. These individuals were selected, as both high-coverage (30X) WGS data, (The 1000 Genomes Consortium, 2015), and high-resolution SBT *HLA* data (Gourraud *et al.*, 2014) were available. The original BAM files were aligned to GRCh37, and the information they contained had to be realigned to GRCh38 in preparation for *HLA* genotyping. The number of reads that aligned to the classical *HLA* class I loci within the GRCh37-aligned BAM files ranged between approximately 500 and 1500, across all three loci (Table 3.1). Three individuals (NA19238, NA19239 and NA19240) had a higher number of aligned reads, compared to the other individuals. As the region specified during the data retrieval encompassed the *HLA* class I loci, which also included the non-classical loci, as well as decoy sequences and unmapped reads, it was also necessary to determine the total read counts, before and after preprocessing. The total number of reads ranged between 5 000 000 and 8 500 000 across individuals (Table 3.2). The number of reads that were removed during preprocessing, ranged between 55 000 and 140 000 per individual, which accounted for between 0.7 and three percent of the total reads.

When the reads were aligned to GRCh38, two strategies were used. Firstly, reads were aligned to the linear haploid GRCh38 reference sequence, then secondly, to GRCh38, with alternate loci and allele sequences obtained from the IMGT/HLA database. The number of reads that aligned to the classical *HLA* class I loci through the non-alt-aware alignment ranged between 500 and 1200 across all three loci (Table 3.3). The alt-aware alignment strategy had a similar number of reads

which aligned to the GRCh38 reference sequence, with a greater number of reads aligning to *HLA* sequences, and alternate loci. The number of reads that aligned to *HLA* sequences ranged between 13 000 and 64 000. Within *HLA-B* and *HLA-C*, the number of reads that aligned to alternate loci ranged between 750 and 1700. The number of reads that aligned to alternate loci within *HLA-A* ranged between 15 000 and 56 000.

Table 3.1: The number of reads aligning to each of the three classical *HLA* class I genes in the original GRCh37-aligned BAM files obtained for each of the 12 individuals for whom both SBT and WGS data were available

	HLA-A	*HLA-B*	*HLA-C*
HG00096	617	568	571
HG00268	730	630	607
HG00419	605	613	602
HG01051	602	583	545
HG01112	500	546	544
NA18939	860	666	693
NA19238	1055	803	852
NA19239	1202	842	734
NA19240	1577	817	864
NA19625	520	483	448
NA19648	565	537	475
NA20502	746	560	503

HLA-A Chr6: 29 910 247 - 29 913 661
HLA-B Chr6: 31 321 649 - 31 324 989
HLA-C Chr6: 31 236 526 - 31 239 913

Table 3.2: The total number of reads in the original GRCh37-aligned BAM files, including reads aligned to the *HLA* class I region, decoy sequences and unmapped reads. Counts are provided both before and after performing sanitization with RevertSam

	Number of Reads		
	Original	Removed	Final
HG00096	6 124 694	55 072	6 069 62
HG00268	6 474 278	64 292	6 409 986
HG00419	5 585 369	86 907	5 498 462
HG01051	5 468 179	60 345	5 407 834
HG01112	6 055 236	40 642	6 014 594
NA18939	6 206 023	62 163	6 143 860
NA19238	8 384 863	63 831	8 321 032
NA19239	8 464 120	64 012	8 400 108
NA19240	8 511 349	70 521	8 440 828
NA19625	4 661 980	106 026	4 555 954
NA19648	4 756 252	140 362	4 896 614
NA20502	5 069 546	114 146	4 955 400

Table 3.3: The number of reads, from both the alt-aware and non-alt-aware alignment protocols provided by BWA-mem, aligning to each of the three classical *HLA* class I genes in GRCh38. The alt-aware alignment counts include reads aligning to *HLA* allele sequences, and alternate loci. Counts were obtained from the GRCh38-aligned BAM files for each of the 12 individuals for whom both SBT and WGS data were available

	HLA-A			*HLA-B*			*HLA-C*			GRCh38 Alt non-aware		
	GRCh38	*HLA*	Alt-loci	GRCh38	*HLA*	Alt-loci	GRCh38	*HLA*	Alt-loci	*HLA-A*	*HLA-B*	*HLA-C*
HG00096	371	15 170	37 235	321	18 383	1557	385	18 330	1344	612	572	569
HG00268	444	19 316	38 166	429	21 962	1281	400	16 795	1233	731	637	602
HG00419	234	17 482	18 739	456	24 136	1400	169	17 259	1000	601	617	598
HG01051	277	19 011	17 651	345	20 124	1224	245	22 381	1474	596	575	498
HG01112	232	18 287	35 795	327	21 434	1458	196	21 312	1716	500	549	544
NA18939	566	27 308	38 028	394	24 756	1472	483	20 821	1379	855	677	693
NA19238	833	32 696	49 556	374	34 727	1833	320	26 374	1548	1050	811	849
NA19239	539	48 388	53 361	355	33 133	1599	206	23 704	1663	1192	853	732
NA19240	1141	64 030	56 385	405	37 383	1711	349	28 025	1714	1567	821	877
NA19625	254	17 672	15 267	397	20 628	1175	332	16 801	1160	519	477	445
NA19648	513	15 164	26 678	420	20 628	932	322	15 453	791	551	533	466
NA20502	480	24 827	31 855	403	22 310	986	323	13 536	749	738	557	500

Specific locations used to obtain reads counts available in Supplementary Table A3

In addition to calculating read counts, the depth of coverage across individual loci were calculated for each of the 12 individuals (Figure B1). Within *HLA-A*, the read depth of samples HG01112 and NA19625 were below an average of 30X (Figure B1). NA19239 also contained two regions of decreased read depth across this gene, however, these were located within intronic regions. Conversely, NA19239 and NA19240 both had an average read depth of above 75X across the same region. Within *HLA-B*, the read depths of HG01051 and NA19625 were found to be below average, while all other samples had a read depth of approximately 30X. Within *HLA-C*, HG01051, HG01112, NA19625 and NA19648 had read depths below an average of 30X (Figure B1). The other samples had read depth above 30X.

The SBT *HLA* genotyping data for these individuals were obtained from Gourraud *et al.* (2014) and were converted to ambiguous allele format (Table 3.4). This was done in order to facilitate comparisons between the four *HLA* genotyping tools, as Kourami only reports genotypes in an ambiguous format. The majority of the alleles within the dataset could be assigned ambiguity codes, however, three alleles (*HLA-A*36:01, HLA-B*38:01:01* and *HLA-B*67:01:01*) had unique nucleotide sequences across exons 2 and 3. For these 12 individuals, the list of possible *HLA* alleles corresponding to each ambiguity code was obtained from IMGT/HLA database, and is reported in Supplementary tables B3-B5.

Table 3.4: High-resolution SBT *HLA* genotyping data for the 12 individuals for whom WGS were also available, following conversion to their relevant ambiguity codes. Original SBT genotyping data were obtained from Gourraud *et al.* (2014)

Sample	Population	*HLA-A*	*HLA-B*	*HLA-C*
HG00096	British in England	*A*01:01:01G*	*B*08:01:01G*	*C*07:01:01G*
	and Scotland	*A*29:02:01G*	*B*44:03:01G*	*C*16:01:01G*
HG00268	Finnish in Finland	*A*03:01:01G*	*B*07:02:01G*	*C*07:02:01G*
		*A*25:01:01G*	*B*18:01:01G*	*C*12:03:01G*
HG00419	Southern Han	*A*02:06:01G*	*B*13:01:01G*	*C*03:04:04*
	Chinese	*A*24:02:01G*	*B*40:01:01G*	*C*03:04:01G*
HG01051	Puerto Ricans from	*A*02:02:01G*	*B*15:16:01G*	*C*12:03:01G*
	Puerto Rico	*A*24:02:01G*	*B*35:03:01G*	*C*14:02:01G*
HG01112	Colombians from	*A*02:01:01G*	*B*38:01:01*	*C*05:01:01G*
	Medellin	*A*26:01:01G*	*B*44:02:01G*	*C*12:03:01G*
NA18939	Japanese in Tokyo,	*A*11:01:01G*	*B*27:04:01G*	*C*07:02:01G*
	Japan	*A*31:01:02G*	*B*67:01:01*	*C*12:02:01G*
NA19238	Yoruba in Ibadan,	*A*30:01:01G*	*B*53:01:01G*	*C*04:01:01G*
	Nigeria	*A*36:01*	*B*57:03:01G*	*C*18:01:01G*
NA19239	Yoruba in Ibadan,	*A*02:01:01G*	*B*35:01:01G*	*C*04:01:01G*
	Nigeria	*A*68:02:01G*	*B*52:01:02G*	*C*16:01:01G*
NA19240	Yoruba in Ibadan,	*A*30:01:01G*	*B*35:01:01G*	*C*04:01:01G*
	Nigeria	*A*68:02:01G*	*B*57:03:01G*	*C*18:01:01G*
NA19625	Yoruba in Ibadan,	*A*02:01:01G*	*B*07:02:01G*	*C*07:01:01G*
	Nigeria	*A*23:01:01G*	*B*44:03:02G*	*C*12:03:01G*
NA19648	Mexican Ancestry,	*A*03:01:01G*	*B*07:02:01G*	*C*01:02:01G*
	Los Angeles, USA	*A*11:01:01G*	*B*51:01:01G*	*C*07:02:01G*
NA20502	Toscani in Italia	*A*01:01:01G*	*B*07:02:01G*	*C*04:01:01G*
		*A*31:01:02G*	*B*35:02:01G*	*C*07:02:01G*

Ambiguous allele codes are shown with the suffix "G".

3.2 Computational Time and Memory Use

In order to compare the differences in time required to align reads in an alt-aware and non-alt-aware manner, the reads were first aligned to GRCh38.p7 (non-alt-aware), and then GRCh38.p7 with alternate loci and *HLA* sequences obtained from the IMGT/HLA database (alt-aware); Figure 3.1). It was found that aligning reads in an alt-aware manner was slightly slower than simply aligning reads to the haploid reference sequence. This was the case at all of the threads which were specified.

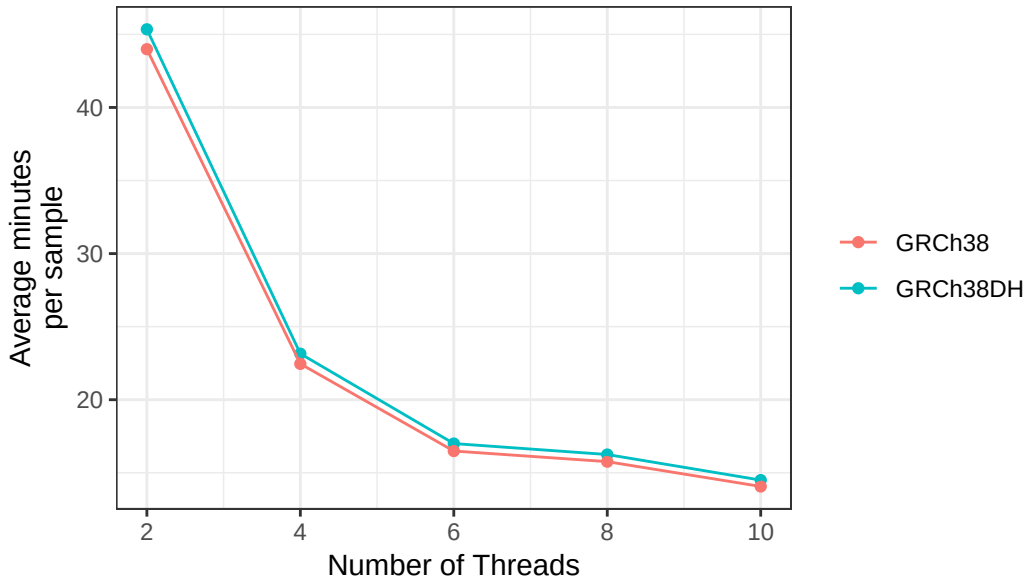


Figure 3.1: Average run time per thread utilized when aligning reads to GRCh38 with (GRCh38DH) and without (GRCh38) the inclusion of alternate loci and sequences obtained from the IMGT/HLA database (Robinson *et al.*, 2015)

The run times were also recorded for the full pipeline, which consisted of read alignment, *HLA*-specific read extraction, and the genotyping step (Figure 3.2). HISAT-Genotype performance was least affected when different number of threads were specified. When two threads were specified, BWAki and xHLA had similar run times, however, as more threads were made available, the decrease in run time of xHLA was more substantial than that of BWAkit. HISAT-Genotype and Kourami also had similar run times when two threads were specified, however, Kourami had the shortest run times thereafter. For all four tools, run times decreased as more threads were made available, however the improvement in performance plateaued at around six threads.

Similarly, differences in peak RAM usage between the four tools, across different numbers of threads, were compared (Figure 3.3). The peak RAM usage was measured across the whole pipeline for each individual tool. The peak RAM usage by HISAT-Genotype was consistent when measured across the different number of threads. Conversely, the other three tools followed a similar pattern, displaying higher peak RAM usage as more threads were made available.

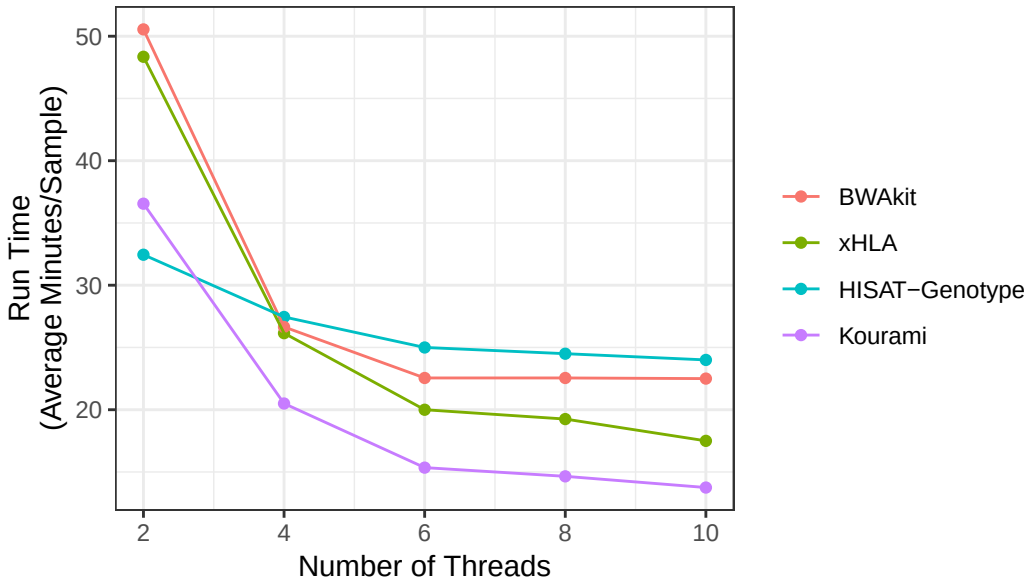


Figure 3.2: Average run time per thread utilized by BWAkit, xHLA, Kourami and HISAT-Genotype. The run time for the alignment step (performed separately using BWA-mem) is included for Kourami and xHLA

As the tools utilize different amounts of RAM throughout their respective pipelines, the RAM usage was also measured across the duration of the pipeline (Figure 3.4). For this, genotyping was performed on one individual (NA19239). HISAT-Genotype demonstrated the highest RAM usage, which occurred in the initial steps of the pipeline. This high RAM usage was seen until approximately 19 minutes, where a subsequent drop in RAM usage could be seen. Both Kourami and BWAkit initially utilized the same amount of RAM. Thereafter, however, Kourami, demonstrated a drop in RAM usage at approximately 13 minutes, which was not seen in BWAkit. A similar drop in RAM usage was seen for xHLA. These drops corresponded with the end of the alignment step for HISAT-Genotype, Kourami and xHLA.

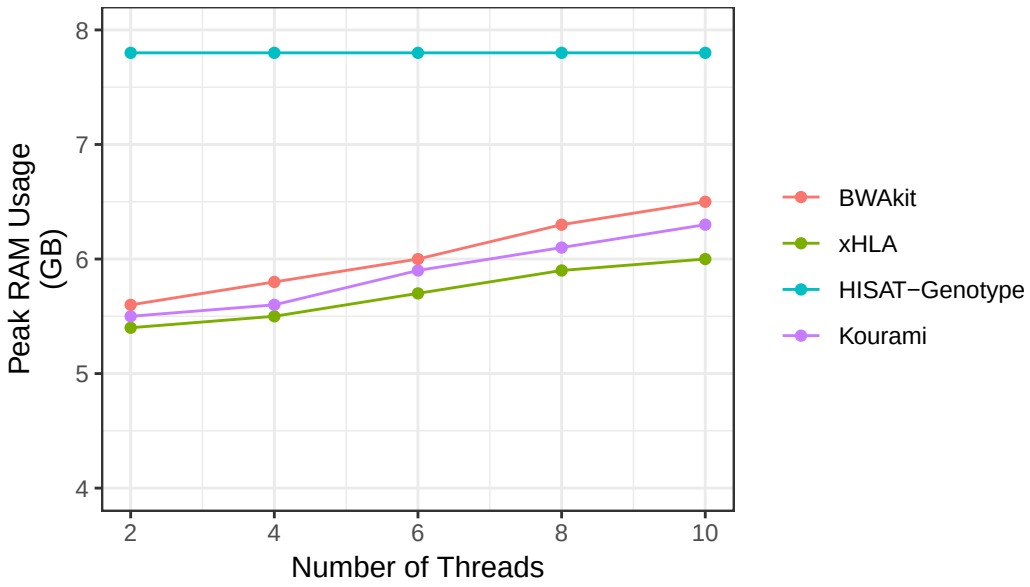


Figure 3.3: Peak RAM usage per thread utilized by BWAkit, HISAT-Genotype, Kourami and xHLA. The peak RAM usage for the alignment step (performed separately using BWA-mem) is included for Kourami and xHLA

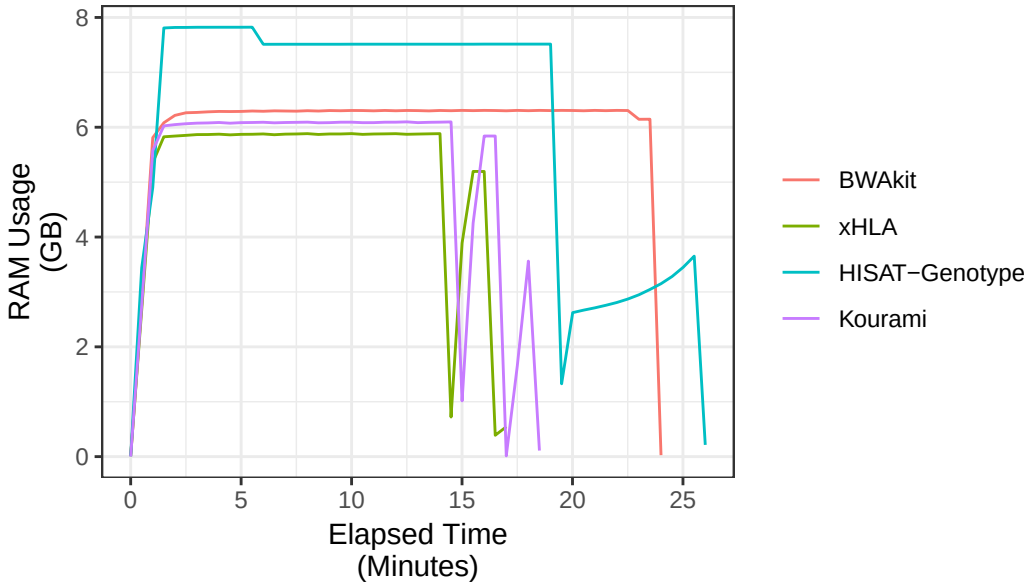


Figure 3.4: RAM usage over time for the four tools: BWAkit, xHLA, Kourami and HISAT-Genotype. RAM usage was measured across the entire pipeline, using eight threads for one individual (NA19239)

3.3 Evaluation of *HLA* Genotyping Accuracy

3.3.1 Allele-level *HLA* Genotyping Accuracy

In order to evaluate the accuracy of the four *HLA* genotyping tools, BWAkit, xHLA, Kourami, and HISAT-Genotype, at the allele level, the genotypes reported for the 12 individuals were compared to the corresponding high-resolution SBT genotyping data. The genotypes reported by BWAkit, xHLA and HISAT-Genotype were converted to their corresponding ambiguity codes, to facilitate comparisons with Kourami. The accuracy of the tools was reported at the two-, four- and six-digit resolution (Figure 3.5 and Tables B7, B8 and B9). Overall, BWAkit was found to be the most inaccurate tool, across all resolutions. Furthermore, the accuracy of this tool decreased as genotyping resolution increased. BWAkit was not able to assign genotypes to two individuals (HG01051 and NA19625) within *HLA-B* (Table 3.5). Furthermore, BWAkit incorrectly genotyped nine alleles at the two-digit resolution, 14 alleles at the four-digit resolution, and a further four alleles at the six-digit resolution.

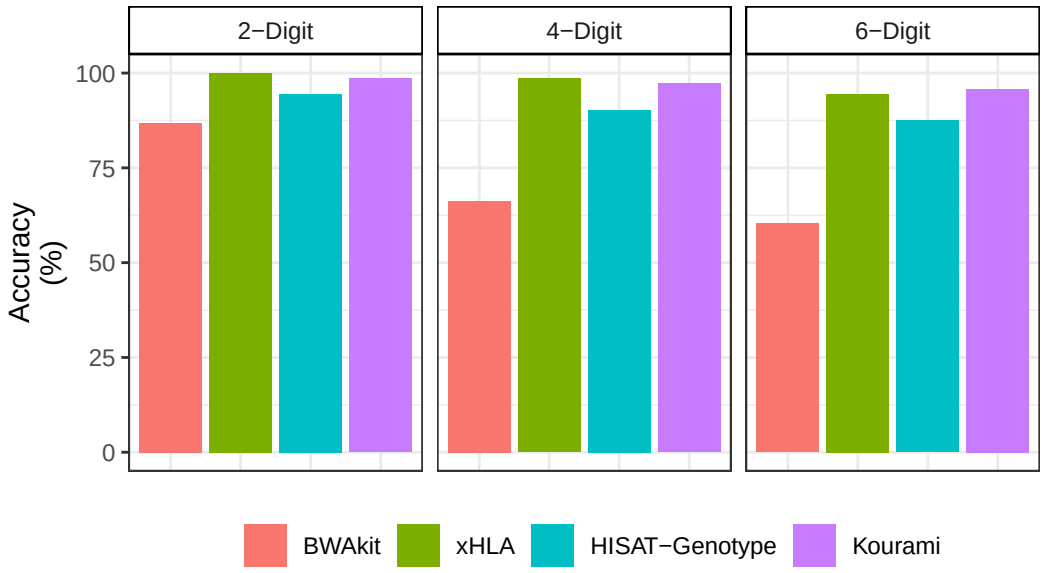


Figure 3.5: Accuracy of *HLA* Genotyping by the four tools: BWAkit, xHLA, Kourami and HISAT-Genotype, at the two-, four- and six-digit resolution

At the two-digit resolution, xHLA was the most accurate, correctly genotyping every allele (Figure 3.5). Kourami displayed 98.6 percent accuracy, while

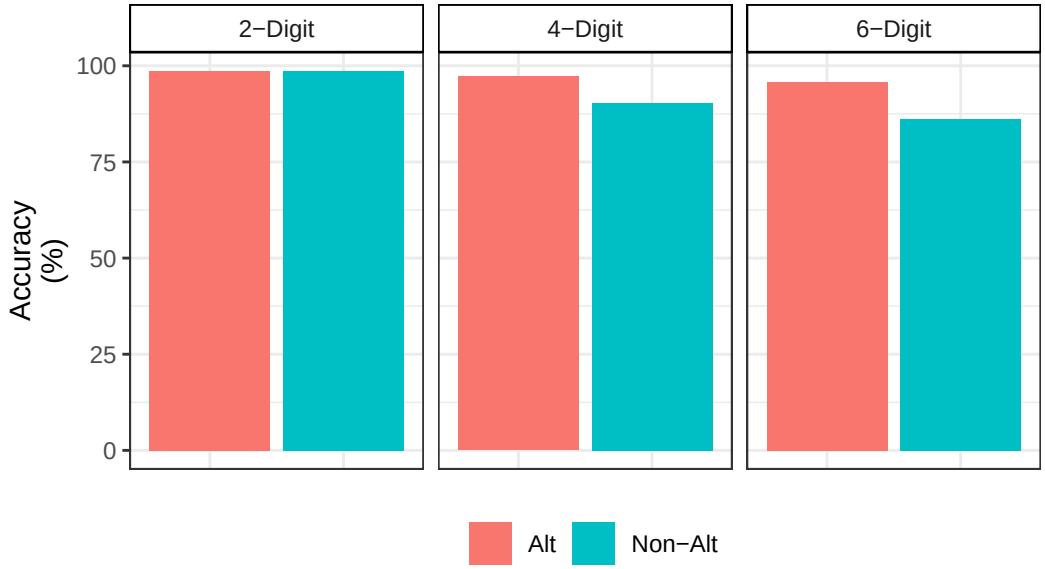


Figure 3.6: Genotyping accuracy of Kourami comparing the effects of inclusion and exclusion of alternate loci and sequences obtained from the IMGT/HLA database (Robinson *et al.*, 2015)

HISAT-Genotype was found to be 94.4 percent accurate. A similar trend was seen at the four-digit resolution, where xHLA was the most accurate (98.6 percent), followed by Kourami (97.2 percent) and HISAT-Genotype (90.3 percent). At the six-digit resolution, however, Kourami was found to be the most accurate (95.8 percent), followed by xHLA (94.4 percent) and HISAT-Genotype (87.5 percent). As Kourami is capable of utilizing alt-aligned and non-alt-aligned inputs, both approaches were tested. At the two-digit resolution there was no difference in the overall accuracy of the two approaches (Figure 3.6), however, at the four- and six-digit resolutions, the alt-aware approach resulted in increased accuracy.

With respect to individual alleles assigned by these tools (Table 3.5), xHLA incorrectly assigned two alleles at the four-digit resolution, and a further three at the six-digit resolution. All the incorrectly assigned alleles occurred within *HLA-C*. Kourami incorrectly assigned one allele at the two-digit resolution, one at the four-digit resolution, and a further one at the six-digit resolution. Two of these incorrectly assigned alleles occurred in one individual (HG01051) within *HLA-C*. HISAT-Genotype incorrectly assigned four alleles at the two-digit resolution, seven at the four-digit resolution, and nine at the six-digit resolution. The

incorrectly assigned alleles occurred across all three loci, but these included incorrect genotypes for two individuals homozygous at the *HLA-C* locus.

Table 3.5: Incorrectly assigned alleles reported by the four tools: BWAkit, xHLA, Kourami and HISAT-Genotype. Only genotypes not concordant with high-resolution SBT genotyping data are reported

Sample	Tool	*HLA-A*	*HLA-B*	*HLA-C*
HG00096	SBT	*01:01:01G/29:02:01G*	08:01:01G/44:03:01G	*07:01:01G/16:01:01G*
	BWAkit	*1:01:05/31:21*	08:50/+[3]	*+/16:21*
	xHLA			*07:01:05/+*
	Kourami (L)[1]	*+/29:18*		
HG00268	SBT	*03:01:01G/25:01:01G*	07:02:01G/18:01:01G	*07:02:01G/12:03:01G*
	BWAkit	*03:12/+*	07:07/+	*07:31:01/+*
	Kourami (L)			*+/12:03:16*
	HISAT-Genotype		07:50/+	
HG00419	SBT	*02:06:01G/24:02:01G*	13:01:01G/40:01:01G	*03:04:01G/03:04:04*
	BWAkit		+/40:49	*+/03:04:01G*
	xHLA			*+/03:04:04*
HG01051	SBT	*02:02:01G/24:02:01G*	15:16:01G/35:03:01G	*12:03:01G/14:02:01G*
	BWAkit	*02:05:04/02:05:04*	NULL/NULL[4]	*12:03:04/12:03:04*
	xHLA			*12:02:01G/+*
	Kourami (L)	*+/24:229*		
	Kourami (A)[2]			*12:03:20/12:55*
	HISAT-Genotype	*24:03:01G/24:10:01*	+/35:03:19	*+/12:03:01G*
HG01112	SBT	*02:01:01G/26:01:01G*	38:01:01/44:02:01G	*05:01:01G/12:03:01G*
	BWAkit	*02:55/02:55*		*05:52/08:25*
NA18939	SBT	*11:01:01G/31:01:02G*	27:04:01G/67:01:01	*07:02:01G/12:03:01G*
	BWAkit			*+/07:314*
	xHLA			*07:02:04/+*
	Kourami (L)			*+/12:14:01*
NA19238	SBT	*303:01:01G/36:01:00*	53:01:01G/57:03:01G	*04:01:01G/18:01:01G*
	BWAkit			*+/18:03*
NA19240	SBT	*30:01:01G/68:02:01G*	53:01:01G/57:01:01G	*04:01:01G/18:01:01G*
	BWAkit			*+/18:08*
NA19625	SBT	*02:01:01G//23:01:01G*	07:02:01G/44:03:02G	*07:01:01G/12:03:01G*
	BWAkit	*23:04/23:04*	NULL/NULL	*16:02:11/16:02:11*
	Kourami (L)	*02:571/+*		*+/12:03:16*
	Kourami (A)	*02:571/+*		
	HISAT-Genotype	*23:01:01G/+*	07:36/44:03:33	*+/07:01:01G*
NA19648	SBT	*03:01:01G/11:01:01G*	07:02:01G/51:01:01G	*01:02:01G/07:02:01G*
	BWAkit	*11:01:01G/+*		*+/07:29*
	Kourami (L)	*11:50Q/11:14*		
NA20502	SBT	*01:01:01G/31:01:02G*	07:02:01G/35:02:01G	*04:01:01G/07:02:01G*
	BWAkit	*01:143/+*		
	Kourami (L)	*+/31:21*		*+/07:02:12*

[1] Kourami (L) - Linear alignment without alternate loci.
[2] Kourami (A) - Alt-aware alignment including alternate haplotypes and IMGT/HLA sequences.
[3] + indicates the same allele as SBT.
[4] NULL - No allele called.
Sample NA19239 was correctly genotyped by all tools

3.3.2 SNP-level Genotyping Accuracy

As alleles which group together at the four-digit resolution can have differences in their nucleotide sequences, the SNP-level accuracy of the assigned alleles and SBT

alleles were compared (Supplementary Figures: B5 - B32). This is due to the fact that synonymous nucleotide changes affect allele assignment at the six-digit resolution, and non-synonymous changes affect genotyping at the four-digit (and possibly the two-digit) resolution. A MSA of the incorrectly assigned allele and the corresponding SBT allele sequences was created per locus, for each individual. Thereafter, the effect of the incorrect nucleotide on the accuracy of the relevant allele call was determined. (Table3.6). It was found that BWAkit produced the most SNP-level errors, followed by HISAT-Genotype and Kourami. xHLA produced the least number of errors. Furthermore, the majority of SNP-level errors produced by the graph-based tools (Kourami and HISAT-Genotype) were due to the tool falsely assigning a homozygous nucleotide at positions that were in fact heterozygous (Table 3.6).

To visualize the nucleotide differences between these alleles, a Neighbor-Joining dendogram was constructed. Within *HLA-A* (Figure 3.7), xHLA correctly assigned each allele. It was further observed that the differences between the sequences of the alleles incorrectly assigned by Kourami were more similar to the correct SBT alleles, than the alleles incorrectly assigned by the other two tools (BWAkit and HISAT-Genotype). This was especially true when the alt-aware method was used by Kourami. Within *HLA-B* (Figure 3.8), both Kourami and xHLA correctly assigned each allele. Again, alleles assigned by both BWAkit and HISAT-Genotype were similar to the SBT alleles, however, there were numerous nucleotide differences between the alleles. Within *HLA-C* (Figure 3.9)), all four tools incorrectly assigned alleles, however, the incorrect alleles assigned by xHLA were closest in sequence to the SBT alleles.

Table 3.6: Comparison of number and type of nucleotide variants between incorrect *HLA* allele and SBT allele sequences from 12 individuals from the 1000 Genomes Project

Tool	Incorrect Allele Calls (n = 72)	Incorrect Base Calls	Error Result		Error Type	
			Synonymous	Non-synonymous	Heterozygous	Homozygous
BWAkit	27	57	8	49	12	45
xHLA	4	5	3	2	2	3
Kourami	3	16	3	13	-	16
HISAT-Genotype	9	39	9	30	3	36

Number of incorrect allele calls at a six-digit resolution

Data summarized from MSA of incorrect alleles, and SBT alleles.

MSA data located in Supplementary Figures: B5 - B32

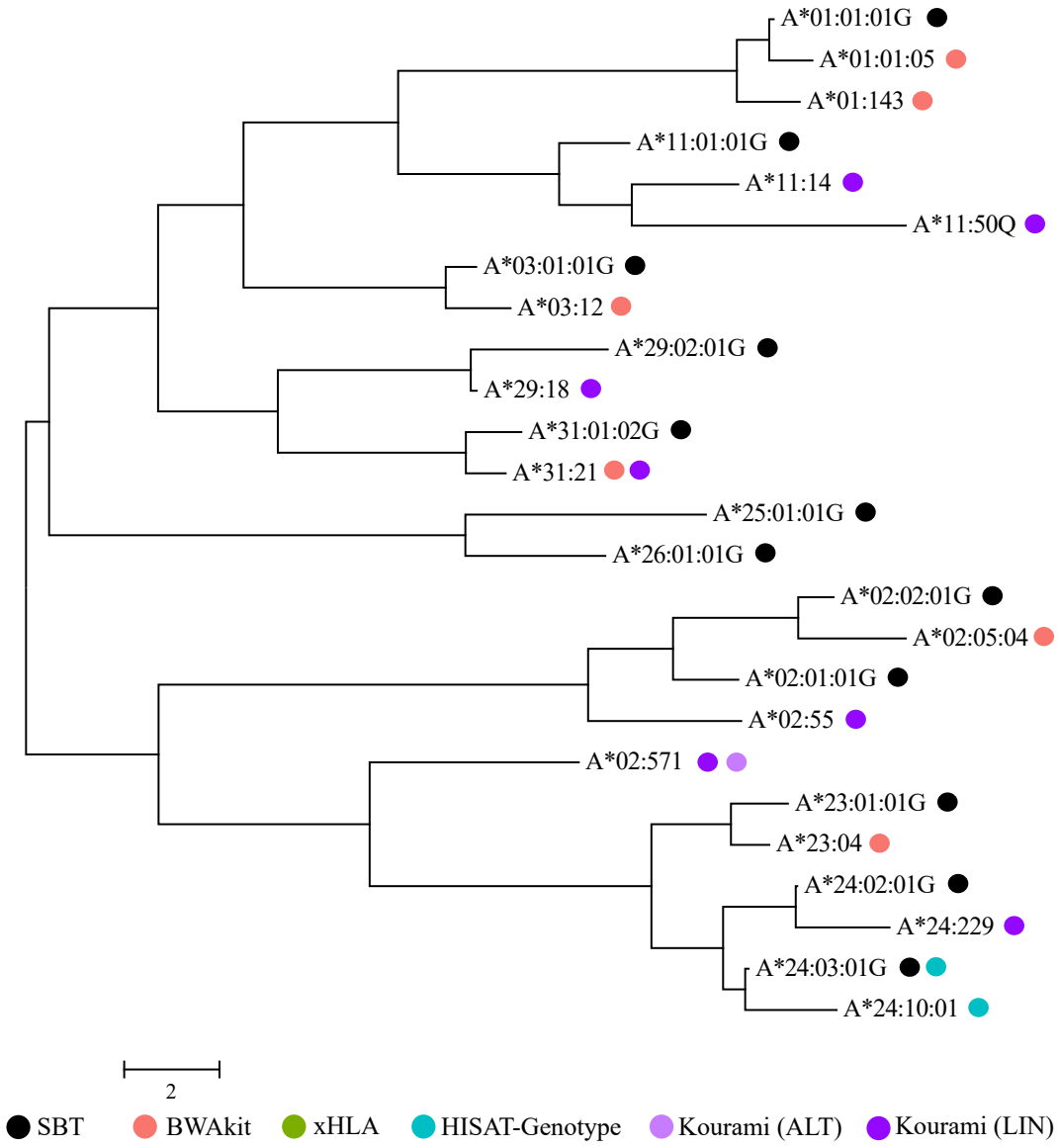


Figure 3.7: Dendogram constructed using the Neighbor-Joining method, depicting the relationship between the *HLA-A* alleles incorrectly assigned by the four genotyping methods and the alleles genotyped by SBT. Branch lengths are indicative of the number of nucleotide differences between alleles across exons 2 and 3. Dendogram produced using MEGA (v. 7.0.26; Kumar *et al.* 2015)

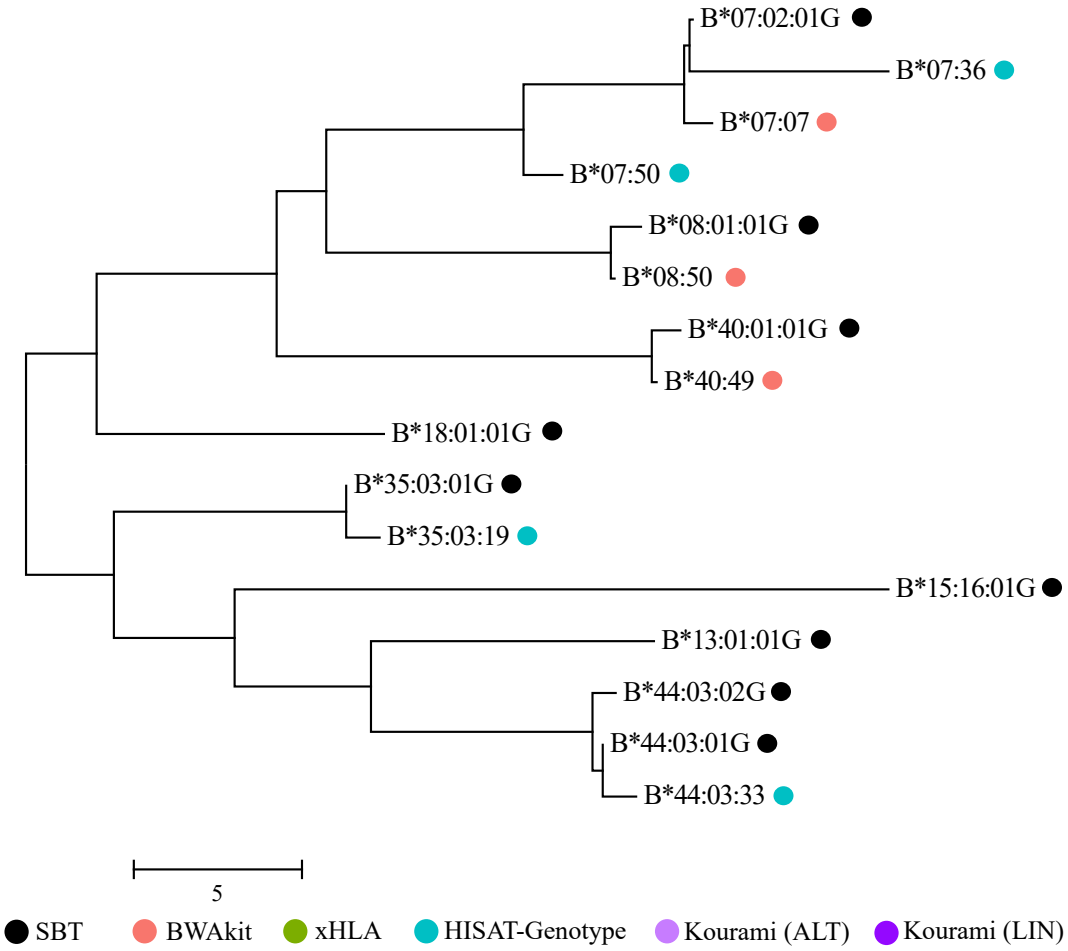


Figure 3.8: Dendogram constructed using the Neighbor-Joining method, depicting the relationship between the *HLA-B* alleles incorrectly assigned by the four genotyping methods and the alleles genotyped by SBT. Branch lengths are indicative of the number of nucleotide differences between alleles across exons 2 and 3. Dendogram produced using MEGA (v. 7.0.26; Kumar *et al.* 2015)

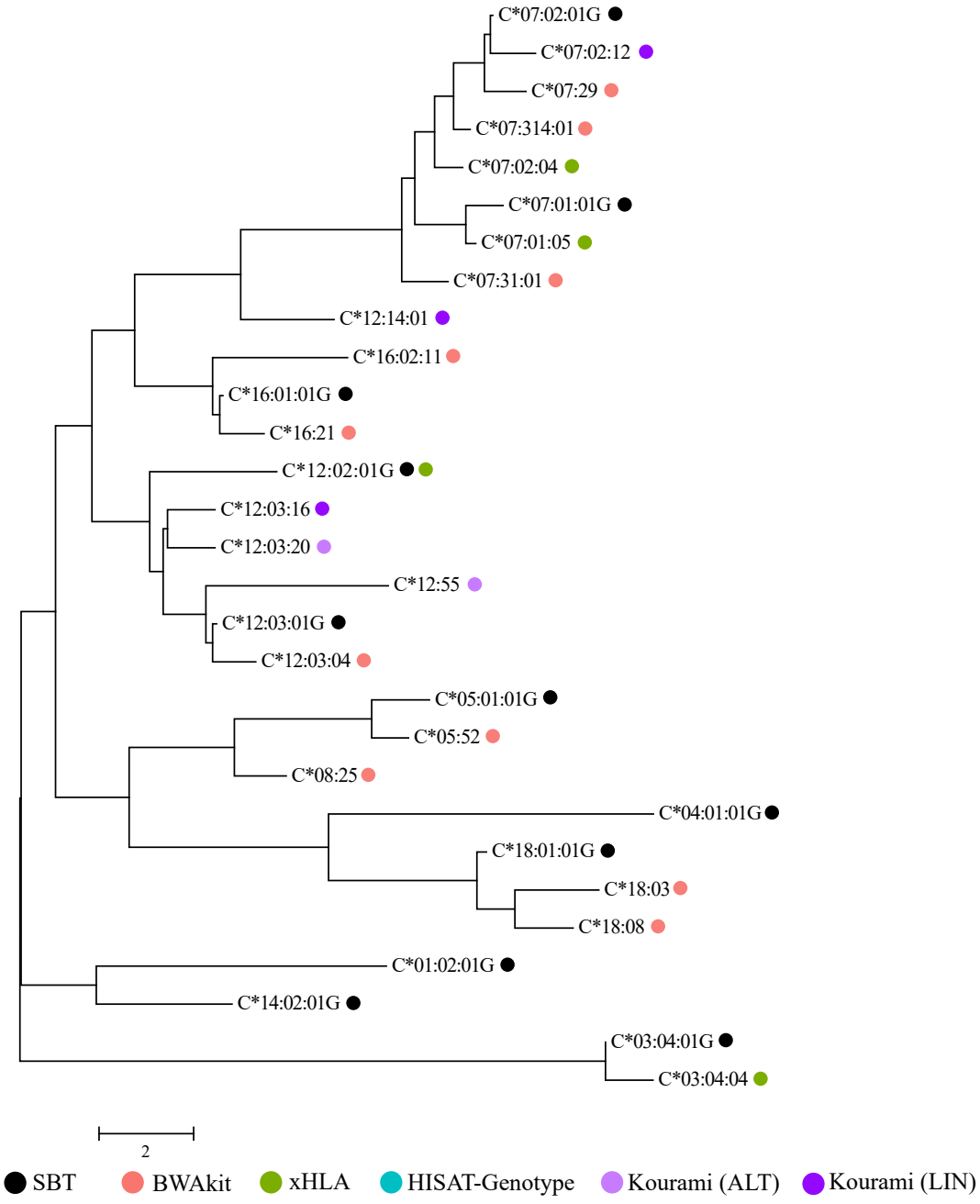


Figure 3.9: Dendogram constructed using the Neighbor-Joining method, depicting the relationship between the *HLA-C* alleles incorrectly assigned by the four genotyping methods and the alleles genotyped by SBT. Branch lengths are indicative of the number of nucleotide differences between alleles across exons 2 and 3. Dendogram produced using MEGA (v. 7.0.26; Kumar *et al.* 2015)

3.4 Analysis of *HLA* Variability within the 1000 Genomes Project

3.4.1 Allele-level Variability in the 1000 Genomes Project

HLA variability within 1267 individuals with high-resolution SBT *HLA* genotype data was examined in order to determine whether variability within populations can have an affect on *HLA* genotyping accuracy. This larger cohort was divided into four super-populations, dependent on reported ancestry and the ambiguous allele counts were reported (Table 3.7). A total of 219 alleles were present in the dataset, of which, the majority of were from *HLA-B*. With regards to *HLA-A*, all four super-populations had a similar number of variable, and unique alleles. Within *HLA-B*, the Asian and South American populations had the highest allelic variability, while the African population had the most unique *HLA-C* alleles.

Table 3.7: *HLA* allele counts for the 1267 individuals for whom high-resolution *HLA* SBT genotype were available. Both the total number and number of unique alleles per super-population are reported. Data was obtained from Gourraud *et al.* (2014)

Super-population	Population	Number of Individuals	Number of Alleles		
			A	B	C
African	ASW	90	25	38	24
	LWK	90	24	29	21
	YRI	90	23	28	16
	Total	272	34 (6)	55 (11)	28 (3)
Asian	CHB + JPT	181	25	41	20
	CHD	90	17	34	18
	CHS	100	21	39	23
	Total	376	32 (7)	51 (18)	29 (1)
European	CEPH	111	20	30	18
	FIN	100	16	23	15
	GBR	96	18	28	19
	TSI	90	25	35	22
	Total	396	32 (6)	48 (6)	25 (1)
South American	CLM	70	29	40	18
	MXL	89	22	47	21
	PUR	70	28	41	23
	Total	229	36 (7)	70 (18)	27 (1)
	Grand total	1267	64	112	43

ASW - African American. LWK - Luhya in Webuye, Kenya, YRI - Yoruban in Nigeria. CHB - Han Chinese in Beijing, China. JPT - ASW - African American. LWK - Luhya in Webuye, Kenya, YRI - Yoruban in Nigeria. CHB - Han Chinese in Beijing, China. JPT - Japanese in Tokyo, Japan. CHD - Chinese in Metropolitan Denver. CHS - Southern Han Chinese. CEPH - Utah residents with Northern and Western European Ancestry. FIN - Finnish in Finland. GBR - British in England and Scotland. TSI - Toscani in Italy. CLM - Colombian from Medellin, Colombia. MXL - Mexican Ancestry from Los Angeles - California, USA. PUR - Puerto Rican, Puerto Rico.

3.4.2 Nucleotide-level Variability in the 1000 Genomes Project

In order to evaluate the *HLA* nucleotide variability from within The 1000 Genomes Project, the number of alternate nucleotides per variable position were counted (Table 3.8). Within *HLA-A*, each super-population possessed a similar number of variable sites, however, the European super-population possessed more unique variable sites than the other super-populations. Within *HLA-B*, all four

super-populations again had similar numbers of variable sites. This locus also had the highest degree of variation, when compared to the other loci. In this instance, the Asian population possessed the most unique variable sites. *HLA-C* contained the least number of variable sites out of the three loci. Again, the Asian super-population contained the highest number of unique variable sites, in contrast with the African and South American super-populations, who did not possess any unique variable sites. This is represented visually in figures 3.10 - 3.12). From this, it was seen that the variable positions were similar across all four super-populations. Also, the degree of variability per site was consistent across super-populations.

Table 3.8: Number of variable and population-specific variable nucleotide sites within *HLA-A*, *HLA-B* and *HLA-C* from with the 1000 Genomes Project

Super-population	Number of Variable Sites (n = 546)			Number of Unique Variable Sites		
	HLA-A	*HLA-B*	*HLA-C*	*HLA-A*	*HLA-B*	*HLA-C*
African	75	87	54	1	3	-
Asian	72	85	56	3	6	5
European	75	83	53	5	1	1
South American	77	86	55	3	-	-

A PCA was performed to further analyze the *HLA* class I variation between super-populations. Before LD pruning, there were 63 variants within *HLA-A*, 67 within *HLA-B*, and 49 within *HLA-C*. After pruning, there were 61 within *HLA-A* and *HLA-B*, and no variants were found to be in complete LD within *HLA-C*. The PCA plots demonstrated that the the alleles did not cluster within super-populations across all three loci evaluated (*HLA-A*, *HLA-B* and *HLA-C*) (Figure 3.13). Furthermore, while PC1 accounted for above 15 percent of variability across all three loci (Figure 3.13), the correlation to variation was small (-0.2<PC1<0.1 and -0.1<PC2<0.25).

The pairwise differences between alleles were determined at each variable position, for all three loci, in order to calculate the intragenic distances (Figure 3.14). From this, the mean intragenic distance across each the four super-populations was calculated, and an ANOVA test was performed. The mean

intragenic distances were not found to differ significantly between super-populations ($p > 0.05$), at any of the three loci (Figure 3.14). However, it was seen that the mean intragenic distance of each super-population within *HLA-B* was greater than those observed at *HLA-A* and *HLA-C*. When the mean intragenic distances between loci were compared, the ANOVA results were found to be significant. A Tukey HSD test was performed to determine which loci were significantly different. It was found that each allele was significantly different from each other (Figure 3.18).

To further investigate *HLA* variability, both within and between loci and super-populations, the intragenic distance between alleles was visualized (figures B33 - B44). Within *HLA-A*, it can be seen that the African and European super-population contained smaller clusters of similar alleles (Figure B33 & B35), than the Asian and South American super-populations (Figure B34 & B36). While there were slight differences between the super-populations, when all of the alleles were grouped, a similar heatmap was formed (Figure 3.15. This indicates that certain *HLA-A* alleles, such as *HLA-A*02*, are common to all four super-populations.

Within *HLA-B*, less clustering is observed as compared to *HLA-A*. Once again, there was evidence of clustering in the African and European super-populations (Figure B37 & B39), which was less pronounced in the Asian and South American populations (Figure B38 & B40). This was possibly due to the greater number of alleles present in the Asian and South American populations. When all of the alleles were grouped, again, a similar pattern emerged, indicating the presence of common alleles between super-populations (Figure 3.16). Within *HLA-C* however, there was a similar pattern of clustering across all four super-populations (Figure 3.17), possibly due to the low number of alleles present in each super-population (Figure B41, B42, B43 & B44).

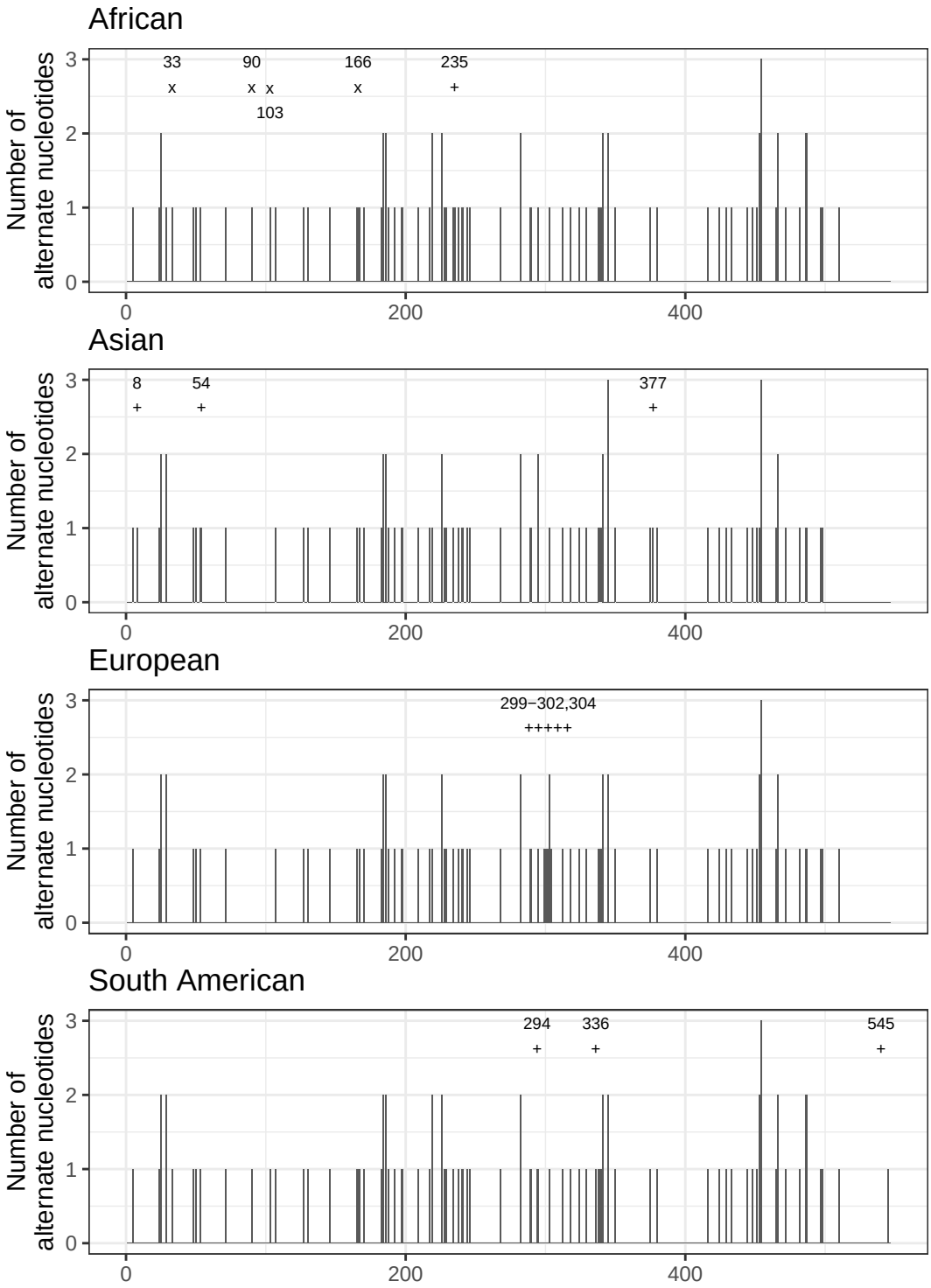


Figure 3.10: The number of alternate nucleotides per position within exons 2 and 3 of *HLA-A* from 1267 individuals from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotyping data were available. Four super-populations were analyzed: African, Asian, European and South American. + - population-specific variable sites. X - population-specific variable sites when compared between African, Asian and European super-populations

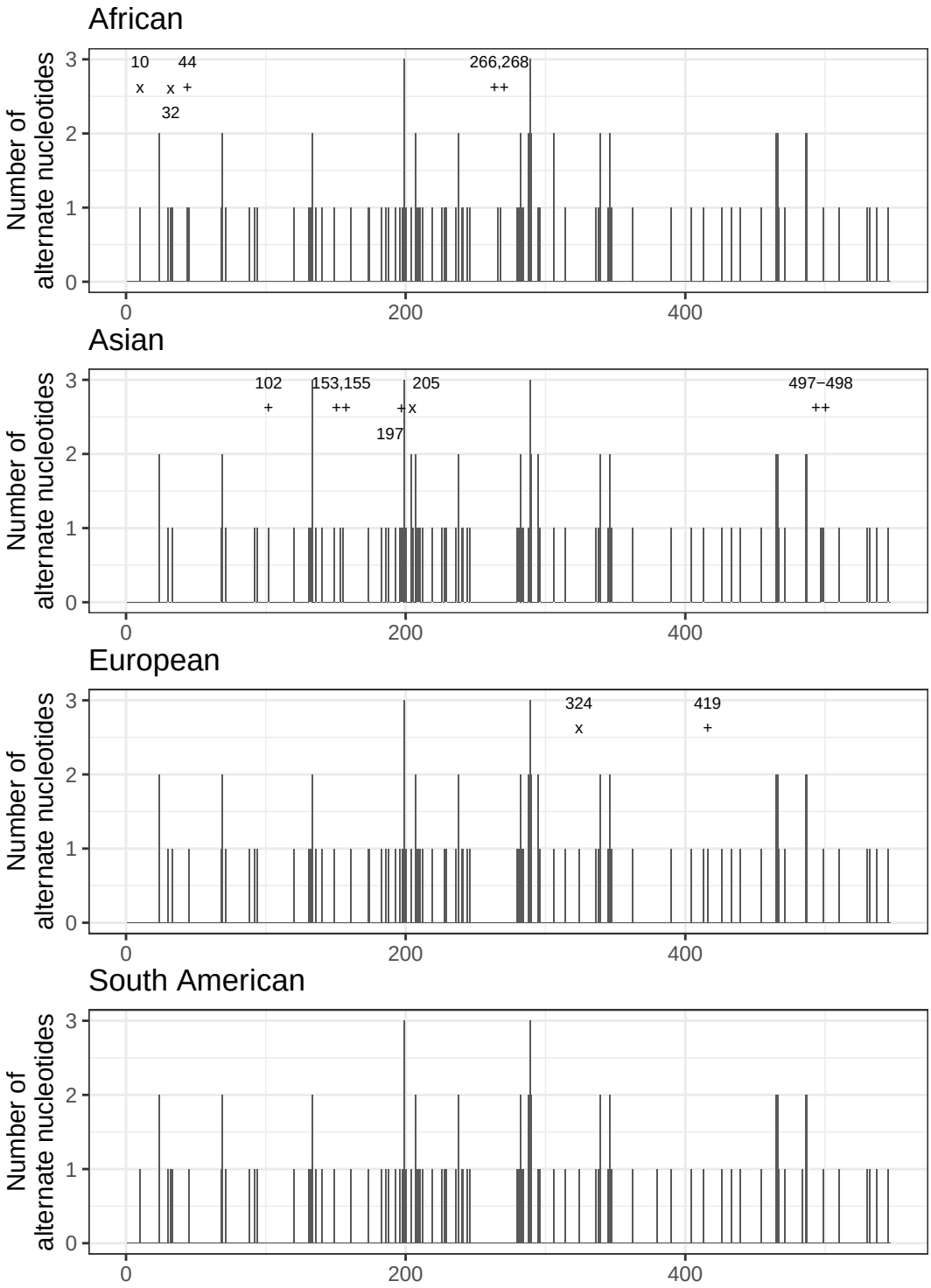


Figure 3.11: The number of alternate nucleotides per position within exons 2 and 3 of *HLA-B* from 1267 individuals from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotyping data were available. Four super-populations were analyzed: African, Asian, European and South American. + - population-specific variable sites. X - population-specific variable sites when compared between African, Asian and European super-populations

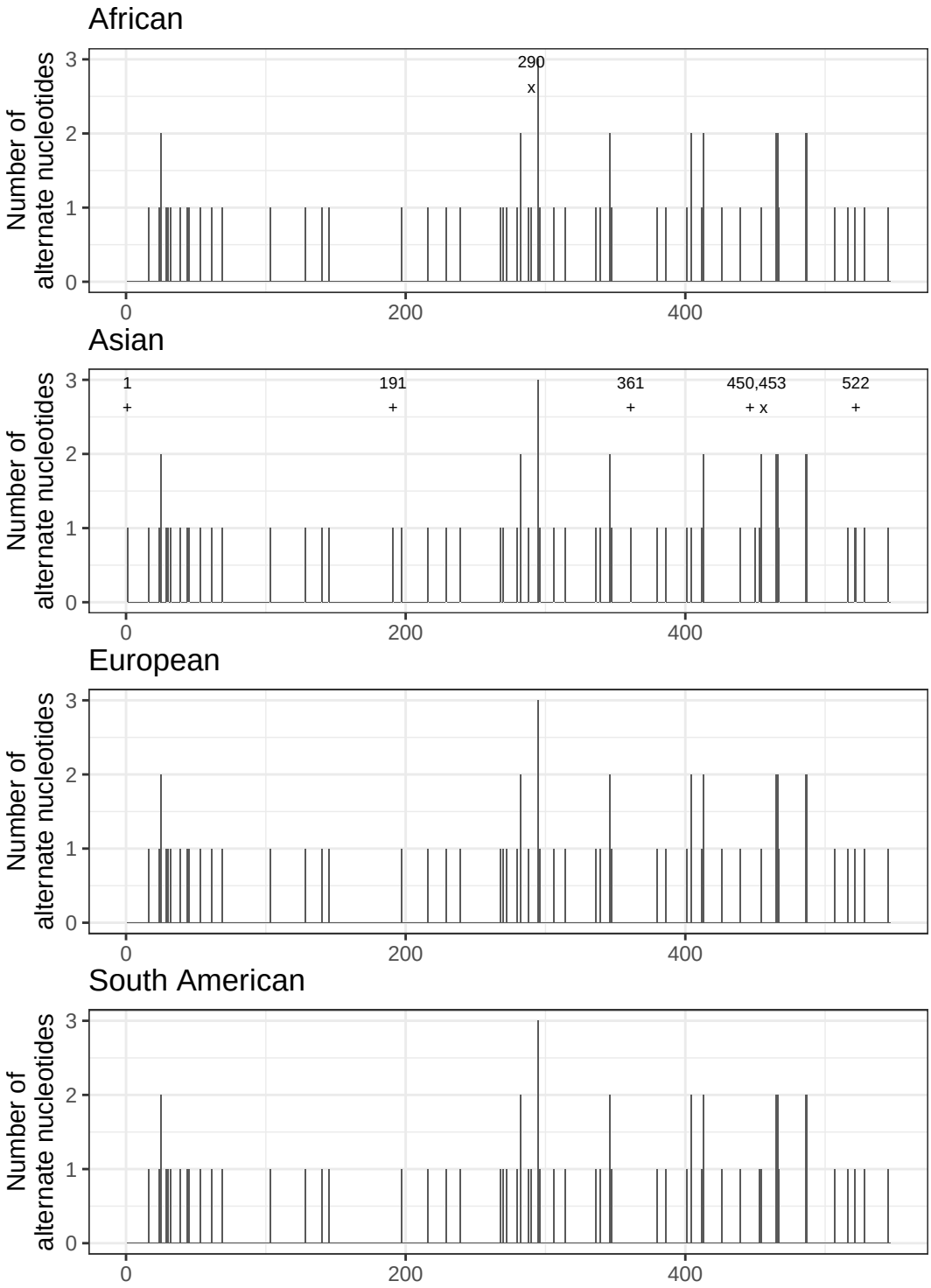


Figure 3.12: The number of alternate nucleotides per position within exons 2 and 3 of *HLA-C* from 1267 individuals from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotyping data were available. Four super-populations were analyzed: African, Asian, European and South American. + - population-specific variable sites. X - population-specific variable sites when compared between African, Asian and European super-populations

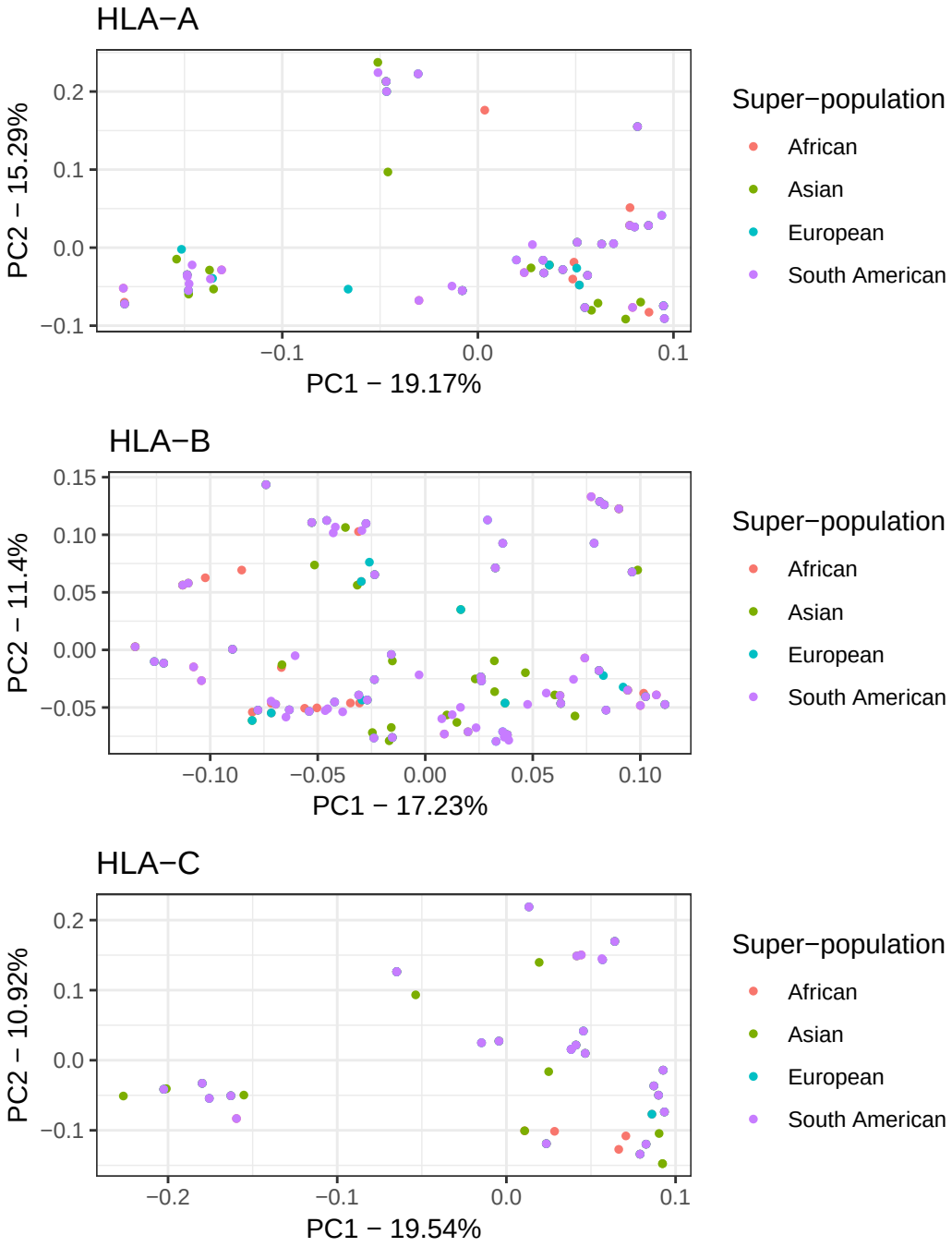


Figure 3.13: Principal Component Analysis of SNP-level *HLA* variability in 1267 individuals from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available

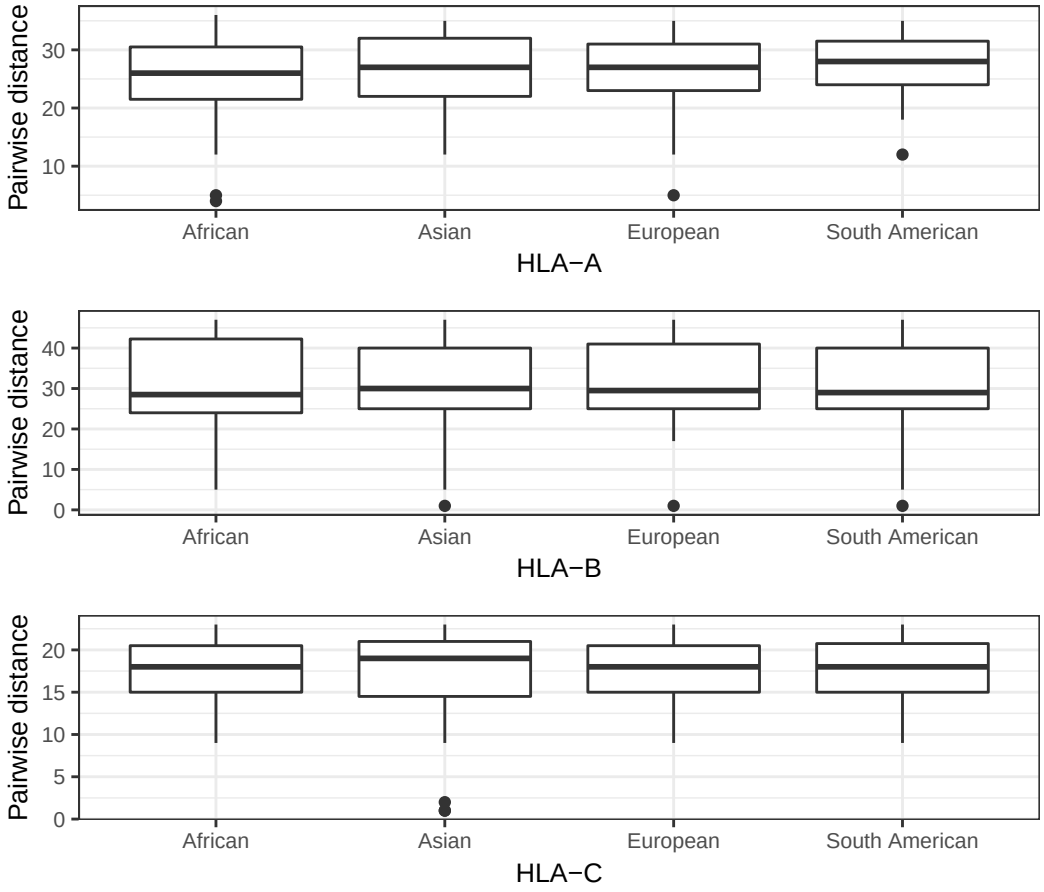


Figure 3.14: Box-plot of intragenic distances representing SNP-level variability of the four super-populations across *HLA-A*, *HLA-B*, and *HLA-C*. α = 0.05

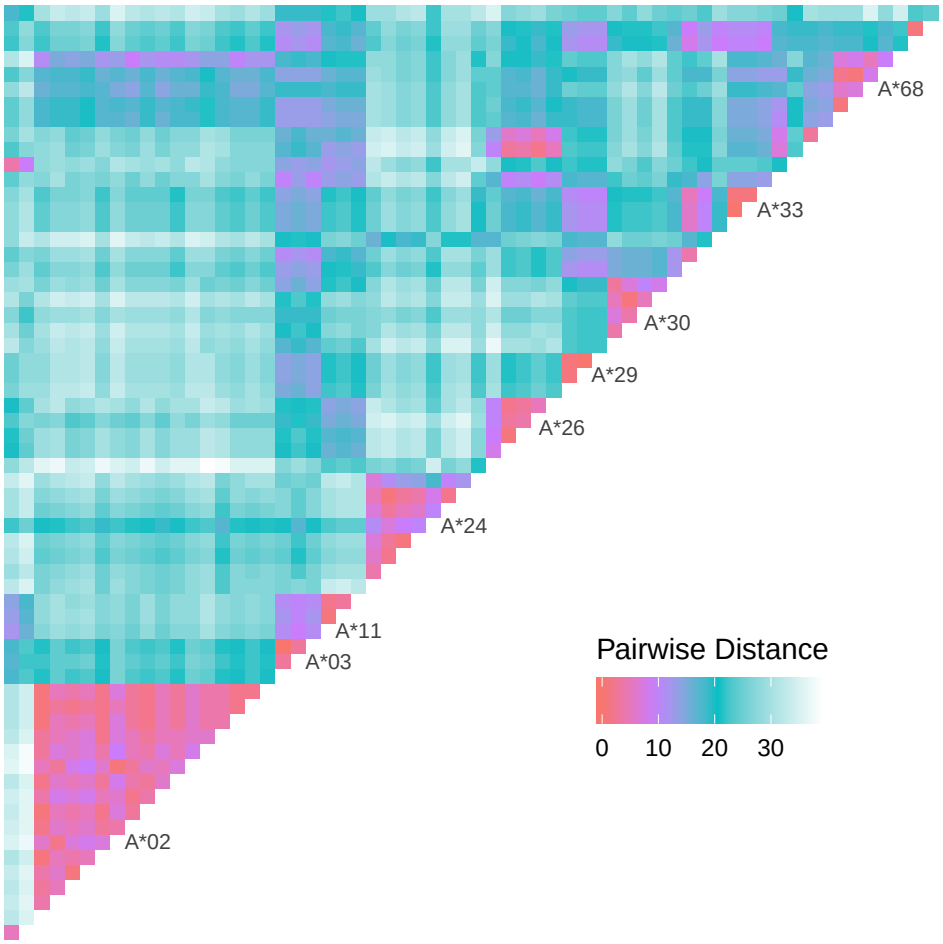


Figure 3.15: Intragenic distances between the *HLA-A* alleles observed in 1267 individuals from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available. Groups with fewer than three observed alleles are not shown: A*01, A*23, A*25, A*32, A*34, A*69, A*74, A*80

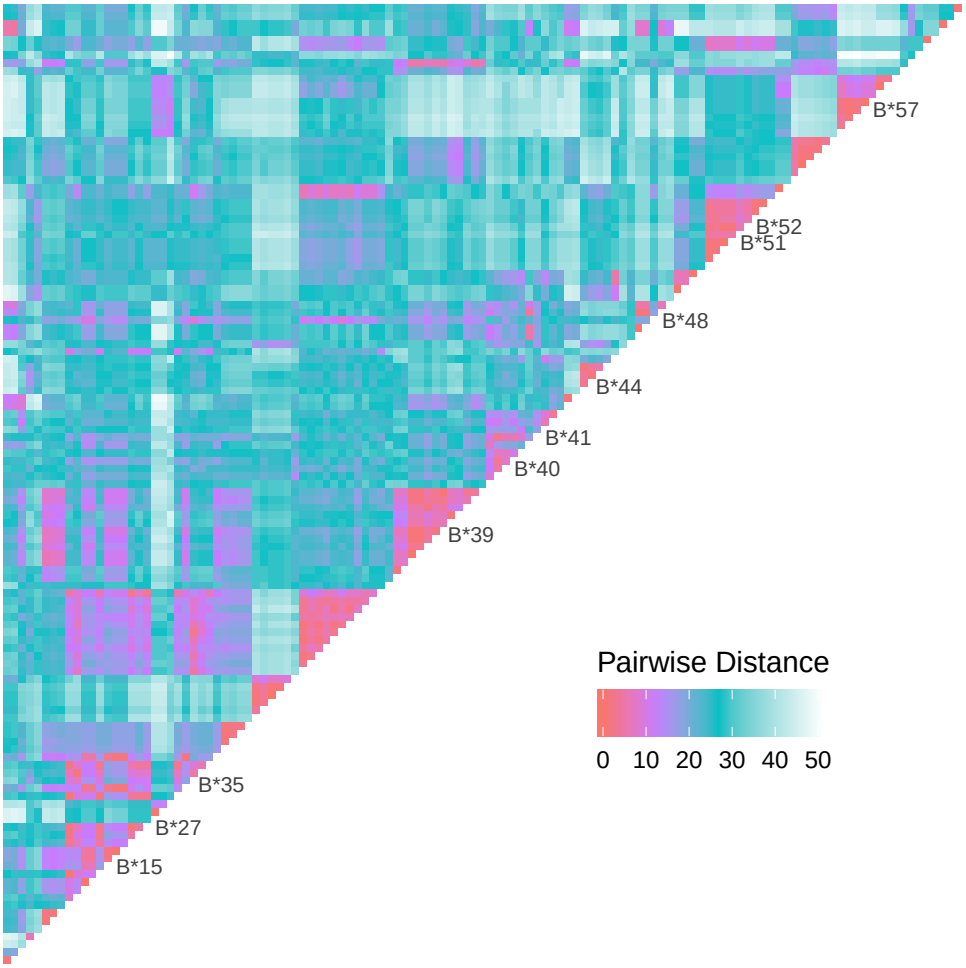


Figure 3.16: Intragenic distances between the *HLA-B* alleles observed in 1267 individuals from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available. Groups with fewer than three observed alleles are not shown:: B*07, B*08, B*13, B*37, B*38, B*42, B*45, B*46, B*47, B*49, B*50, B*53, B*58, B*59, B*67, B*73, B*78, B*81, B*82

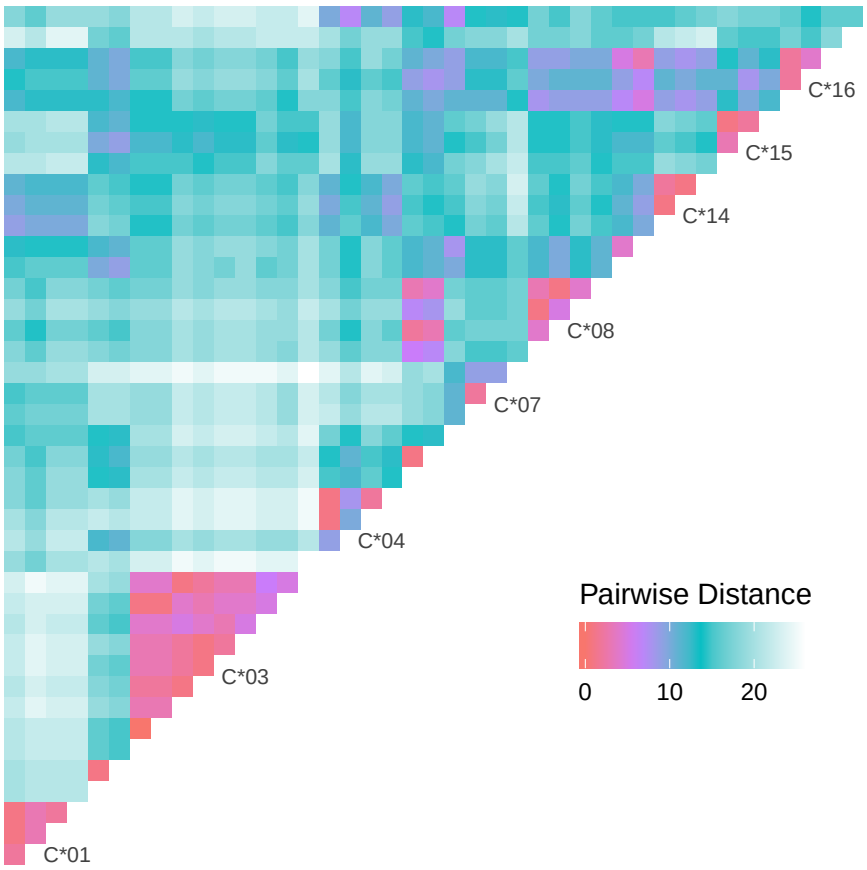


Figure 3.17: Intragenic distances between the *HLA-C* alleles observed in 1267 individuals from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available. Groups with fewer than three observed alleles are not shown:: C*02, C*05, C*12, C*17, C*18

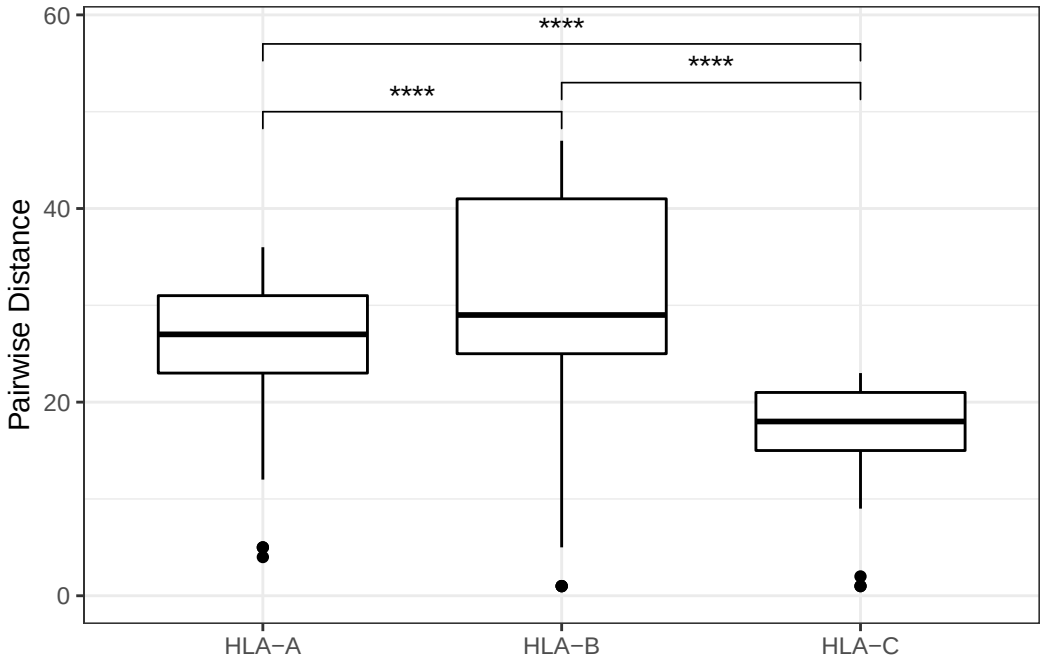


Figure 3.18: Boxplot of mean intragenic distance to reference for *HLA-A*, *HLA-B* and *HLA-C*. Significance reported from a Tukey HSD test. ****p<10^{-4}. $\alpha = 0.05$

Chapter 4
Discussion

This study aimed to compare the accuracy and limitations of four tools, namely, BWAkit, xHLA, HISAT-Genotype and Kourami. These tools have an advantage over previous tools, as they do not require high-end computer hardware, and can be run on a modern laptop or desktop computer. Through this research, 12 individuals were genotyped across their classical *HLA* class I loci. These 12 individuals were selected based on the availability of high-coverage WGS data from the 1000 Genomes Project, as well as high-resolution SBT *HLA* data. Furthermore, these 12 individuals have been utilized in numerous *HLA* genotyping and imputation studies, to benchmark the accuracy of a number of tools, including two of the tools (xHLA and Kourami) used in this study (Dilthey *et al.*, 2011; Huang *et al.*, 2015; Kawaguchi *et al.*, 2017; Kim *et al.*, 2018; Lee and Kingsford, 2018; Liu *et al.*, 2013a; Xie *et al.*, 2017; Zheng *et al.*, 2014). The four tools, used in this study, were evaluated in regards to run times and RAM usage across their respective pipelines, and accuracy, both at the allele- and SNP-level. From this, it was found that Kourami correctly assigned the most alleles at the six-digit resolution. At the two- and four-digit resolutions, xHLA was found to be the most accurate. Furthermore, with the exception of HISAT-Genotype, the tools were most accurate when genotyping *HLA-B*, and least accurate at *HLA-C*.

4.1 Allele-level Accuracy

xHLA utilized a nucleotide-to-protein alignment, which is the possible reason for the increased accuracy of xHLA at the two- and four-digit resolution when compared to the other tools, as the two- and four- digit resolution is indicative of peptide-binding, and therefore, the amino acid sequence of the encoding allele. Seven individuals (HG01112, NA19238, NA19239, NA19240, NA19625, NA19648 and NA20502) from this study were used by Xie *et al.* (2017) to

benchmark xHLA. Xie *et al.* (2017) found that the xHLA algorithm had a 99 - 100 percent accuracy when compared to SBT data at a four-digit resolution. From the results of this study, these individuals were accurately called at the four and six-digit resolution, which is concordant with Xie *et al.* (2017). With regards to the other three individuals that had incorrectly assigned genotypes, the incorrect allele calls were located at *HLA-C*. Upon further analyses, two of the three incorrect alleles were due to errors within the first codon of exon 2. Due to the exonic structure of *HLA*, there are incomplete codons at the intron-exon boundaries. As xHLA performs a nucleotide to protein alignment, the incomplete codons appear to result in a decrease in accuracy, as can be seen in HG00096 and NA18939, both of which had errors at MSA position 2, which resulted in non-synonymous changes. The algorithm utilized by xHLA does attempt to circumvent the affect of incomplete codons, by performing an additional nucleotide based alignment at the boundaries, however, it appears not to be as robust as the nucleotide-to-protein alignment.

The cohort used by Lee and Kingsford (2018) to test Kourami had eight individuals in common with the cohort used in this study (HG01112, NA18939, NA19238, NA19239, NA19240, NA19625, NA19648 and NA20502). The same genotyping results were found for seven of the individuals, with the exception of NA19625, in which Lee and Kingsford (2018) found Kourami correctly assigned genotypes to this individual, whereas this study found one allele was incorrectly genotyped. Lee and Kingsford (2018) utilized GRCh38 aligned CRAM files, whereas this study utilized GRCh37-aligned files, of which a subset of the reads were realigned to GRCh38. Therefore, the CRAM-aligned file for NA19625 was obtained and the analysis rerun. This resulted in the correct genotyping result. When the aligned read counts were determined, using the same parameters as previously used in this study, the CRAM file contained more reads than the BAM file used in this study. In the data acquisition process, utilized in this research, the specified regions for data extraction may have excluded reads, as these reads may have aligned to other regions. Furthermore, some reads pertaining to the *HLA* region may have been removed during preprocessing.

Kourami incorrectly assigned two alleles to HG01051, at *HLA-C*. The extracted sequences identified the correct nucleotides across exon 2 and exon 3. The error that resulted in the incorrect allele allocation appears to be due to a phasing issue between exons 2 and 3. This was observed, in the MSA, where the extracted sequence 1 is identical to *C*14:02:01G* along exon 2 and is inverted with extracted sequence 2, which aligns to *C*12:03:01G*. To identify possible causes of the incorrect phasing, the read depth was analyzed across exons 2 and 3, and the intronic region separating the two exons. While there was sufficient read depth across the two exons, the intronic region had a decrease in read depth, and a homozygous 5 bp deletion. This decrease in read depth in conjunction with the deletion may have resulted in the incorrect phasing of exons.

The cohort utilized by Kim *et al.* (2018), with HISAT-Genotype, included 17 related individuals, which were obtained from Illumina's Platinum Genomes, and had an average read depth of 50X. Through this, Kim *et al.* (2018) found that HISAT-Genotype correctly identified 100 percent of the alleles in the classical *HLA* class I and II genes. From this study, the accuracy of HISAT-Genotype was found to be lower. This decrease in accuracy is possibly due to the decrease in read depth, as the data was found to contain drops in read depth to approximately 20X in some instances. This was evident with HG01051, in which HISAT-Genotype correctly assigned two out of the six alleles.

Through the genotyping step within HISAT-Genotype, the number of genotyped alleles are not limited to two. This was found to be the case for HG00096, NA19648 and NA20502, in which more then two alleles were reported. Therefore, assembly was required to determine the two most probable alleles present. Assembly is further required to detect if a novel allele is present. HISAT-Genotype further produced two homozygous allele calls, both within *HLA-C*. From this study, the assembly resulted in the same genotyping calls as the alignment calls, including the incorrectly assigned genotypes. This additional step did not result in increased RAM use but did slightly increased the run time of the pipeline.

BWAkit does not contain published results, however, the authors do state that the accuracy of the tool is questionable, and other tools may be more applicable (https://github.com/lh3/bwa/tree/master/bwakit). From the results, BWAkit was the least accurate at all allele-resolutions. This was further evident at the SNP-level, in which BWAkit, again, had the most errors. BWAkit further was the only tool which did not produce a call for every allele. Similarly to HISAT-Genotype (but to a greater degree), BWAkit incorrectly assigned homozygous genotypes.

One critique of xHLA is the reliance upon an outdated IMGT/HLA database. Whilst the authors do not specify which exact version of the database is used, the data was obtained over two years ago. Kourami and BWAkit use version v. 3.24, and HISAT-Genotype uses v. 3.31. The differences in the number of alleles included in each database versions are large, with 13 580 alleles present in v. 3.21, 14 473 in v. 3.24 and 17 874 in v. 3.31. An advantage with Kourami, is that it is possible for the user to update the database to the latest one available(Lee and Kingsford, 2018). Another factor to consider is that in the older databases, many alleles do not have full-length sequence data and many alleles have been renamed and updated (Robinson *et al.*, 2015).

4.2 SNP-Level Accuracy

While numerous studies have utilized allele-level accuracy to benchmark the accuracy of *HLA* genotyping tools (Kim *et al.*, 2018; Lee and Kingsford, 2018; Szolek *et al.*, 2014; Xie *et al.*, 2017), a further factor to consider is the nucleotide sequence accuracy. This is because an allele is defined by its specific sequence, and two alleles with the same four-digit allele identifier may differ by numerous SNVs. These errors, which result in the incorrect allele call, can impact the results of studies, particularly when predicting antigen-binding, as previous research has found that single non-synonymous changes in the antigen-recognition site can affect the peptide-binding repertoire, and in particular, *HLA-B* is more susceptible to changes to the predicted binding repertoires from this (van Deutekom and Kesmir, 2015). Therefore, simply analyzing the accuracy of genotyping at

different resolutions (two-digit to six-digit) may not be applicable, depending on the research, and taking the pairwise alignment accuracy into account is important.

When the sequences of the incorrectly genotyped alleles were aligned to their SBT allele counterparts, it was found that xHLA was the most accurate, followed by Kourami, HISAT-Genotype and BWAkit. Further analysis of the alignments indicated that the graph-based tools - Kourami and HISAT-Genotype, were susceptible to falsely calling heterozygous positions as homozygous, which produced the majority of the errors. In regards to synonymous and non-synonymous SNP-level errors, xHLA did not appear to produce any bias, whereas the graph-based methods produced a far greater amount of non-synonymous errors when compared to synonymous errors. The increased SNP-level accuracy from xHLA may be due to the protein alignment step in the algorithm. The protein alignment step results in a reduced number of alleles used in the nucleotide alignment step, and this is beneficial, as there are less comparisons required thereafter. When looking at genotyping accuracy, it is important to remember that even closely related alleles may have different associations. An example of this is B*27:05 (Associated with Ankylosing Spondylitis) and B*27:09 (No association; Fiorillo *et al.* 1998). This has been found to be the case in numerous *HLA* alleles which differ by one amino acid, with predicted differences in peptide-binding repertoires (van Deutekom and Kesmir, 2015). As these differences are attributed to single amino acid changes between alleles, it demonstrates the importance of utilizing a nucleotide-to-protein step, as this drastically decreased the number of inferred synonymous changes to the alleles called by xHLA when compared to HISAT-Genotype and Kourami.

4.3 Effects of Alt-aware Alignments

An interesting observation in this research was the differences in the number of reads which aligned to the linear GRCh38 reference sequence when compared to alternate loci and the IMGT/HLA allele sequences. This increase in number of reads was efficiently utilized by Kourami, which had the highest accuracy at the six-digit resolution when the alternate sequences were utilized. Furthermore, when

comparing the alt-aligned and non-alt-aligned approaches used by Kourami, the non-alt-aligned approach resulted in decreased *HLA* genotyping accuracy, due to the decreased availability of aligned reads. In the alignment step, through BWA-mem, the mismatch, gap open and extension penalties prevent reads pertaining to the *HLA* region from aligning to the linear GRCh38 reference sequence. While it may be possible to relax these parameters to allow variant reads to align to the GRCh38 reference sequence, this could be detrimental in other areas, as it could result in increased mis-mapped reads, or reads that incorrectly align to certain regions. The *HLA* region contains numerous coding genes and psuedogenes, which are thought to have arisen through duplication (Pierini and Lenz, 2018), as there is high sequence similarity between these loci. This increases the chance of mis-mapped reads in this region.

Therefore, as xHLA and Kourami are capable of genotyping *HLA* from a pre-aligned BAM file, the approach to alignment can dictate which tool to use. If alternate loci and sequences obtained from the IMGT/HLA database were utilized, using xHLA is not recommended, as many of the reads that would map to the GRCh38 linear reference sequence would instead be mapped to either the alternate loci or the IMGT/HLA sequences. This would result in a decreased number of available reads for xHLA to utilize, and possibly decreased accuracy. The same can be stated for Kourami, with a linearly aligned BAM file, as was the case with this research, a far greater amount of reads align to alternate loci and IMGT/HLA sequences then the GRCh38 reference sequence.

Concerning the assigned alleles from the two different approaches, differing samples and alleles were incorrectly genotyped, with the exception of NA19625, at *HLA-A*, where both approaches assigned *HLA-A*02:571* as opposed to *HLA-A*02:01:01G*. Furthermore, where a reference allele was present (HG00096 - *HLA-A*, HG00268; NA18939; NA19625 - *HLA-C*), the linear approach correctly assigned the reference allele, and incorrectly assigned the other allele, indicating a possible reference sequence bias, and the importance of an alt-aware approach. Kourami uses a graph-based method, which has previously been shown to decrease reference strand bias (Garrison *et al.*, 2018; Paten *et al.*, 2017), and therefore, the

reference strand bias is possibly due to the weighting of the POG. As each read is aligned to the POG, the path that the read aligns to is weighted, and therefore, the alignment score for that path increases for each subsequent read that aligns to it. With access to reads that are similar to the linear reference sequence, this would introduce the reference sequence bias. With regards to xHLA, which also utilized a linear alignment approach, a contrasting pattern emerged, where if a reference allele was present (HG00096 and NA18939, *HLA-C*), an incorrect allele was assigned. This was not present for every reference allele present in the dataset, however, it appears to suggest that within the xHLA algorithm, there are parameters in place to reduce reference sequence bias.

4.4 Effects of Read Depth

The NGS data was obtained from high coverage (30X) WGS data, however, the 30X read depth is obtained from an average across the whole genome. Within the three *HLA* class I loci tested, the read depth was found to vary. One sample, HG01051, was found to have the lowest read depth across all three loci. Subsequently, all the tools produced errors when assigning genotypes to this individual, however, each tool assigned a different incorrect allele. Furthermore, HG01112 had a low read depth across *HLA-A*, and the tools (with the exclusion of BWAkit) correctly assigned genotypes to this individual at this locus. This indicates that read depth is not entirely indicative of genotyping accuracy. For this, depth of coverage across the whole gene may be a better indicator of possible genotyping accuracy, as HG01112 had decreased read depth within an intronic region.

4.5 Computational Time and Memory Use

As there are different alignment strategies available, in which alternate loci and sequences obtained from the IMGT/HLA database are optional, the running time between the two approaches were tested. It was found that the non-alt-aware approach was slightly quicker than the alt-aware approach, at all threads measured. The differences in time, however, were relatively small, and therefore, the

recommended approach is more dependent on the downstream analysis, than the running time differences between the two approaches.

When performing computationally intensive tasks, RAM is often a limiting factor. HISAT-Genotype utilized the most RAM, across the different number of threads tested, when compared to BWAkit, xHLA and Kourami. It should be noted that these four tools utilize less RAM than previous tools (Dilthey *et al.*, 2016; Szolek *et al.*, 2014), and are capable of running on a modern desktop computer. The reason HISAT-Genotype utilizes more RAM than the other three tools is due to the use of a hierarchical FM index, which is larger than a standard FM index. This may be problematic, as majority of computers contain 8 GB of RAM, followed closely by 16 GB of RAM (https://store.steampowered.com/hwsurvey/Steam-Hardware-Software-Survey-Welcome-to-Steam).

When utilizing a single thread, HISAT-Genotype was the quickest, followed by Kourami, xHLA and BWAkit. As more threads were specified, BWAkit, xHLA and Kourami required far less time to produce *HLA* genotypes. The increased thread count with BWAkit resulted in decreased time as well, however, the speed reduction appeared to plateau before Kourami and xHLA, possibly due to the use of concurrent commands used by BWAkit, which reduced available system resources. This was further evident when the RAM use was measured over time, as even though the three tools utilized the same alignment tool, BWA-mem, BWAkit was found to be slower. It was noted that specifying more threads resulted in a decrease in run times for BWAkit, xHLA and Kourami, whereas, HISAT-Genotype did not experience this decrease. This was due to the multi-threaded application within HISAT-Genotype, which is used only for indexing during the read alignment and extraction step. HISAT-Genotype states that the local FM indices, which make up the hierarchical FM index are small enough to utilize the CPU cache, which greatly reduces run times, as accessing data stored in the CPU cache memory is much faster than accessing data stored in RAM (Adam *et al.*, 1985). What these results indicate, however, is the importance of multi-threaded use.

In regards to the RAM usage across the different pipelines, it can be seen that all four tools utilize a large amount of RAM in the initial stages. This correlates to the alignment step of all four tools. The first decreases in RAM usage experienced by xHLA and Kourami correlate to the end of read extraction, and the beginning of the genotyping step, which for both of these tools required little time. As the four tools can utilize different approaches to genotyping *HLA*, in which Kourami and xHLA are capable of utilizing a BAM file directly as input (Lee and Kingsford, 2018; Xie *et al.*, 2017). The direct use of BAM files can result in a drastic decrease in computational time, as aligning reads accounts for majority of the running time. Furthermore, whether the BAM file consists of the whole genome, or is subset to the *HLA* region, does not influence the time required to extract reads as an indexed BAM file allows for quick read extraction. Through this, xHLA and Kourami are capable of generating genotypes from a 30X whole genome BAM file in approximately three minutes (Lee and Kingsford, 2018; Xie *et al.*, 2017).

4.6 Population-specific Variability

The allele counts within super-populations demonstrated that the South American individuals contained the highest number of alleles at *HLA-A* and *HLA-B*. Previous research has identified a large proportion of admixture within the South American populations, and therefore, the introduction of *HLA* alleles may have occurred through this. In particular, African admixture has previously been identified as a large component of the admixture within the South American populations (Ruiz-Linares *et al.*, 2014; Meyer *et al.*, 2018). From the cohort used in this research, the African super-population had the second most *HLA-A* and *HLA-B* alleles, and the variability within the African cohort may have added to the variability within the South Americans. The variable positions within each super-population were similar across the three loci tested, with similar patterns of per-position variation. To observe whether the alleles from different super-populations group together, a PCA was performed. The PCA demonstrated that there was no clear grouping of alleles within super-populations, however, the PCA did show that there are clusters of alleles shared by the super-populations. To further demonstrate this, the intragenic distances of the alleles were visualized.

The heatmaps indicated that there are indeed common clusters of allele families in each super-population, such as *HLA-A*02*, and that the specific alleles differ between populations. To test whether there was significant mean intragenic differences between super-populations, an ANOVA test was performed. The results indicated that there was no significant differences in mean intragenic distances between super-populations. This differs from previous research, in which whole genome analysis found clear grouping of autosomal SNPs within super-populations (Parsonnet and Xu, 1999; Wang *et al.*, 2010). This is possibly due to a combination of the unique evolutionary processes which resulted in the variation within *HLA*.

A previous comparison of *HLA* imputation tools found that the *HLA* genotyping accuracy in African cohorts was lower than in other populations (Karnes *et al.*, 2017). This was not observed in this cohort, as three of the four African samples had a 100 percent call rate accuracy from xHLA, HISAT-Genotype and Kourami. This, however, was more likely due to the higher read depth of the three samples. Furthermore, as the African individuals used in this cohort possess alleles which have been found to occur in numerous populations at a relatively high frequency (González-Galarza *et al.*, 2015) and complete sequence data for the entire allele is available from the IMGT/HLA database (Robinson *et al.*, 2015), these results are not indicative of *HLA* genotyping accuracy in an African cohort.

To further observe factors which could influence *HLA* genotyping accuracy, the three loci were analyzed. An ANOVA test was performed, to determine whether there was a significant mean intragenic distance difference between the loci. This was found to be significant, and therefore, a Tukey HSD test was performed to determine which loci differed. It was found that all three loci have significant differences in mean intragenic distances. This is concordant with previous research (Robinson *et al.*, 2017). To further demonstrate this, the intragenic heatmaps showed clear groupings of allele families in *HLA-A* and *HLA-C*, with smaller groups of alleles in *HLA-B*. This further demonstrates that the differences in variation between loci is an important factor to consider when genotyping *HLA*.

Therefore, as the PCA plots, and the ANOVA did not find significant variation between the super-populations, unique variant positions within *HLA* loci were analyzed. It has previously been found that 95 percent of positions within exons 2 and 3 exhibit variation, with majority of the positions exhibiting three alternate forms (including alternate nucleotides and indels). This, however, was from a study which analyzed all the alleles present in the IMGT/HLA database (v. 3.25.0; Robinson *et al.* 2015), and therefore, the variation unique to super-populations were analyzed. The variable positions were plotted in bargraphs, and the number of alternate nucleotides were used as an indication of variance per position. The four different super-populations exhibited a similar pattern to the variation, possibly due to the amino acids encoded at those positions, and their association with peptide-binding repertoires. When the unique positions were analyzed, none of the unique positions occurred within positions reported to encode peptides which interact with antigens or T-cells (van Deutekom and Kesmir, 2015). However, as van Deutekom and Kesmir (2015) found, single non-synonymous variations within *HLA-B* have a greater affect of peptide-binding repertoires then in *HLA-A* or *HLA-C*. This can effect the outcome of functional *HLA* studies, where an incorrect allele call at the two and four-digit resolution can impact the predicted binding repertoire. As the number of unique positions within the Asian super-population outnumber the other three super-populations, especially in *HLA-B*, this suggests that the peptide-binding repertoires within the Asian super-population can vary when compared to other super-populations.

Another strategy is to utilize pangenomes, in which a population- or super-population-specific reference assembly is constructed. This could be beneficial, as many *HLA* alleles have only been found within certain populations. This has previously been reported to be accurate in an Icelandic population (Eggertsson *et al.*, 2017), however, the population-specific variability analysis indicates that this might not be applicable, especially as there are numerous shared alleles.

4.7 Limitations

While this study compared and contrasted the accuracy of four genotyping tools, and considered the implications of population-specific effects on *HLA* genotyping accuracy, there were some limitations. The first is that only a subset of the WGS data were obtained. This was due to the computational requirements involved in managing a WGS BAM file. This did reduce the number of reads that were available to each tool, as was observed with sample NA19625, as Kourami was capable of correctly assigning alleles to this individual with data from a WGS BAM file. Furthermore, the concordance between the high-resolution SBT *HLA* genotype data and assigned genotypes from the four tools were only compared using the ambiguous allele format. This was due to both Kourami genotyping to the ambiguous allele resolution, and the sequencing of only exons 2 and 3 to generate the SBT data. Therefore, the true six-digit resolution accuracy was not determined. A further limitation, when comparing RAM usage, is with the one sample (NA19239) which was used. This sample was selected as all tools correctly genotyped this individual. This would decrease the bias, as the tools could have longer run times with samples that are incorrectly genotyped. Therefore, while using this sample resulted in decreased bias, the time usage comparison should be more reliant upon the average time per sample. Lastly, in regards to limitations with the *HLA* genotyping, is that the genotyping accuracy was not measured when the read depths were artificially varied, and therefore conclusions based on the read depth effect on accuracy could not be made.

With regards to the population-specific *HLA* variability, the cohort from Gourraud *et al.* (2014) consisted of 1267 individuals, and therefore, these individuals possessed far fewer alleles then are currently available in the IMGT/HLA database. Therefore, these individuals do not fully capture the possible variation present. In addition, the populations were grouped into super-populations. A more apt method may have been to compare populations, to observe whether geographically-linked populations grouped together into the super-populations, and thereafter, analyze the population-specific variability. While this may have introduced bias, through

intentionally grouping populations, it could also prevent admixed populations from obscuring the results.

4.8 Future work

Therefore, future work should involve using these four tools on a larger cohort. As BWAkit, xHLA and Kourami were less accurate when genotyping *HLA-C*, a larger cohort could allow for more significant conclusions to be drawn as to whether this specific locus has an affect on genotyping accuracy. Through a larger cohort, a wider variety of alleles may be tested. Furthermore, the ability of the graph-based tools to genotype novel alleles has not been tested outside of using simulated read data, and therefore, testing the tools using an individual with an allele not included in the tools IMGT/HLA database could be beneficial. Lastly, this research could be performed on *HLA* class II, as the class II loci are less variable then class I, and therefore, the decreased variability may affect the genotyping accuracy, as seen in *HLA-C*, in this study.

4.9 Conclusion

What these results indicate, is that *HLA* genotyping from Illumina paired-end high read depth (30X) WGS data is possible. This is particularly apparent from the results as each allele was correctly genotyped at least once. Furthermore, from the accuracy of xHLA and Kourami, using an alt-aware approach, followed by nucleotide-to-protein alignments is vital for accurate genotyping. Furthermore, in order to genotype a novel allele, a graph-based step is required. Finally, future improvements to *HLA* genotyping tools could introduce differing parameters when genotyping different loci. As the results indicated, three of the tools had decreased accuracy at *HLA-C*, and increased accuracy at *HLA-B*. Furthermore, the loci-specific variability may affect the accuracy of the assigned genotypes. Therefore, a consensus approach may be required for accurate genotyping, until either these programs are modified, or a future tool is released which builds upon the algorithms utilized by these tools.

References

Adam, M., Meadows, S. and McCaslin, R. (1985). Cache memory architecture for microcomputer speed-up board.

Aguilar, A., Roemer, G., Debenham, S., Binns, M., Garcelon, D. and Wayne, R. K. (2004). High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences*, vol. 101, no. 10, pp. 3490–3494.

Apanius, V., Penn, D., Slev, P. R., Ruff, L. R. and Potts, W. K. (1997). The Nature of Selection on the Major Histocompatibility Complex. *Critical Reviews^{TM} in Immunology*, vol. 17, no. 2, pp. 179–224.

Apps, R., Del Prete, G. Q., Chatterjee, P., Lara, A., Brumme, Z. L., Brockman, M. A., Neil, S., Pickering, S., Schneider, D. K., Piechocka-Trocha, A., Walker, B. D., Thomas, R., Shaw, G. M., Hahn, B. H., Keele, B. F., Lifson, J. D. and Carrington, M. (2016). HIV-1 Vpu Mediates HLA-C Downregulation. *Cell Host and Microbe*, vol. 19, no. 5, pp. 686–695.

Barone, J. C., Saito, K., Beutner, K., Campo, M., Dong, W., Goswami, C. P., Johnson, E. S., Wang, Z. X. and Hsu, S. (2015). HLA-genotyping of clinical specimens using Ion Torrent-based NGS. *Human Immunology*, vol. 76, no. 12, pp. 903–909.

Barter, R. and Yu, B. (2017). superheat: A Graphical Tool for Exploring Complex Datasets Using Heatmaps.

Bauer, D. C., Zadoorian, A., Wilson, L. O. W., Melbourne Genomics Health Alliance, M. G. H. and Thorne, N. P. (2016). Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Briefings in Bioinformatics*, vol. bbw097, pp. 1–9.

Bentley, G., Higuchi, R., Hoglund, B., Goodridge, D., Sayer, D., Trachtenberg, E. A. and Erlich, H. A. (2009). High-resolution, high-throughput HLA

genotyping by next-generation sequencing. *Tissue Antigens*, vol. 74, no. 5, pp. 393–403.

Bernatchez, L. and Landry, C. (2003). MHC studies in nonmodel vertebrates: What have we learned about natural selection in 15 years?

Bjorkman, P., Saper, M., B, S., Bennett, W., Strominger, J. L. and Wiley, D. C. (1987). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature*, vol. 329, no. 8, pp. 506–512.

Bjorkman, P. J. and Parham, P. (1990). Structure, function, and diversity of class I major histocompatibility complex molecules, vol. 59. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA.

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J. and Meyer, D. (2015). Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3: Genes—Genomes—genetics*, vol. 5, no. 5, pp. 931–941.

Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, vol. 34, no. 5, pp. 525–527.

Buchfink, B., Xie, C. and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, vol. 12, no. 1, pp. 59–60.

Buhler, S. and Sanchez-Mazas, A. (2011). HLA DNA sequence variation among human populations: Molecular signatures of demographic and selective events. *PLoS ONE*, vol. 6, no. 2.

Burrows, M. and Wheeler, D. J. (1994). Approximate pattern matching using the Burrows-Wheeler transform. Tech. rep., CADigital Equipment Corporation.

Carapito, R., Radosavljevic, M. and Bahram, S. (2016). Next-Generation Sequencing of the HLA locus: Methods and impacts on HLA typing, population genetics and disease association studies. *Human Immunology*, vol. 77, no. 11, pp. 1016–1023.

Carreno, B. M., Winter, C. C., Taurog, J. D., Hansen, T. H. and Biddison, W. E. (1993). Residues in pockets B and F of HLA-B27 are critical in the presentation of an influenza A virus nucleoprotein peptide and influence the stability of peptide-MHC complexes. *International Immunology*, vol. 5, no. 4, pp. 353–360.

Chan, S. H. and Tan, T. (1989). HLA and allopurinol drug eruption. *Dermatology*, vol. 179, no. 1, pp. 32–33.

Churko, J. M., Mantalas, G. L., Snyder, M. P. and Wu, J. C. (2013). Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circulation Research*, vol. 112, no. 12, pp. 1613–1623.

Colbert, R. A. (2000). HLA-B27 misfolding: A solution to the spondyloarthropathy conundrum? *Molecular Medicine Today*, vol. 6, no. 6, pp. 224–230.

Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W. and Ramachandran, S. (2015). A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences*, vol. 112, no. 5, pp. 1265–1272.

D'amato, M., Fiorillo, M. T., Carcassi, C., Mathieu, A., Zuccarelli, A., Bitti, P. P. and Sorrentino, R. (1995). Relevance of Residue 116 of HLA-B27 in Determining Susceptibility to Ankylosing Spondylitis. *European Journal of Immunology*, vol. 25, no. 11, pp. 3199–3201.

De Bakker, P. I., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., Ke, X., Monsuur, A. J., Whittaker, P., Delgado, M., Morrison, J., Richardson, A., Walsh, E. C., Gao, X., Galver, L., Hart, J., Hafler, D. A., Pericak-Vance, M., Todd, J. A., Daly, M. J., Trowsdale, J., Wijmenga, C., Vyse, T. J., Beck, S., Murray, S. S., Carrington, M., Gregory, S., Deloukas, P. and Rioux, J. D. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics*, vol. 38, no. 10, pp. 1166–1172.

Diez, R. A. (1990). HLA-B27 and agranulocytosis by levamisole.

Dilthey, A. T., Gourraud, P.-A. A., Mentzer, A. J., Cereb, N., Iqbal, Z. and McVean, G. (2016). High-Accuracy HLA Type Inference from Whole-Genome

Sequencing Data Using Population Reference Graphs. *PLoS Computational Biology*, vol. 12, no. 10, p. e1005151.

Dilthey, A. T., Moutsianas, L., Leslie, S. and McVean, G. (2011). HLA*IMP-an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics*, vol. 27, no. 7, pp. 968–972.

dos Santos Francisco, R., Buhler, S., Nunes, J. M., Bitarello, B. D., França, G. S., Meyer, D. and Sanchez-Mazas, A. (2015). HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. *Immunogenetics*, vol. 67, no. 11-12, pp. 651–663.

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A. A. A., Jonasdottir, A. A. A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K. and Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, vol. 49, no. 11, pp. 1654–1660.

Erlich, H. (2012). HLA DNA typing: past, present, and future. *Tissue Antigens*, vol. 80, no. 1, pp. 1–11.

Erlich, R. L., Jia, X., Anderson, S., Banks, E., Gao, X., Carrington, M., Gupta, N., DePristo, M. a., Henn, M. R., Lennon, N. J. and de Bakker, P. I. W. (2011). Next-generation sequencing for HLA typing of class I loci. *BMC Genomics*, vol. 12, no. 1, p. 42.

Fellay, J., Ge, D., Shianna, K. V., Colombo, S., Ledergerber, B., Cirulli, E. T., Urban, T. J., Zhang, K., Gumbs, C. E., Smith, J. P., Castagna, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Günthard, H. F., Mallal, S., Mussini, C., Dalmau, J., Martinez-Picado, J., Miro, J. M., Obel, N., Wolinsky, S. M., Martinson, J. J., Detels, R., Margolick, J. B., Jacobson, L. P., Descombes, P., Antonarakis, S. E., Beckmann, J. S., O'Brien, S. J., Letvin, N. L., McMichael, A. J., Haynes, B. F., Carrington, M., Feng, S., Telenti, A. and Goldstein, D. B. (2009). Common genetic variation and the control of HIV-1 in humans. *PLoS Genetics*, vol. 5, no. 12.

Ferragina, P., Manzini, G., Mäkinen, V. and Navarro, G. (2004). An Alphabet-Friendly FM-Index., vol. 3246. Heidelberg: Springer, Berlin, Heidelberg.

Fiorillo, M. T., Greco, G., Maragno, M., Potolicchio, I., Monizio, A., Dupuis, M. L. and Sorrentino, R. (1998). The naturally occurring polymorphism Asp116->His116, differentiating the ankylosing spondylitis-associated HLA-B*2705 from the non-associated HLA-B*2709 subtype, influences peptide-specific CD8 T cell recognition. *European Journal of Immunology*, vol. 28, no. 8, pp. 2508–2516.

Fogdell-Hahn, A., Ligers, A., Grønning, M., Hillert, J. and Olerup, O. (2000). Multiple sclerosis: A modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. *Tissue Antigens*, vol. 55, no. 2, pp. 140–148.

Francke, U. and Pellegrino, M. A. (1977). Assignment of the major histocompatibility complex to a region of the short arm of human chromosome 6. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 3, pp. 1147–51.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv*, p. 1207.3907.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B. and Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, vol. 36, no. 9, pp. 875–879.

Germain, R. N. and Margulies, D. H. (1993). The biochemistry and cell biology of antigen processing and presentation. *Annual Review of Immunology*, vol. 11, no. 1, pp. 403–450.

González-Galarza, F. F., Takeshita, L. Y., Santos, E. J., Kempson, F., Maia, M. H. T., Da Silva, A. L. S., Teles E Silva, A. L., Ghattaoraya, G. S., Alfirevic, A., Jones, A. R. and Middleton, D. (2015). Allele frequency net 2015 update: New features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Research*, vol. 43, no. D1, pp. 784–788.

Goulder, P. J., Bunce, M., Krausa, P., McIntyre, K., Crowley, S., Morgan, B., Edwards, A., Giangrande, P., Phillips, R. E. and McMichael, a. J. (1996). Novel, cross-restricted, conserved, and immunodominant cytotoxic T lymphocyte epitopes in slow progressors in HIV type 1 infection. *AIDS research and human retroviruses*, vol. 12, no. 18, pp. 1691–1698.

Gourraud, P. A., Khankhanian, P., Cereb, N., Yang, S. Y., Feolo, M., Maiers, M., Rioux, J. D., Hauser, S. and Oksenberg, J. (2014). HLA diversity in the 1000 genomes dataset. *PLoS ONE*, vol. 9, no. 7, p. e97282.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. and Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*, vol. 328, no. 5979, pp. 710–722.

Gumperz, J. E., Barber, L. D., Valiante, N. M., Percival, L., Phillips, J. H., Lanier, L. L. and Parham, P. (1997). Conserved and variable residues within the Bw4 motif of HLA-B make separable contributions to recognition by the NKB1 killer cell-inhibitory receptor. *Journal of immunology*, vol. 158, no. 11, pp. 5237–41.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S. and Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, vol. 10, no. 3, p. R32.

Hendel, H., Caillat-Zucman, S., Lebuanec, H., Carrington, M., O'Brien, S., Andrieu, J. M., Schächter, F., Zagury, D., Rappaport, J., Winkler, C., Nelson,

G. W. and Zagury, J. F. (1999). New class I and II HLA alleles strongly associated with opposite patterns of progression to AIDS. *Journal of immunology (Baltimore, Md. : 1950)*, vol. 162, no. 11, pp. 6942–6946.

Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J. and Greenwood, B. M. (1991a). Common West African HLA antigens are associated with protection from severe malaria. *Nature*, vol. 352, no. 6336, pp. 595–600.

Hill, A. V., Allsopp, C. E., McMichael, A. J., Kwiatkowski, D., Anstey, N. M. and Greenwood, B. M. (1991b). HLA class I typing by PCR: HLA-B27 and an African B27 subtype. *The Lancet*, vol. 337, no. 8742, pp. 640–642.

Holcomb, C. L., Höglund, B., Anderson, M. W., Blake, L. A., Böhme, I., Egholm, M., Ferriola, D., Gabriel, C., Gelber, S. E., Goodridge, D., Hawbecker, S., Klein, R., Ladner, M., Lind, C., Monos, D., Pando, M. J., Pröll, J., Sayer, D. C., Schmitz-Agheguian, G., Simen, B. B., Thiele, B., Trachtenberg, E. A., Tyan, D. B., Wassmuth, R., White, S. and Erlich, H. A. (2011). A multi-site study using high-resolution HLA genotyping by next generation sequencing. *Tissue Antigens*, vol. 77, no. 3, pp. 206–217.

Holmans, P. (2001). Ramifications of HLA class I polymorphism and population genetics for vaccine development. *Genetic Epidemiology*, vol. 20, no. 1, pp. 87–106.

Homburger, J. R., Moreno-Estrada, A., Gignoux, C. R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B. A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C. D., Gravel, S., Alarcón-Riquelme, M. E. and Bustamante, C. D. (2015). Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genetics*, vol. 11, no. 12, p. e1005602.

Huang, Y., Yang, J., Ying, D., Zhang, Y., Shotelersuk, V., Hirankarn, N., Sham, P. C., Lau, Y. L. and Yang, W. (2015). HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Medicine*, vol. 7, no. 1, p. 25.

Hughes, A. L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, vol. 335, no. 6186, pp. 167–170.

Jain, M. (2011). A next-generation approach to the characterization of a non-model plant transcriptome. *Current Science*, vol. 101, no. 11, pp. 1435–1439.

Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W. M., Concannon, P. J., Rich, S. S., Raychaudhuri, S. and de Bakker, P. I. (2013). Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS ONE*, vol. 8, no. 6, p. e64683.

Jose Casas, A. C., Casas, J., López, J. A., Delgado, R. G., Guerrero, M. L. F. and Górgolas, M. (2015). Long-Term Efficacy of Nevirapine Plus Co-Formulated Abacavir/Lamividune as Simplification Therapy in HIV-Infected Patients with Undetectable Viral Load. *Journal of AIDS & Clinical Research*, vol. 06, no. 05, pp. 1–5.

Jung, J. W., Kim, J. Y., Park, I. W., Choi, B. W. and Kang, H. R. (2018). Genetic markers of severe cutaneous adverse reactions.

Karnes, J. H., Shaffer, C. M., Bastarache, L., Gaudieri, S., Glazer, A. M., Steiner, H. E., Mosley, J. D., Mallal, S., Denny, J. C., Phillips, E. J. and Roden, D. M. (2017). Comparison of HLA allelic imputation programs. *PLoS ONE*, vol. 12, no. 2, p. e0172444.

Karolchik, D. (2003). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, vol. 32, no. 90001, pp. 493D–496.

Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. and Matsuda, F. (2017). HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Human Mutation*, vol. 38, no. 7, pp. 788–797.

Kim, D., Langmead, B. and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, vol. 12, no. 4, pp. 357–360.

Kim, D., Paggi, J. M. and Salzberg, S. (2018). HISAT-genotype: Next Generation Genomic Analysis Platform on a Personal Computer. *bioRxiv*, p. 266197.

Klepiela, P., Leslie, A. J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., Rathnavalu, P., Moore, C., Pfafferott, K. J., Hilton, L., Zimbwa, P., Moore, S., Allen, T., Brander, C., Addo, M. M., Altfeld, M., James, I., Mallal, S., Bunce, M., Barber, L. D., Szinger, J., Day, C., Klenerman, P., Mullins, J. J., Korber, B., Coovadia, H. M., Walker, B. D., Goulder, P. J. R., Kiepiela, P., Leslie, A. J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., Rathnavalu, P., Moore, C., Pfafferott, K. J., Hilton, L., Zimbwa, P., Moore, S., Allen, T., Brander, C., Addo, M. M., Altfeld, M., James, I., Mallal, S., Bunce, M., Barber, L. D., Szinger, J., Day, C., Klenerman, P., Mullins, J. J., Korber, B., Coovadia, H. M., Walker, B. D. and Goulder, P. J. R. (2004). Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature*, vol. 432, no. 7018, pp. 769–774.

Kristensen, T. and Mossin, J. (1982). Competitive inhibition of T-cell mediated lympholysis by platelets. *Tissue antigens*, vol. 19, no. 5, pp. 321–8.

Kumar, S., Stecher, G. and Tamura, K. (2015). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets.

Kurtz, S. (1999). Reducing the space requirement of suffix trees. *Software - Practice and Experience*, vol. 29, no. 13, pp. 1149–1171.

Langmead, B. (2010). Aligning Short Sequencing Reads with Bowtie. *Current Protocols in Bioinformatics*, vol. 32, no. 1, pp. 11.7.1–11.7.14.

Lee, H. and Kingsford, C. (2017). Graph-Guided Assembly For Novel HLA Allele Discovery. *bioRxiv*, p. 138826.

Lee, H. and Kingsford, C. (2018). Kourami: Graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biology*, vol. 19, no. 1, p. 16.

Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, vol. 49, no. 1, pp. 49–67.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, vol. 00, no. 00, p. 3.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079.

Li, H., Ruan, J. and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, vol. 18, no. 11, pp. 1851–1858.

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K. and Wang, J. (2009b). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967.

Li, Y., Yao, Y., Yang, M., Shi, L., Li, X., Yang, Y., Zhang, Y. and Xiao, C. (2013). Association between HLA-B*46 allele and Graves disease in Asian populations: a meta-analysis. *International journal of medical sciences*, vol. 10, no. 2, pp. 164–70.

Lindenbaum, P. (2015). JVarkit: java-based utilities for Bioinformatics.

Lippert, R. A. (2005). Space-Efficient Whole Genome Comparisons with Burrows–Wheeler Transforms. *Journal of Computational Biology*, vol. 12, no. 4, pp. 407–415.

Liu, C., Yang, X., Duffy, B., Mohanakumar, T., Mitra, R. D., Zody, M. C. and Pfeifer, J. D. (2013a). ATHLATES: Accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Research*, vol. 41, no. 14, pp. e142–e142.

Liu, X., Han, S., Wang, Z., Gelernter, J. and Yang, B. Z. (2013b). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, vol. 8, no. 9, p. e75619.

Ljunggren, H. G. and Karre, K. (1990). In search of the 'missing self': MHC molecules and NK cell recognition. *Immunology Today*, vol. 11, no. C, pp. 237–244.

Lonjou, C., Borot, N., Sekula, P., Ledger, N., Thomas, L., Halevy, S., Naldi, L., Bouwes-Bavinck, J. N., Sidoroff, A., De Toma, C., Schumacher, M., Roujeau,

J. C., Hovnanian, A. and Mockenhaupt, M. (2008). A European study of HLA-B in Stevens-Johnson syndrome and toxic epidermal necrolysis related to five high-risk drugs. *Pharmacogenetics and Genomics*, vol. 18, no. 2, pp. 99–107.

Lucas, A., Lucas, M., Strhyn, A., Keane, N. M., McKinnon, E., Pavlos, R., Moran, E. M., Meyer-Pannwitt, V., Gaudieri, S., D'Orsogna, L., Kalams, S., Ostrov, D. A., Buus, S., Peters, B., Mallal, S. and Phillips, E. (2015). Abacavir-reactive memory T cells are present in drug naïve individuals. *PLoS ONE*, vol. 10, no. 2, p. e0117160.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W. and Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, vol. 1, no. 1, p. 18.

MacDonald, K. S., Fowke, K. R., Kimani, J., Dunand, V. A., Nagelkerke, N. J., Ball, T. B., Oyugi, J., Njagi, E., Gaur, L. K., Brunham, R. C., Wade, J., Luscher, M. A., Krausa, P., Rowland-Jones, S., Ngugi, E., Bwayo, J. J. and Plummer, F. A. (2000). Influence of HLA supertypes on susceptibility and resistance to human immunodeficiency virus type 1 infection. *Journal of Infectious Diseases*, vol. 181, no. 5, pp. 1581–1589.

Major, E., Rigó, K., Hague, T., Bérces, A. and Juhos, S. (2013). HLA typing from 1000 Genomes whole genome and whole exome illumina data. *PLoS ONE*, vol. 8, no. 11, p. e78410.

Mallal, S., Nolan, D., Witt, C., Masel, G., Martin, A. M., Moore, C., Sayer, D., Castley, A., Mamotte, C., Maxwell, D., James, I. and Christiansen, F. T. (2002). Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet*, vol. 359, no. 9308, pp. 727–732.

Mandelboim, O., Reyburn, H. T., Sheu, E. G., Valés-Gómez, M., Davis, D. M., Pazmany, L. and Strominger, J. L. (1997). The binding site of NK receptors on HLA-C molecules. *Immunity*, vol. 6, no. 3, pp. 341–350.

Maretty, L., Jensen, J. M., Petersen, B., Sibbesen, J. A., Liu, S., Villesen, P., Skov, L., Belling, K., Theil Have, C., Izarzugaza, J. M. G., Grosjean, M., Bork-Jensen, J., Grove, J., Als, T. D., Huang, S., Chang, Y., Xu, R., Ye, W., Rao, J., Guo, X., Sun, J., Cao, H., Ye, C., van Beusekom, J., Espeseth, T., Flindt, E., Friborg, R. M., Halager, A. E., Le Hellard, S., Hultman, C. M., Lescai, F., Li, S., Lund, O., Løngren, P., Mailund, T., Matey-Hernandez, M. L., Mors, O., Pedersen, C. N. S., Sicheritz-Pontén, T., Sullivan, P., Syed, A., Westergaard, D., Yadav, R., Li, N., Xu, X., Hansen, T., Krogh, A., Bolund, L., Sørensen, T. I. A., Pedersen, O., Gupta, R., Rasmussen, S., Besenbacher, S., Børglum, A. D., Wang, J., Eiberg, H., Kristiansen, K., Brunak, S. and Schierup, M. H. (2017). Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*, vol. 548, no. 7665, pp. 87–91.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. a., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. a., Volkmer, G. a., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, vol. 437, no. 7057, pp. 376–80.

Marsh, S. G. E., 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. and Trowsdale, J. (2010). Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, vol. 75, no. 4, pp. 291–455.

Martínez, B., Caraballo, L., Hernández, M., Valle, R., Avila, M. and Iglesias Gamarra, A. (1999). HLA-B27 subtypes in patients with ankylosing spondylitis (As) in Colombia. *Revista de investigacion clinica; organo del Hospital de Enfermedades de la Nutricion*, vol. 51, no. 4, pp. 221–6.

Matsui, M., Hioe, C. E. and Frelinger, J. A. (1993). Roles of the six peptide-binding pockets of the HLA-A2 molecule in allorecognition by human cytotoxic T-cell clones. *Proceedings of the National Academy of Sciences*, vol. 90, no. 2, pp. 674–678.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, vol. 20, no. 9, pp. 1297–1303.

Merino, E., Galocha, B., Vázquez, M. N. and López De Castro, J. A. (2008). Disparate folding and stability of the ankylosing spondylitis-associated HLA-B*1403 and B*2705 proteins. *Arthritis and Rheumatism*, vol. 58, no. 12, pp. 3693–3704.

Meyer, D., Vitor, V. R., Bitarello, B. D., Débora, D. Y. and Nunes, K. (2018). A genomic perspective on HLA evolution.

Miretti, M. M., Walsh, E. C., Ke, X., Delgado, M., Griffiths, M., Hunt, S., Morrison, J., Whittaker, P., Lander, E. S., Cardon, L. R., Bentley, D. R., Rioux, J. D., Beck, S. and Deloukas, P. (2005). A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *American journal of human genetics*, vol. 76, no. 4, pp. 634–46.

Momburg, F. and Tan, P. (2002). Tapasin - The keystone of the loading complex optimizing peptide binding by MHC class I molecules in the endoplasmic reticulum. *Molecular Immunology*, vol. 39, no. 3-4, pp. 217–233.

Moretta, A., Bottino, C., Vitale, M., Pende, D., Biassoni, R., Mingari, M. C. and Moretta, L. (1996). RECEPTORS FOR HLA CLASS-I MOLECULES IN HUMAN NATURAL KILLER CELLS. *Annual Review of Immunology*, vol. 14, no. 1, pp. 619–648.

Nejentsev, S., Howson, J. M. M., Walker, N. M., Szeszko, J., Field, S. F., Stevens, H. E., Reynolds, P., Hardy, M., King, E. E. E., Masters, J., Hulme, J., Maier, L. M., Smyth, D., Bailey, R., Cooper, J. D., Ribas, G., Campbell, R. D., Clayton, D. G., Todd, J. A., Burton, P. R., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Donnelly, P., Barrett, J. H. J. C., Davison, D., Easton, D., Evans, D., Leung, H. T., Marchini, J. L., Morris, A. P., Spencer, C. C., Tobin, M. D., Attwood, A. P., Boorman, J. P., Cant, B., Everson, U., Hussey, J. M., Jolley, J. D., Knight, A. S., Koch, K., Meech, E., Nutland, S., Prowse, C. V., Taylor, N. C., Walters, G. R., Watkins, N. A., Winzer, T., Jones, R. W., McArdle, W. L., Ring, S. M., Strachan, D. P., Pembrey, M., Breen, G., St Clair, D., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E. K., Grozeva, D., Hamshere, M. L., Holmans, P. A., Jones, I. R., Kirov, G., Moskvina, V., Nikolov, I., O'Donovan, M. C., Owen, M. J., Collier, D. A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A. H., Ferrier, I. N., Ball, S. G., Balmforth, A. J., Barrett, J. H. J. C., Bishop, D. T., Iles, M. M., Maqbool, A., Yuldasheva, N., Hall, A. S., Braund, P. S., Dixon, R. J., Mangino, M., Stevens, S., Thompson, J. R., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C. W., Nimmo, E. R., Satsangi, J., Fisher, S. A., Forbes, A., Lewis, C. M., Onnie, C. M., Prescott, N. J., Sanderson, J., Mathew, C. G., Barbour, J., Mohiuddin, M. K., Todhunter, C. E., Mansfield, J. C., Tariq, A., Cummings, F. R., Jewell, D. P., Webster, J., Brown, M. A. M. J., Lathrop, G. M., Connell, J., Dominiczak, A., Braga Marcano, C. A., Burke, B., Dobson, R., Gungadoo, J., Lee, K. L., Munroe, P. B., Newhouse, S. J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., Bruce, I. N., Donovan, H., Eyre, S., Gilbert, P. D., Hider, S. L., Hinks, A. M., John, S. L., Potter, C., Silman, A. J., Symmons, D. P., Thomson, W., Worthington, J., Dunger, D. B., Widmer, B., Frayling, T. M., Freathy, R. M., Lango, H., Perry, J. R., Shields, B. M., Weedon, M. N., Hattersley, A. T., Hitman, G. A., Walker,

M., Elliott, K. S., Groves, C. J., Lindgren, C. M., Rayner, N. W., Timpson, N. J., Zeggini, E., Newport, M., Sirugo, G., Lyons, E., Vannberg, F., Hill, A. V., Bradbury, L. A., Farrar, C., Pointon, J. J., Wordsworth, P., Brown, M. A. M. J., Franklyn, J. A., Heward, J. M., Simmonds, M. J., Gough, S. C., Seal, S., Stratton, M. R., Rahman, N., Ban, M., Goris, A., Sawcer, S. J., Compston, A. A., Conway, D., Jallow, M., Rockett, K. A., Bryan, C., Bumpstead, S. J., Chaney, A., Downes, K., Ghori, J., Gwilliam, R., Hunt, S. E., Inouye, M., Keniry, A., King, E. E. E., McGinnis, R., Potter, S., Ravindrarajah, R., Whittaker, P., Withers, D., Cardin, N. J., Ferreira, T., Pereira-Gale, J., Hallgrimsdóttir, I. B., Howie, B. N., Su, Z., Yik, Y. T., Vukcevic, D., Bentley, D., Compston, A. A. and Wellcome Trust Case Control Consortium (2007). Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature*, vol. 450, no. 7171, pp. 887–892.

Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, vol. 12, no. 6, pp. 443–451.

Ober, C., Weitkamp, L. R., Cox, N., Dytch, H., Kostyu, D. and Elias, S. (1997). HLA and Mate Choice in Humans. *American journal of human genetics*, vol. 61, pp. 497–504.

O'Callaghan, C. A. and Bell, J. I. (1998). Structure and function of the human MHC class Ib molecules HLA-E, HLA-F and HLA-G. *Immunological Reviews*, vol. 163, no. 1, pp. 129–138.

Omar, M. A., Hammond, M. G. and Asmal, A. C. (1984). HLA-A, B, C and DR antigens in young South African blacks with Type 1 (insulin-dependent) diabetes mellitus. *Diabetologia*, vol. 26, no. 1, pp. 20–23.

Parsonnet, . J. and Xu, . Y. (1999). Supporting Online Material Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Genomics*, vol. 22, p. 329.

Paten, B., Novak, A. M., Eizenga, J. M. and Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research*, vol. 27, no. 5, pp. 665–676.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, vol. 14, no. 4, pp. 417–419.

Phillips, E., Bartlett, J. A., Sanne, I., Lederman, M. M., Hinkle, J., Rousseau, F., Dunn, D., Pavlos, R., James, I., Mallal, S. A. and Haas, D. W. (2013). Associations between HLA-DRB1*0102, HLA-B*5801, and hepatotoxicity during initiation of nevirapine-containing regimens in South Africa.

Pierini, F. and Lenz, T. L. (2018). Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection. *Molecular Biology and Evolution*, vol. 35, no. 9, pp. 2145–2158.

Popov, E. A., Levitan, B. N., Alekseev, L. P., Pronina, O. A. and Suchkov, S. V. (2005). Immunogenetic HLA markers of chronic viral hepatitis. *Terapevticheskii arkhiv*, vol. 77, no. 2, pp. 54–9.

Prugnolle, F., Manica, A., Charpentier, M., Guégan, J. F., Guernier, V. and Balloux, F. (2005). Pathogen-driven selection and worldwide HLA class I diversity. *Current Biology*, vol. 15, no. 11, pp. 1022–1027.

R Core Development Team (2018). R: A Language and Environment for Statistical Computing. *Vienna: R Foundation for Statistical Computing*.

Rammensee, H. G., Falk, K. and Rötzschke, O. (1993). Peptides Naturally Presented by MHC Class I Molecules. *Annual Review of Immunology*, vol. 11, no. 1, pp. 213–244.

Reich, D. E., Cargili, M., Boik, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R. and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, vol. 411, no. 6834, pp. 199–204.

Robinson, J., Guethlein, L. A., Cereb, N., Yang, S. Y., Norman, P. J., Marsh, S. G. and Parham, P. (2017). Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genetics*, vol. 13, no. 6, p. e1006862.

Robinson, J., Halliwell, J., Hayhurst, J., Flicek, P., Parham, P. and Marsch, S. (2015). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research*, vol. 43, pp. 423–432.

Robinson, J., Halliwell, J. A., McWilliam, H., Lopez, R., Parham, P. and Marsh, S. G. E. (2013). The IMGT/HLA database. *Nucleic Acids Research*, vol. 41, no. D1, pp. 1222–1227.

Ruggeri, L., Capanni, M., Casucci, M., Volpi, I., Tosti, A., Perruccio, K., Urbani, E., Negrin, R. S., Martelli, M. F. and Velardi, A. (1999). Role of natural killer cell alloreactivity in HLA-mismatched hematopoietic stem cell transplantation. *Blood*, vol. 94, no. 1, pp. 333–9.

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., Gómez-Valdés, J., León-Mimila, P., Hunemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Burley, M. W., Konca, E., de Oliveira, M. Z., Veronez, M. R., Rubio-Codina, M., Attanasio, O., Gibbon, S., Ray, N., Gallo, C., Poletti, G., Rosique, J., Schuler-Faccini, L., Salzano, F. M., Bortolini, M. C., Canizales-Quinteros, S., Rothhammer, F., Bedoya, G., Balding, D. and Gonzalez-José, R. (2014). Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLoS Genetics*, vol. 10, no. 9, p. e1004572.

Sadasivan, B., Lehner, P. J., Ortmann, B., Spies, T. and Cresswell, P. (1996). Roles for calreticulin and a novel glycoprotein, tapasin, in the interaction of MHC class I molecules with TAP. *Immunity*, vol. 5, no. 2, pp. 103–114.

Sampaio-Barros, P. D., Bertolo, M. B., Kraemer, M. H., Neto, J. F. and Samara, A. M. (2001). Primary ankylosing spondylitis: patterns of disease in a Brazilian population of 147 patients. *The Journal of rheumatology*, vol. 28, no. 3, pp. 560–5.

Santamaria, P., Lindstrom, A. L., Boyce-Jacino, M. T., Myster, S. H., Barbosa, J. J., Faras, A. J. and Rich, S. S. (1993). HLA class I sequence-based typing. *Human Immunology*, vol. 37, no. 1, pp. 39–50.

Saper, M. A., Bjorkman, P. J. and Wiley, D. C. (1991). Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. *Journal of Molecular Biology*, vol. 219, no. 2, pp. 277–319.

Scheet, P. and Stephens, M. (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics*, vol. 78, no. 4, pp. 629–644.

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caolle, C., Bajorek, E., Black, S., Chan, Y. M., Denys, M., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Gomez, M., Gonzales, E., Haydu, L., Lopez, F., Ramirez, L., Retterer, J., Rodriguez, A., Rogers, S., Salazar, A., Tsai, M. and Myers, R. M. (2004). Quality assessment of the human genome sequence. *Nature*, vol. 429, no. 6990, pp. 365–368.

Sette, A. and Sidney, J. (1998). HLA supertypes and supermotifs: A functional perspective on HLA polymorphism.

Shankarkumar, U., Pawar, A. and Ghosh, K. (2008). Implications of HLA sequence-based typing in transplantation. *Journal of postgraduate medicine*, vol. 54, no. 1, pp. 41–4.

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., Levin, A. M., Eng, C., Yazdanbakhsh, M., Wilson, J. G., Marrugo, J., Lange, L. A., Williams, L. K., Watson, H., Ware, L. B., Olopade, C. O., Olopade, O., Oliveira, R. R., Ober, C., Nicolae, D. L., Meyers, D. A., Mayorga, A., Knight-Madden, J., Hartert, T., Hansel, N. N., Foreman, M. G., Ford, J. G., Faruque, M. U., Dunston, G. M., Caraballo, L., Burchard, E. G., Bleecker, E. R., Araujo, M. I., Herrera-Paz, E. F., Campbell, M., Foster, C., Taub, M. A., Beaty, T. H., Ruczinski, I., Mathias, R. A., Barnes, K. C. and Salzberg, S. L. (2018). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*.

Shih, H. C., Liu, S. C., Chang, C. P., Tschen, J. S., Chiu, H. Y., Liu, H. C. and Chang, J. G. (2001). Positive association of ankylosing spondylitis with

homozygous HLA-B2704, but protection with B2705 in Taiwan Chinese. *The Kaohsiung journal of medical sciences*, vol. 17, no. 10, pp. 509–16.

Shiina, T., Inoko, H. and Kulski, J. K. (2004). An update of the HLA genomic region, locus information and disease associations: 2004.

Sidney, J., Peters, B., Frahm, N., Brander, C. and Sette, A. (2008). HLA class I supertypes: a revised and updated classification. *BMC immunology*, vol. 9, p. 1.

Simmonds, M. J., Howson, J. M., Heward, J. M., Carr-Smith, J., Franklyn, J. A., Todd, J. A. and Gough, S. C. (2007). A novel and major association of HLA-C in Graves' disease that eclipses the classical HLA-DRB1 effect. *Human Molecular Genetics*, vol. 16, no. 18, pp. 2149–2153.

Smith, A. D., Xuan, Z. and Zhang, M. Q. (2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, vol. 9, no. 1, p. 128.

Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M. and Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics (Oxford, England)*, vol. 30, no. 23, pp. 3310–6.

Taub, Floyd, E., Deleo, J. M. and Thompson, E. B. (1983). Sequential Comparative Hybridizations Analyzed by Computerized Image Processing Can Identify and Quantitate Regulated RNAs. *DNA*, vol. 2, no. 4, pp. 309–327.

The 1000 Genomes Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, vol. 467, no. 7319, pp. 1061–1073.

The 1000 Genomes Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, vol. 491, no. 7422, pp. 56–65.

The 1000 Genomes Consortium (2015). A global reference for human genetic variation. *Nature*, vol. 526, no. 7571, pp. 68–74.

The International Genome Sequencing Consortium and Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, vol. 431, no. 7011, pp. 931–945.

The International HIV Controllers (2011). The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science (New York, N.Y.)*, vol. 330, no. 6010, pp. 1551–1557.

Thio, C. L., Thomas, D. L., Karacki, P., Gao, X., Marti, D., Kaslow, R. A., Goedert, J. J., Hilgartner, M., Strathdee, S. A., Duggal, P., O'Brien, S. J., Astemborski, J. and Carrington, M. (2003). Comprehensive analysis of class I and class II HLA antigens and chronic hepatitis B virus infection. *Journal of virology*, vol. 77, no. 22, pp. 12083–7.

Tishkoff, S. A. and Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews Genetics*, vol. 3, no. 8, pp. 611–621.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, vol. 13, no. 1, pp. 36–46.

Turcatti, G., Romieu, A., Fedurco, M. and Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research*, vol. 36, no. 4, pp. e25–e25.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. and DePristo, M. A. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, vol. 11.10, no. Supplement 45, pp. 1–33.

van Deutekom, H. W. and Kesmir, C. (2015). Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most? *Immunogenetics*, vol. 67, no. 8, pp. 425–436.

Vitale, M., Bottino, C., Sivori, S., Sanseverino, L., Castriconi, R., Marcenaro, E., Augugliaro, R., Moretta, L. and Moretta, A. (1998). NKp44, a novel triggering

surface molecule speci cally expressed by activated natural killer cells, is involved in non-major histocompatibility complex-restricted tumor cell lysis. *J. Exp. Med*, vol. 187, no. 12, pp. 2065–2072.

Wang, C., Szpiech, Z. A., Degnan, J. H., Jakobsson, M., Pemberton, T. J., Hardy, J. A., Singleton, A. B. and Rosenberg, N. A. (2010). Comparing spatial maps of human population-genetic variation using procrustes analysis. *Statistical Applications in Genetics and Molecular Biology*, vol. 9, no. 1, p. Article 13.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. and Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. *G3: Genes—Genomes—Genetics*, vol. 5, no. 8, pp. 1543–1550.

Wearsch, P. A. and Cresswell, P. (2007). Selective loading of high-affinity peptides onto major histocompatibility complex class I molecules by the tapasin-ERp57 heterodimer. *Nature Immunology*, vol. 8, no. 8, pp. 873–881.

Wolfe, D., Dudek, S., Ritchie, M. D. and Pendergrass, S. A. (2013). Visualizing genomic information across chromosomes with PhenoGram. *BioData Mining*, vol. 6, no. 1, p. 18.

Wu, Y.-F., Wang, L.-Y., Lee, T.-D., Lin, H. H., Hu, C.-T., Cheng, M.-L. and Lo, S.-Y. (2004). HLA phenotypes and outcomes of hepatitis B virus infection in Taiwan. *Journal of Medical Virology*, vol. 72, no. 1, pp. 17–25.

Xie, C., Yeo, Z. X., Wong, M., Piper, J., Long, T., Kirkness, E. F., Biggs, W. H., Bloom, K., Spellman, S., Vierra-Green, C., Brady, C., Scheuermann, R. H., Telenti, A., Howard, S., Brewerton, S., Turpaz, Y. and Venter, J. C. (2017). Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proceedings of the National Academy of Sciences*, vol. 114, no. 30, pp. 8059–8064.

Xuan, J., Yu, Y., Qing, T., Guo, L. and Shi, L. (2013). Next-generation sequencing in the clinic: Promises and challenges. *Cancer Letters*, vol. 340, no. 2, pp. 284–295.

Yeo, T. W., De Jager, P. L., Gregory, S. G., Barcellos, L. F., Walton, A., Goris, A., Fenoglio, C., Ban, M., Taylor, C. J., Goodman, R. S., Walsh, E., Wolfish, C. S., Horton, R., Traherne, J., Beck, S., Trowsdale, J., Caillier, S. J., Ivinson, A. J., Green, T., Pobywajlo, S., Lander, E. S., Pericak-Vance, M. A., Haines, J. L., Daly, M. J., Oksenberg, J. R., Hauser, S. L., Compston, A., Hafler, D. A., Rioux, J. D. and Sawcer, S. (2007). A second major histocompatibility complex susceptibility locus for multiple sclerosis. *Annals of Neurology*, vol. 61, no. 3, pp. 228–236.

Zeniya, M., Watanabe, F., Aizawa, Y. and Toda, G. (1993). Immunogenetic background of hepatitis B virus infection and autoimmune hepatitis in Japan. *Gastroenterologia Japonica*, vol. 28 Suppl 4, pp. 69–75; discussion 76–80.

Zerbino, D. R. and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, vol. 18, no. 5, pp. 821–829.

Zheng, X., Gogarten, S., Lawrence, M., Stilp, A., Conomos, M., Weir, B., Laurie, C. and Levine, D. (2017). SeqArray – A storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*.

Zheng, X., Levine, D., Shen, J., Gogarten, S., Laurie, C. and Weir, B. (2012). A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics*, vol. 28, no. 24, pp. 3326–3328.

Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R. and Weir, B. S. (2014). HIBAG - HLA genotype imputation with attribute bagging. *Pharmacogenomics Journal*, vol. 14, no. 2, pp. 192–200.

Appendix A - List of Tools and Sample Locations

Table A1: List of programs used

Program	Version	Reference	URL
BWAkit	0.7.11	Li (2013)	https://github.com/lh3/bwa/tree/master/bwakit
BWA	0.7.11	Li (2013)	https://github.com/lh3/bwa
HISAT-Genotype	1.01b	Kim *et al.* (2018)	http://ccb.jhu.edu/hisat-genotype/index.php/Main_Page
Kourami	0.9.6	Lee and Kingsford (2018)	https://github.com/Kingsford-Group/kourami
xHLA	04/10/2017	Xie *et al.* (2017)	https://github.com/humanlongevity/HLA
SAMtools	1.9	Li *et al.* (2009a)	https://github.com/samtools/samtools
PICARD	2.18.0		https://broadinstitute.github.io/picard/
R	3.5.1	R Core Development Team (2018)	https://www.r-project.org

Table A2: Sample file locations

Sample	Data Location
HG00096	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ HG00096/high_coverage_alignment/HG00096.wgs.ILLUMINA.bwa.GBR. high_cov_pcr_free.20140203.bam
HG00268	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ HG00268/high_coverage_alignment/HG00268.wgs.ILLUMINA.bwa.FIN. high_cov_pcr_free.20140203.bam
HG00419	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ HG00419/high_coverage_alignment/HG00419.wgs.ILLUMINA.bwa.CHS. high_cov_pcr_free.20140203.bam
HG01051	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ HG01051/high_coverage_alignment/HG01051.wgs.ILLUMINA.bwa.PUR. high_cov_pcr_free.20140203.bam
HG01112	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ HG01112/high_coverage_alignment/HG01112.wgs.ILLUMINA.bwa.CLM. high_cov_pcr_free.20140203.bam
NA18939	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ NA18939/high_coverage_alignment/NA18939.wgs.ILLUMINA.bwa.JPT. high_cov_pcr_free.20140203.bam
NA19238	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ NA19238/high_coverage_alignment/NA19238.mapped.ILLUMINA.bwa. YRI.high_coverage_pcr_free.20130924.bam
NA19239	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ NA19239/high_coverage_alignment/NA19239.mapped.ILLUMINA.bwa. YRI.high_coverage_pcr_free.20130924.bam
NA19240	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ NA19240/high_coverage_alignment/NA19240.mapped.ILLUMINA.bwa. YRI.high_coverage_pcr_free.20130924.bam
NA19625	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ NA19625/high_coverage_alignment/NA19625.wgs.ILLUMINA.bwa.ASW. high_cov_pcr_free.20140203.bam
NA19648	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ NA19648/high_coverage_alignment/NA19648.wgs.ILLUMINA.bwa.MXL. high_cov_pcr_free.20140203.bam
NA20502	ftp://anonymous@ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ NA20502/high_coverage_alignment/NA20502.wgs.ILLUMINA.bwa.TSI. high_cov_pcr_free.20140203.bam

Table A3: Locations of *HLA-A*, *HLA-B* and *HLA-C* within chromosome six, and alternate loci.

	HLA-A	HLA-B	HLA-C
GRCh37	Chr6: 29 910 247 - 29 913 661	Chr6: 31 321 649 - 31 324 989	Chr6: 31 236 526 - 31 239 913
GRCh38:	Chr6: 29 942 470 - 29 945 884	Chr6: 31 353 866 - 31 357 245	Chr6: 31 268 749 - 31 272 136
chr6_GL000250v2_alt	1 200 217 - 1 203 632		
chr6_GL000251v2_alt	1 421 892 - 1 425 307	2 834 226 -2 838 475	2 749 675 - 2 753 062
chr6_GL000252v2_alt	1 144 879 - 1 201 452		2 526 549 - 2 529 926
chr6_GL000253v2_alt	1 144 120 - 1 200 442	2 662 478 - 2 665 857	2 577 801 - 2 581 178
chr6_GL000254v2_alt	1 144 513 - 1 289 979	2 695 838 - 2 699 230	2 611 478 - 2 614 855
chr6_GL000255v2_alt	1 196 245 - 1 200 858	2 609 563 - 2 613 543	2 524 181 - 2 527 558
chr6_GL000256v2_alt	1 187 504 - 1 243 714	2 656 104 - 2 659 483	2 570 707 - 2 574 084

Appendix B - Supplementary Data

Ambiguous Allele codes and the encoded Alleles

Table B3: Ambiguous *HLA-A* allele codes, and list of alleles represented by ambiguous allele code. Data obtained from the IMGT/HLA database (Robinson *et al.*, 2015).

Ambiguous allele	Alleles
*A*01:01:01G*	*A*01:01:01:01 A*01:01:01:02N A*01:01:01:03 A*01:01:01:04 A*01:01:01:05 A*01:01:01:06 A*01:01:01:07 A*01:01:01:08 A*01:01:01:09 A*01:01:01:10 A*01:01:01:11 A*01:01:01:12 A*01:01:01:13 A*01:01:01:14 A*01:01:01:15 A*01:01:01:16 A*01:01:01:17 A*01:01:01:18 A*01:01:01:19 A*01:01:01:20 A*01:01:01:21 A*01:01:38L A*01:01:51 A*01:01:83 A*01:01:84 A*01:01:91 A*01:01:93 A*01:01:94 A*01:01:95 A*01:04:01:01N A*01:04:01:02N A*01:22N A*01:32 A*01:37:01:01 A*01:37:01:02 A*01:45 A*01:56N A*01:81 A*01:87N A*01:103 A*01:107 A*01:109 A*01:132 A*01:141 A*01:142 A*01:155 A*01:177 A*01:212 A*01:217 A*01:234 A*01:237 A*01:246 A*01:248Q A*01:249 A*01:251 A*01:252 A*01:253 A*01:261 A*01:274 A*01:276 A*01:277 A*01:280 A*01:281*
*A*02:01:01G*	*A*02:01:01:01 A*02:01:01:02L A*02:01:01:03 A*02:01:01:04 A*02:01:01:05 A*02:01:01:06 A*02:01:01:07 A*02:01:01:08 A*02:01:01:09 A*02:01:01:10 A*02:01:01:11 A*02:01:01:12 A*02:01:01:13 A*02:01:01:14 A*02:01:01:15 A*02:01:01:16 A*02:01:01:17 A*02:01:01:18 A*02:01:01:19 A*02:01:01:20 A*02:01:01:21 A*02:01:01:22 A*02:01:01:23 A*02:01:01:24 A*02:01:01:25 A*02:01:01:26 A*02:01:01:27 A*02:01:01:28 A*02:01:01:29 A*02:01:01:30 A*02:01:01:31 A*02:01:01:32 A*02:01:01:33 A*02:01:01:34 A*02:01:01:35 A*02:01:01:36 A*02:01:01:37 A*02:01:01:38 A*02:01:01:39 A*02:01:01:40 A*02:01:01:41 A*02:01:01:42 A*02:01:01:43 A*02:01:01:44 A*02:01:01:45 A*02:01:01:46 A*02:01:01:47 A*02:01:01:48 A*02:01:01:49 A*02:01:01:50 A*02:01:01:51 A*02:01:01:52 A*02:01:08 A*02:01:104 A*02:01:11 A*02:01:130 A*02:01:131 A*02:01:132 A*02:01:133 A*02:01:134 A*02:01:135 A*02:01:136 A*02:01:143 A*02:01:148 A*02:01:149 A*02:01:14Q A*02:01:15 A*02:01:151 A*02:01:21 A*02:01:48 A*02:01:50 A*02:01:79 A*02:01:80 A*02:01:89 A*02:01:97 A*02:01:98 A*02:01:99 A*02:09:01:01 A*02:09:01:02 A*02:43N A*02:66 A*02:75 A*02:83N A*02:89 A*02:97:01 A*02:97:02 A*02:132 A*02:134 A*02:140 A*02:241 A*02:252 A*02:256 A*02:266 A*02:291 A*02:294 A*02:305N A*02:327 A*02:329 A*02:356N A*02:357 A*02:397 A*02:411 A*02:446 A*02:455 A*02:469 A*02:481 A*02:538 A*02:559 A*02:607 A*02:608N A*02:614 A*02:629 A*02:642 A*02:665 A*02:675N A*02:685 A*02:686 A*02:687 A*02:689 A*02:690 A*02:691N A*02:692 A*02:704 A*02:716 A*02:719 A*02:720 A*02:722 A*02:724 A*02:726 A*02:739 A*02:740 A*02:742 A*02:744 A*02:753 A*02:755 A*02:761 A*02:762 A*02:763 A*02:765 A*02:776 A*02:779*
*A*02:02:01G*	*A*02:02:01:01 A*02:02:01:02 A*02:02:01:03 A*02:02:01:04 A*02:717*

Continued on next page.

Table B3 – Continued from previous page

Ambiguous allele	Alleles
*A*02:06:01G*	*A*02:06:01:01 A*02:06:01:02 A*02:06:01:03 A*02:06:01:04 A*02:06:01:05 A*02:06:13 A*02:06:15 A*02:06:16 A*02:06:23 A*02:06:25 A*02:126 A*02:428 A*02:506N A*02:625 A*02:718 A*02:737 A*02:759 A*02:760N A*02:767 A*02:768*
*A*03:01:01G*	*A*03:01:01:01 A*03:01:01:02N A*03:01:01:03 A*03:01:01:04 A*03:01:01:05 A*03:01:01:06 A*03:01:01:07 A*03:01:01:08 A*03:01:01:09 A*03:01:01:10 A*03:01:01:11 A*03:01:01:12 A*03:01:01:13 A*03:01:01:14 A*03:01:01:15 A*03:01:01:16 A*03:01:01:17 A*03:01:01:18 A*03:01:01:19 A*03:01:01:20 A*03:01:07 A*03:01:27 A*03:01:56 A*03:01:57 A*03:01:58 A*03:01:63 A*03:01:70 A*03:01:71 A*03:01:72 A*03:01:73 A*03:01:74 A*03:01:75 A*03:01:77 A*03:20 A*03:21N A*03:26 A*03:37 A*03:45 A*03:78 A*03:112 A*03:118 A*03:129N A*03:132 A*03:134 A*03:162N A*03:182 A*03:220 A*03:279N A*03:291 A*03:292 A*03:293 A*03:301 A*03:302 A*03:304 A*03:312 A*03:313 A*03:315 A*03:316*
*A*11:01:01G*	*A*11:01:01:01 A*11:01:01:02 A*11:01:01:03 A*11:01:01:04 A*11:01:01:05 A*11:01:01:06 A*11:01:01:07 A*11:01:01:08 A*11:01:01:09 A*11:01:01:10 A*11:01:01:11 A*11:01:01:12 A*11:01:01:13 A*11:01:01:14 A*11:01:46 A*11:01:47 A*11:01:49 A*11:01:52 A*11:01:53 A*11:01:56 A*11:01:58 A*11:01:59 A*11:01:64 A*11:01:67 A*11:01:79 A*11:01:83 A*11:21N A*11:69N A*11:86 A*11:100 A*11:102 A*11:108 A*11:120 A*11:124 A*11:126 A*11:129 A*11:142 A*11:154 A*11:163 A*11:171 A*11:172 A*11:173 A*11:174 A*11:193 A*11:194 A*11:210N A*11:263 A*11:270 A*11:274 A*11:278 A*11:279 A*11:280 A*11:292 A*11:295 A*11:303 A*11:306*
*A*23:01:01G*	*A*23:01:01:01 A*23:01:01:02 A*23:01:01:03 A*23:01:01:04 A*23:01:05 A*23:01:19 A*23:01:24 A*23:01:25 A*23:07N A*23:17:01:01 A*23:17:01:02 A*23:18 A*23:20 A*23:58 A*23:85 A*23:86 A*23:87 A*23:88 A*23:91N A*23:92*
*A*24:02:01G*	*A*24:02:01:01 A*24:02:01:02L A*24:02:01:03 A*24:02:01:04 A*24:02:01:05 A*24:02:01:06 A*24:02:01:07 A*24:02:01:08 A*24:02:01:09 A*24:02:01:10 A*24:02:01:11 A*24:02:01:12 A*24:02:01:13 A*24:02:01:14 A*24:02:01:15 A*24:02:01:16 A*24:02:01:17 A*24:02:03Q A*24:02:10 A*24:02:101 A*24:02:102 A*24:02:103 A*24:02:108 A*24:02:110 A*24:02:113 A*24:02:13 A*24:02:31 A*24:02:40 A*24:02:43 A*24:02:44 A*24:02:56 A*24:02:65 A*24:02:79 A*24:02:80 A*24:02:81 A*24:02:82 A*24:02:83 A*24:02:84 A*24:02:98 A*24:09N A*24:11N A*24:40N A*24:76 A*24:79 A*24:83N A*24:144 A*24:150 A*24:153 A*24:154 A*24:155N A*24:163N A*24:183N A*24:231 A*24:249 A*24:250 A*24:251 A*24:263 A*24:264 A*24:265 A*24:266 A*24:267 A*24:268 A*24:269 A*24:270 A*24:271 A*24:352 A*24:353 A*24:354 A*24:383 A*24:385 A*24:388N A*24:400 A*24:401 A*24:402 A*24:417 A*24:418 A*24:419 A*24:422 A*24:423*
*A*25:01:01G*	*A*25:01:01:01 A*25:01:01:02 A*25:01:01:03 A*25:01:01:04 A*25:01:01:05 A*25:01:13 A*25:07*
*A*26:01:01G*	*A*26:01:01:01 A*26:01:01:02 A*26:01:01:03N A*26:01:01:04 A*26:01:01:05 A*26:01:01:06 A*26:01:01:07 A*26:01:01:08 A*26:01:01:09 A*26:01:01:10 A*26:01:01:11 A*26:01:01:12 A*26:01:07 A*26:01:25 A*26:01:32 A*26:01:35 A*26:01:40 A*26:01:43 A*26:01:44 A*26:01:46 A*26:01:47 A*26:01:49 A*26:24 A*26:26 A*26:56 A*26:82 A*26:98 A*26:99 A*26:117 A*26:157 A*26:160 A*26:162 A*26:163 A*26:164 A*26:166Q A*26:167 A*26:168*

Continued on next page.

Table B3 – Continued from previous page

Ambiguous allele	Alleles
*A*29:02:01G*	*A*29:02:01:01 A*29:02:01:02 A*29:02:01:03 A*29:02:01:04 A*29:02:01:05 A*29:02:07 A*29:02:20 A*29:02:24 A*29:02:26 A*29:26 A*29:46 A*29:75 A*29:95 A*29:100 A*29:116 A*29:119 A*29:120 A*29:121*
*A*30:01:01G*	*A*30:01:01:01 A*30:01:01:02 A*30:01:02 A*30:24 A*30:81 A*30:95 A*30:112 A*30:114 A*30:115 A*30:130N A*30:132N A*30:135 A*30:136 A*30:137 A*30:138 A*30:141 A*30:142*
*A*31:01:02G*	*A*31:01:02:01 A*31:01:02:02 A*31:01:02:03N A*31:01:02:04 A*31:01:02:05 A*31:01:02:06 A*31:01:02:07 A*31:01:02:08 A*31:01:02:09 A*31:01:02:10 A*31:01:13 A*31:01:28 A*31:14N A*31:23 A*31:46 A*31:48 A*31:55 A*31:56 A*31:59 A*31:71 A*31:72 A*31:81 A*31:95 A*31:111 A*31:119 A*31:125 A*31:128 A*31:132 A*31:135 A*31:143*
*A*68:02:01G*	*A*68:02:01:01 A*68:02:01:02 A*68:02:01:03 A*68:02:01:04 A*68:163*

Table B4: Ambiguous *HLA-B* allele codes, and list of alleles represented by ambiguous allele code. Data obtained from the IMGT/HLA database (Robinson *et al.*, 2015).

Ambiguous allele	Alleles
*B*07:02:01G*	*B*07:02:01:01 B*07:02:01:02 B*07:02:01:03 B*07:02:01:04 B*07:02:01:05 B*07:02:01:06 B*07:02:01:07 B*07:02:01:08 B*07:02:01:09 B*07:02:01:10 B*07:02:01:11 B*07:02:06 B*07:02:09 B*07:02:41 B*07:02:45 B*07:02:52 B*07:02:53 B*07:02:55 B*07:02:59 B*07:02:62 B*07:02:63 B*07:02:64 B*07:02:65 B*07:44N B*07:49N B*07:58 B*07:59 B*07:61 B*07:120 B*07:128 B*07:129 B*07:130 B*07:156 B*07:161N B*07:169 B*07:271 B*07:282 B*07:291 B*07:294 B*07:295 B*07:298 B*07:308 B*07:311 B*07:312 B*07:322 B*07:329 B*07:330N*
*B*08:01:01G*	*B*08:01:01:01 B*08:01:01:02 B*08:01:01:03 B*08:01:01:04 B*08:01:01:05 B*08:01:01:06 B*08:01:01:07 B*08:01:01:08 B*08:01:01:09 B*08:01:01:10 B*08:01:01:11 B*08:01:14 B*08:01:20 B*08:01:38 B*08:01:41 B*08:01:43 B*08:01:44 B*08:19N B*08:109 B*08:173 B*08:178 B*08:182 B*08:183 B*08:191 B*08:194 B*08:207*
*B*13:01:01G*	*B*13:01:01:01 B*13:01:01:02 B*13:01:05 B*13:01:07 B*13:01:08 B*13:01:12 B*13:52 B*13:61 B*13:109*
*B*15:16:01G*	*B*15:16:01:01 B*15:16:01:02 B*15:16:01:03*
*B*18:01:01:01*	*B*18:01:01:02 B*18:01:01:03 B*18:01:01:04 B*18:01:01:05 B*18:01:01:06 B*18:01:01:07 B*18:01:01:08 B*18:01:01:09 B*18:01:01:10 B*18:01:01:11 B*18:01:01:12 B*18:01:01:13 B*18:01:01:14 B*18:01:01:15 B*18:01:01:16 B*18:01:03 B*18:01:25 B*18:01:26 B*18:01:27 B*18:01:29 B*18:17N B*18:53 B*18:81 B*18:119 B*18:124 B*18:131:01:01 B*18:131:01:02 B*18:135 B*18:144 B*18:145 B*18:146*
*B*27:04:01G*	*B*27:04:01 B*27:04:04 B*27:68 B*27:69*

Continued on next page.

Table B4 – Continued from previous page

Ambiguous allele	Alleles
*B*35:01:01:01*	*B*35:01:01:02 B*35:01:01:03 B*35:01:01:04 B*35:01:01:05 B*35:01:01:06 B*35:01:01:07 B*35:01:01:08 B*35:01:01:09 B*35:01:01:10 B*35:01:01:11 B*35:01:01:12 B*35:01:01:13 B*35:01:01:14 B*35:01:01:15 B*35:01:03 B*35:01:23 B*35:01:25 B*35:01:28 B*35:01:40 B*35:01:41 B*35:01:47 B*35:40N B*35:42:01 B*35:57 B*35:94 B*35:134N B*35:161 B*35:227 B*35:241 B*35:245 B*35:250 B*35:332 B*35:336 B*35:347 B*35:348 B*35:359 B*35:365 B*35:370 B*35:376 B*35:380 B*35:383*
*B*35:02:01G*	*B*35:02:01:01 B*35:02:01:02 B*35:02:01:03 B*35:02:05 B*35:220 B*35:379*
*B*35:03:01G*	*B*35:03:01:01 B*35:03:01:02 B*35:03:01:03 B*35:03:01:04 B*35:03:01:05 B*35:03:01:06 B*35:03:01:07 B*35:03:01:08 B*35:03:01:09 B*35:03:01:10 B*35:03:13 B*35:03:23 B*35:70 B*35:279 B*35:298 B*35:344 B*35:364 B*35:371*
*B*38:01:01G*	*B*38:01:01:01 B*38:01:01:02 B*38:68Q*
*B*40:01:01G*	*B*40:01:01 B*40:01:02:01 B*40:01:02:02 B*40:01:02:03 B*40:01:02:04 B*40:01:02:05 B*40:01:02:06 B*40:01:02:07 B*40:01:02:08 B*40:01:02:09 B*40:01:25 B*40:01:36 B*40:01:37 B*40:01:42 B*40:01:45 B*40:01:48 B*40:01:52 B*40:01:54 B*40:01:55 B*40:01:57 B*40:01:58 B*40:55 B*40:141 B*40:150 B*40:151 B*40:179 B*40:221 B*40:236 B*40:241 B*40:247 B*40:264 B*40:272 B*40:278 B*40:299 B*40:301 B*40:329 B*40:338N B*40:353 B*40:383 B*40:386 B*40:395*
*B*44:02:01:01*	*B*44:02:01:02S B*44:02:01:03 B*44:02:01:04 B*44:02:01:05 B*44:02:01:06 B*44:02:01:07 B*44:02:01:08 B*44:02:01:09 B*44:02:01:10 B*44:02:01:11 B*44:02:01:12 B*44:02:01:13 B*44:02:25 B*44:02:27 B*44:02:46 B*44:02:49 B*44:02:52 B*44:02:53 B*44:19N B*44:27:01:01 B*44:27:01:02 B*44:66 B*44:118 B*44:187 B*44:243 B*44:262 B*44:267N B*44:270 B*44:279 B*44:292*
*B*44:03:01G*	*B*44:03:01:01 B*44:03:01:02 B*44:03:01:03 B*44:03:01:04 B*44:03:01:05 B*44:03:01:06 B*44:03:01:07 B*44:03:01:08 B*44:03:01:09 B*44:03:01:10 B*44:03:01:11 B*44:03:01:12 B*44:03:01:13 B*44:03:03 B*44:03:04 B*44:03:39 B*44:278 B*44:280 B*44:281*
*B*51:01:01G*	*B*51:01:01:01 B*51:01:01:02 B*51:01:01:03 B*51:01:01:04 B*51:01:01:05 B*51:01:01:06 B*51:01:01:07 B*51:01:01:08 B*51:01:01:09 B*51:01:01:10 B*51:01:01:11 B*51:01:01:12 B*51:01:01:13 B*51:01:01:14 B*51:01:01:15 B*51:01:01:16 B*51:01:01:17 B*51:01:01:18 B*51:01:01:19 B*51:01:01:20 B*51:01:01:21 B*51:01:01:22 B*51:01:01:23 B*51:01:01:24 B*51:01:01:25 B*51:01:01:26 B*51:01:01:27 B*51:01:01:28 B*51:01:01:29 B*51:01:05 B*51:01:07 B*51:01:23 B*51:01:35 B*51:01:44 B*51:01:45 B*51:01:55 B*51:01:56 B*51:01:57 B*51:01:60 B*51:01:61 B*51:01:64 B*51:11N B*51:30 B*51:32 B*51:48 B*51:51 B*51:142 B*51:164 B*51:165 B*51:166 B*51:169 B*51:193 B*51:219 B*51:224 B*51:229 B*51:230 B*51:232 B*51:234 B*51:237 B*51:248 B*51:249 B*51:250*
*B*52:01:02G*	*B*52:01:02:01 B*52:01:02:02 B*52:01:02:03*
*B*53:01:01:01*	*B*53:01:01:02 B*53:01:12 B*53:01:15 B*53:37 B*53:51*
*B*57:03:01G*	*B*57:03:01:01 B*57:03:01:02 B*57:03:01:03 B*57:94 B*57:101*

Table B5: Ambiguous *HLA-C* allele codes, and list of alleles represented by ambiguous allele code. Data obtained from the IMGT/HLA database (Robinson *et al.*, 2015).

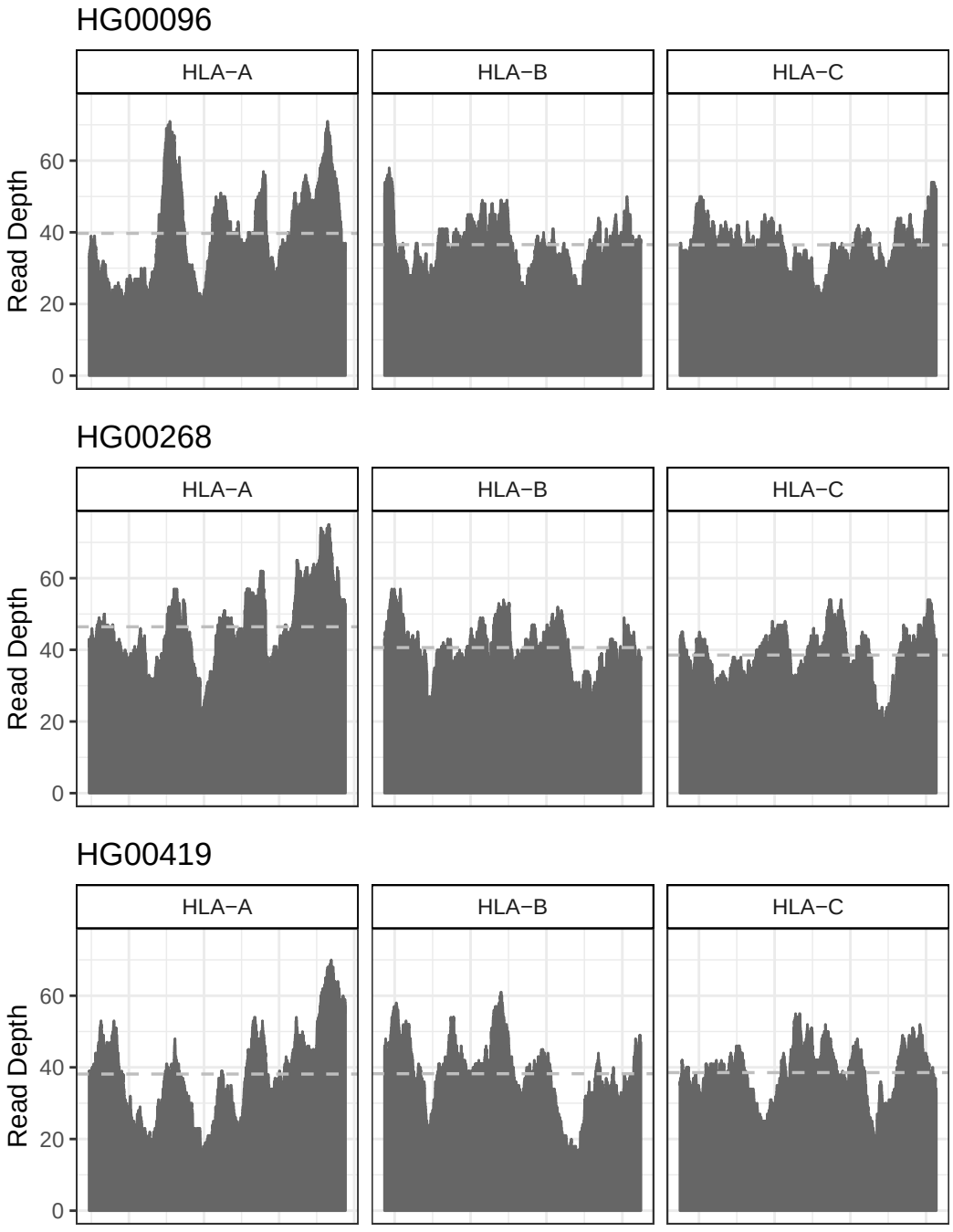
Ambiguous allele	Alleles
*C*01:02:01G*	*C*01:02:01:01 C*01:02:01:02 C*01:02:01:03 C*01:02:01:04 C*01:02:01:05 C*01:02:01:06 C*01:02:01:07 C*01:02:01:08 C*01:02:01:09 C*01:02:01:10 C*01:02:01:11 C*01:02:02 C*01:02:11 C*01:02:12 C*01:02:14 C*01:02:15 C*01:02:23 C*01:02:24 C*01:02:25 C*01:02:30 C*01:02:42 C*01:02:45 C*01:25 C*01:44 C*01:82 C*01:83 C*01:84 C*01:85 C*01:89N C*01:109N C*01:127 C*01:135 C*01:138 C*01:139 C*01:142 C*01:150 C*01:151 C*01:155 C*01:159 C*01:164 C*01:165*
*C*03:04:01G*	*C*03:04:01:01 C*03:04:01:02 C*03:04:01:03 C*03:04:01:04 C*03:04:01:05 C*03:04:01:06 C*03:04:01:07 C*03:04:01:08 C*03:04:01:09 C*03:04:01:10 C*03:04:03 C*03:04:20 C*03:04:36 C*03:04:43 C*03:04:44 C*03:04:55 C*03:04:58 C*03:04:61 C*03:04:63 C*03:100 C*03:101 C*03:105 C*03:106 C*03:211:01 C*03:211:02 C*03:212 C*03:213 C*03:218 C*03:219 C*03:236 C*03:252 C*03:294 C*03:303 C*03:354 C*03:358 C*03:359 C*03:366N C*03:369 C*03:376 C*03:381 C*03:387 C*03:408 C*03:417 C*03:423*
*C*04:01:01G*	*C*04:01:01:01 C*04:01:01:02 C*04:01:01:03 C*04:01:01:04 C*04:01:01:05 C*04:01:01:06 C*04:01:01:07 C*04:01:01:08 C*04:01:01:09 C*04:01:01:10 C*04:01:01:11 C*04:01:01:12 C*04:01:01:13 C*04:01:01:14 C*04:01:01:15 C*04:01:01:16 C*04:01:01:17 C*04:01:01:18 C*04:01:01:19 C*04:01:01:20 C*04:01:01:21 C*04:01:01:22 C*04:01:01:23 C*04:01:01:24 C*04:01:01:25 C*04:01:01:26 C*04:01:102 C*04:01:54 C*04:01:57 C*04:01:69 C*04:01:78 C*04:01:79 C*04:01:82 C*04:01:83 C*04:01:85 C*04:01:86 C*04:01:92 C*04:01:93 C*04:01:94 C*04:01:95 C*04:01:96 C*04:01:97 C*04:09N C*04:28 C*04:30 C*04:41 C*04:79 C*04:82 C*04:84 C*04:106 C*04:144 C*04:146 C*04:161 C*04:162 C*04:165 C*04:195 C*04:226 C*04:267 C*04:274 C*04:275 C*04:277 C*04:287 C*04:289 C*04:295 C*04:298 C*04:306 C*04:308 C*04:310 C*04:318 C*04:320 C*04:321 C*04:322 C*04:327 C*04:328 C*04:329 C*04:330*
*C*05:01:01G*	*C*05:01:01:01 C*05:01:01:02 C*05:01:01:03 C*05:01:01:04 C*05:01:01:05 C*05:01:01:06 C*05:01:01:07 C*05:01:01:08 C*05:01:01:09 C*05:01:01:10 C*05:01:01:11 C*05:01:01:12 C*05:01:01:13 C*05:01:01:14 C*05:01:01:15 C*05:01:04 C*05:01:05 C*05:01:15 C*05:01:35 C*05:01:36 C*05:01:37 C*05:01:38 C*05:01:41 C*05:03 C*05:37 C*05:53 C*05:93 C*05:108 C*05:145 C*05:153N C*05:158 C*05:161 C*05:172 C*05:179 C*05:187*

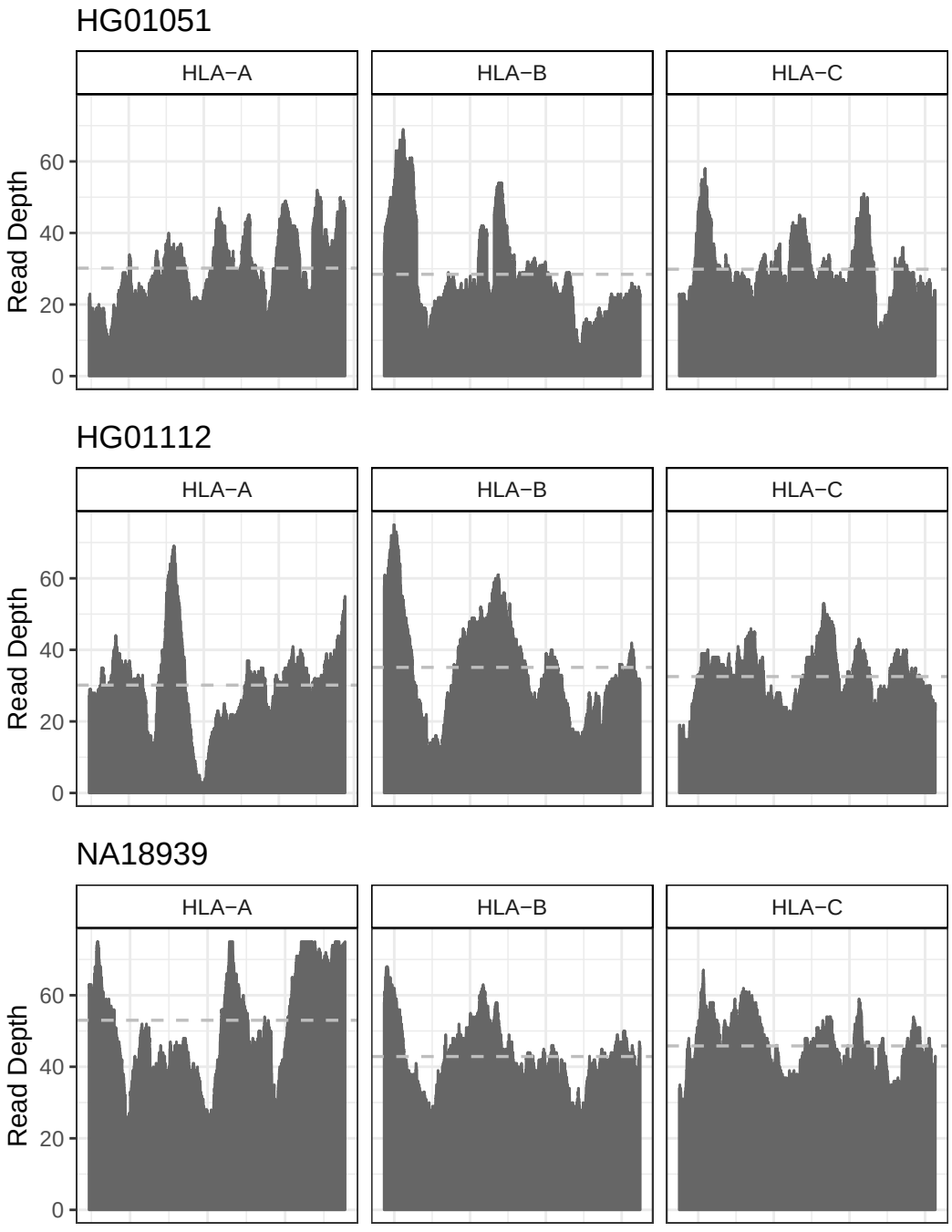
Continued on next page.

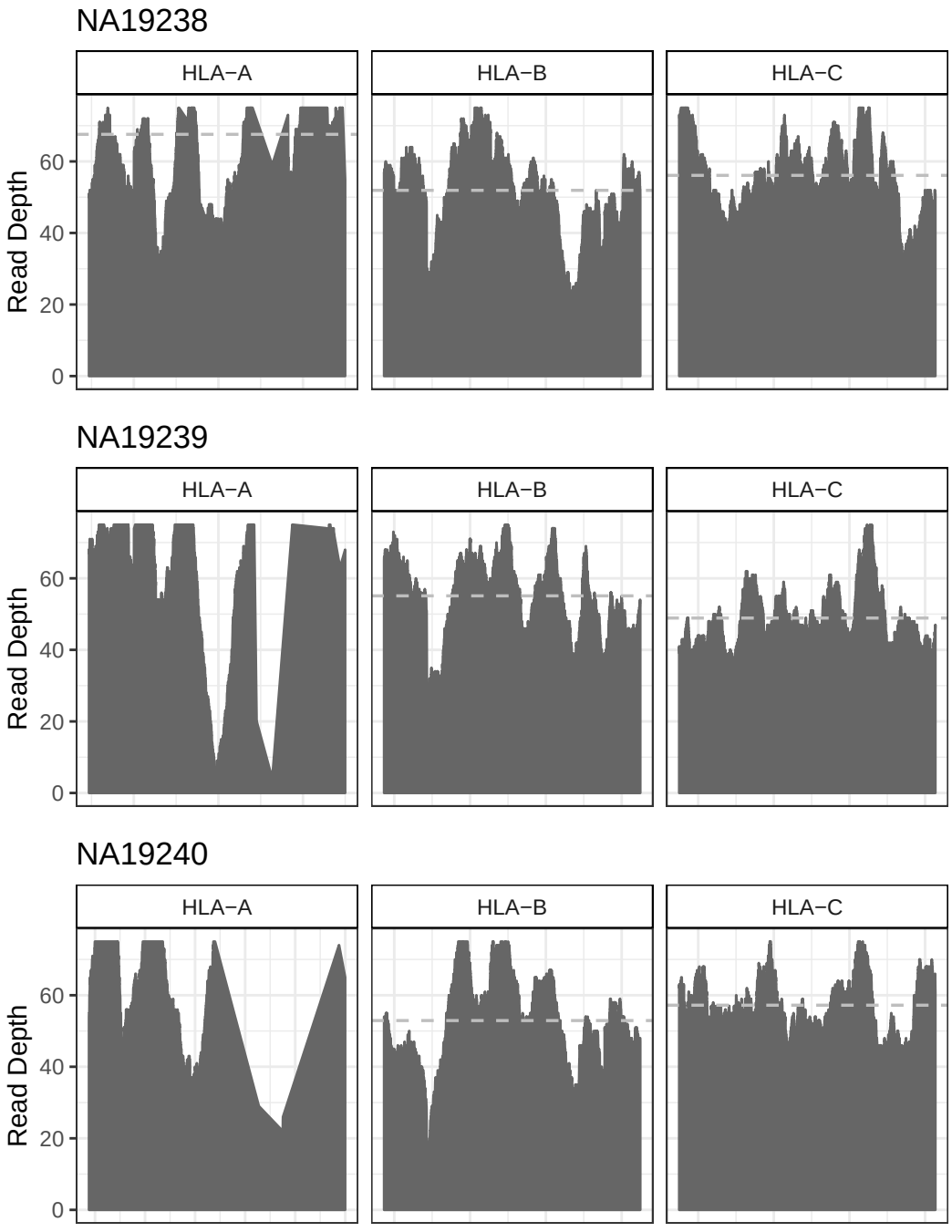
Table B5 – Continued from previous page

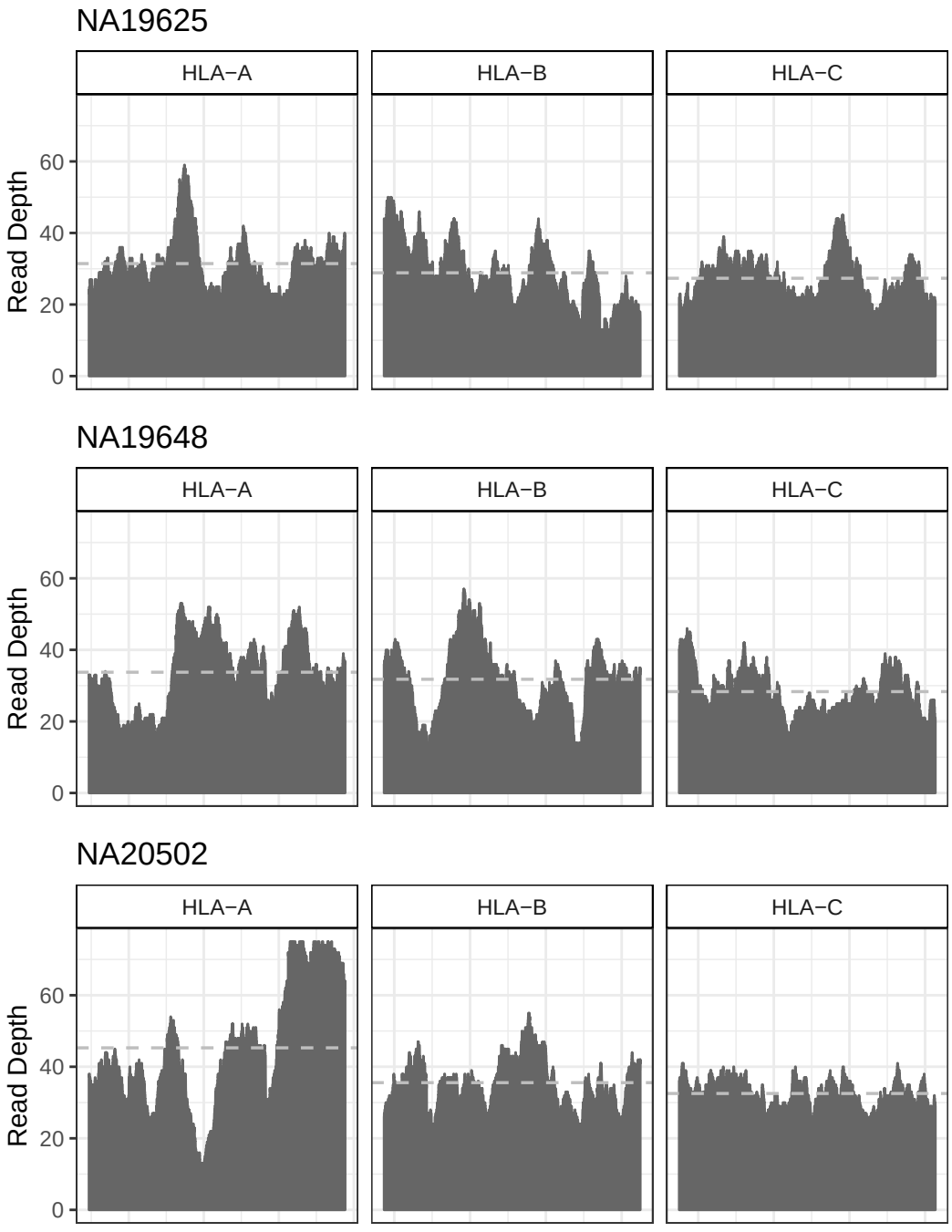
Ambiguous allele	Alleles
*C*07:01:01G*	*C*07:01:01:01 C*07:01:01:02 C*07:01:01:03 C*07:01:01:04 C*07:01:01:05 C*07:01:01:06 C*07:01:01:07 C*07:01:01:08 C*07:01:01:09 C*07:01:01:10 C*07:01:01:11 C*07:01:01:12 C*07:01:01:13 C*07:01:01:14Q C*07:01:01:15 C*07:01:01:16 C*07:01:01:17 C*07:01:01:18 C*07:01:01:19 C*07:01:01:20 C*07:01:01:21 C*07:01:01:22 C*07:01:01:23 C*07:01:01:24 C*07:01:01:25 C*07:01:01:26 C*07:01:01:27 C*07:01:01:28 C*07:01:02 C*07:01:09 C*07:01:19 C*07:01:39 C*07:01:61 C*07:06:01:01 C*07:06:01:02 C*07:18:01:01 C*07:18:01:02 C*07:52 C*07:153 C*07:166 C*07:337 C*07:343:01:01 C*07:343:01:02 C*07:419 C*07:458 C*07:588 C*07:591 C*07:607 C*07:610 C*07:615 C*07:617 C*07:618 C*07:619 C*07:621 C*07:623 C*07:624 C*07:657 C*07:658 C*07:682 C*07:685 C*07:687 C*07:694 C*07:696*
*C*07:02:01G*	*C*07:02:01:01 C*07:02:01:02 C*07:02:01:03 C*07:02:01:04 C*07:02:01:05 C*07:02:01:06 C*07:02:01:07 C*07:02:01:08 C*07:02:01:09 C*07:02:01:10 C*07:02:01:11 C*07:02:01:12 C*07:02:01:13 C*07:02:01:14 C*07:02:01:15 C*07:02:01:16 C*07:02:01:17N C*07:02:01:18 C*07:02:01:19 C*07:02:01:20 C*07:02:01:21 C*07:02:01:22 C*07:02:01:23 C*07:02:01:24 C*07:02:01:25 C*07:02:01:26 C*07:02:01:27 C*07:02:01:28 C*07:02:01:29 C*07:02:01:30 C*07:02:01:31 C*07:02:103 C*07:02:21 C*07:02:23 C*07:02:50 C*07:02:51 C*07:02:53 C*07:02:60 C*07:02:70 C*07:02:79 C*07:02:80 C*07:02:82 C*07:02:85 C*07:02:93 C*07:02:95 C*07:02:96 C*07:50 C*07:66 C*07:74 C*07:159 C*07:160 C*07:167 C*07:245 C*07:308 C*07:348 C*07:349 C*07:350N C*07:359 C*07:446 C*07:486 C*07:500 C*07:533 C*07:544 C*07:566 C*07:592 C*07:593N C*07:594 C*07:595 C*07:596 C*07:608 C*07:612 C*07:661 C*07:665 C*07:666 C*07:667 C*07:675N C*07:676 C*07:684 C*07:686N C*07:688*
*C*12:02:01G*	*C*12:02:01 C*12:02:02:01 C*12:02:02:02 C*12:02:02:03 C*12:02:10 C*12:228 C*12:234 C*12:243*
*C*12:03:01G*	*C*12:03:01:01 C*12:03:01:02 C*12:03:01:03 C*12:03:01:04 C*12:03:01:05 C*12:03:01:06 C*12:03:01:07 C*12:03:01:08 C*12:03:01:09 C*12:03:01:10 C*12:03:01:11 C*12:03:01:12 C*12:03:01:13 C*12:03:06 C*12:03:43 C*12:03:49 C*12:23 C*12:109 C*12:110 C*12:111 C*12:125 C*12:143 C*12:160 C*12:167 C*12:171 C*12:172 C*12:201 C*12:209 C*12:210 C*12:211 C*12:216 C*12:220 C*12:223 C*12:244 C*12:245 C*12:253 C*12:254*
*C*14:02:01G*	*C*14:02:01:01 C*14:02:01:02 C*14:02:01:03 C*14:02:01:04 C*14:02:01:05 C*14:02:01:06 C*14:02:01:07 C*14:02:01:08 C*14:02:07 C*14:02:15 C*14:02:16 C*14:02:22 C*14:02:26 C*14:02:29 C*14:23 C*14:31 C*14:57 C*14:60 C*14:100*
*C*16:01:01G*	*C*16:01:01:01 C*16:01:01:02 C*16:01:01:03 C*16:01:01:04 C*16:01:01:05 C*16:01:01:06 C*16:01:24 C*16:58 C*16:97 C*16:100 C*16:111 C*16:112 C*16:118 C*16:137*
*C*18:01:01G*	*C*18:01 C*18:02:01 C*18:11*

Supplementary Figure B1: Coverage plot of the total read depth across *HLA-A*, *HLA-B* and *HLA-C* in the 12 individuals for whom both high-resolution SBT *HLA* genotype and high coverage WGS data were available. Counts were obtained from reads aligning to the GRCh38 reference assembly and alternate loci




HG01051
HLA−A
HLA−B
HLA−C
Read Depth
0
20
40
60
HG01112
HLA−A
HLA−B
HLA−C
Read Depth
0
20
40
60
NA18939
HLA−A
HLA−B
HLA−C
Read Depth
0
20
40
60


NA19238
HLA−A
HLA−B
HLA−C
Read Depth
0
20
40
60
NA19239
HLA−A
HLA−B
HLA−C
Read Depth
0
20
40
60
NA19240
HLA−A
HLA−B
HLA−C
Read Depth
0
20
40
60


NA19625
HLA−A
HLA−B
HLA−C
Read Depth
60
40
20
0
NA19648
HLA−A
HLA−B
HLA−C
Read Depth
60
40
20
0
NA20502
HLA−A
HLA−B
HLA−C
Read Depth
60
40
20
0

Genotyping Accuracy across 12 samples by four tools: BWAkit, xHLA, Kourami and HISAT-Genotype

Table B6: Overall *HLA* genotyping accuracy of four tools on 12 individuals at two, four and six-digit resolutions.

Tool	Call Rate	Accuracy		
		2-digit	4-digit	6-digit
BWAkit	68/72	86.8(59/68)	66.2(45/68)	60.3(41/68)
xHLA	72/72	100(72/72)	98.6(71/72)	94.4(68/72)
HISAT-Genotype	72/72	94.4(68/72)	90.3(65/72)	87.5(63/72)
Kourami	72/72	98.6(71/72)	97.2(70/72)	95.8(69/72)

Accuracy shown as a percentage of correctly typed alleles and total number of alleles tested is shown in parenthesis.

Table B7: *HLA-A* genotyping accuracy on 12 individuals at two, four and six-digit resolutions.

Tool	Call Rate	Accuracy		
		2-digit	4-digit	6-digit
BWAkit	24/24	79.2(19/24)	62.5(15/24)	54.2(13/24)
xHLA	24/24	100(24/24)	100(24/24)	100(24/24)
HISAT-Genotype	24/24	91.7(22/24)	87.5(21/24)	87.5(21/24)
Kourami	24/24	100(24/24)	95.8(23/24)	95.8(23/24)

Accuracy shown as a percentage of correctly typed alleles and total number of alleles tested is shown in parenthesis.

Table B8: *HLA-B* genotyping accuracy on 12 individuals at two, four and six-digit resolutions.

Tool	Call Rate	Accuracy		
		2-digit	4-digit	6-digit
BWAkit	20/24	100(20/20)	85.0(17/20)	85.0(17/20)
xHLA	24/24	100(24/24)	100(24/24)	100(24/24)
HISAT-Genotype	24/24	100(24/24)	91.7(22/24)	83.3(20/24)
Kourami	24/24	100(24/24)	100(24/24)	100(24/24)

Accuracy shown as a percentage of correctly typed alleles and total number of alleles tested is shown in parenthesis.

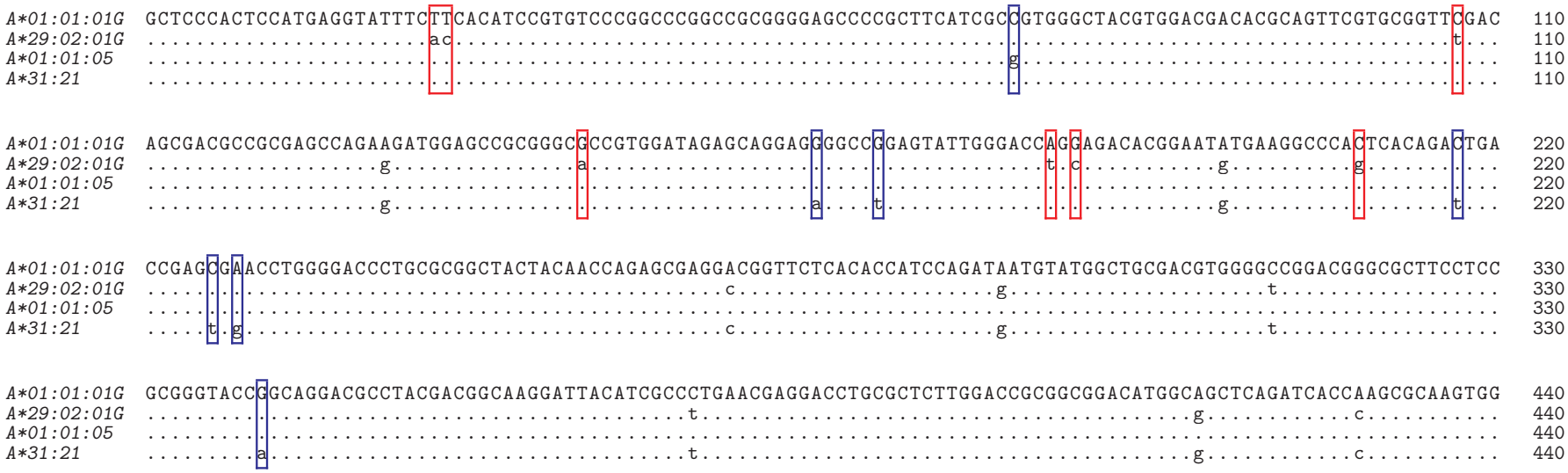
Table B9: *HLA-C* Typing accuracy on 12 individuals at two, four and six-digit resolutions.

Tool	Call Rate	Accuracy		
		2-digit	4-digit	6-digit
BWAkit	24/24	83.3(20/24)	54.2(13/24)	45.8(11/24)
xHLA	24/24	100(24/24)	95.8(23/24)	83.3(20/24)
HISAT-Genotype	24/24	91.7(22/24)	95.5(22/24)	95.5(22/24)
Kourami	24/24	95.8(23/24)	95.8(23/24)	91.7(22/24)

Accuracy shown as a percentage of correctly typed alleles and total number of alleles tested is shown in parenthesis.

Sequence Alignments of Incorrectly assigned Alleles

BWAkit



Supplementary Figure B5: Multiple sequence alignment of *HLA-A* alleles incorrectly assigned by BWAkit for individual HG00096. SBT Alleles: *A*01:01:01G & A*29:02:01G*. BWAkit Alleles: *A*01:01:05 & A*31:21*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Nucleotides enclosed in blue indicate an incorrect heterozygous nucleotide call at that position, where the tool indicated a homozygous position as heterozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B6: Multiple sequence alignment of *HLA-B* alleles incorrectly assigned by BWAkit for individual HG00096. SBT Alleles: *B*08:01:01G* & *B*44:03:01G*. BWAkit Alleles: *B*08:50* & *B*44:03:01G*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

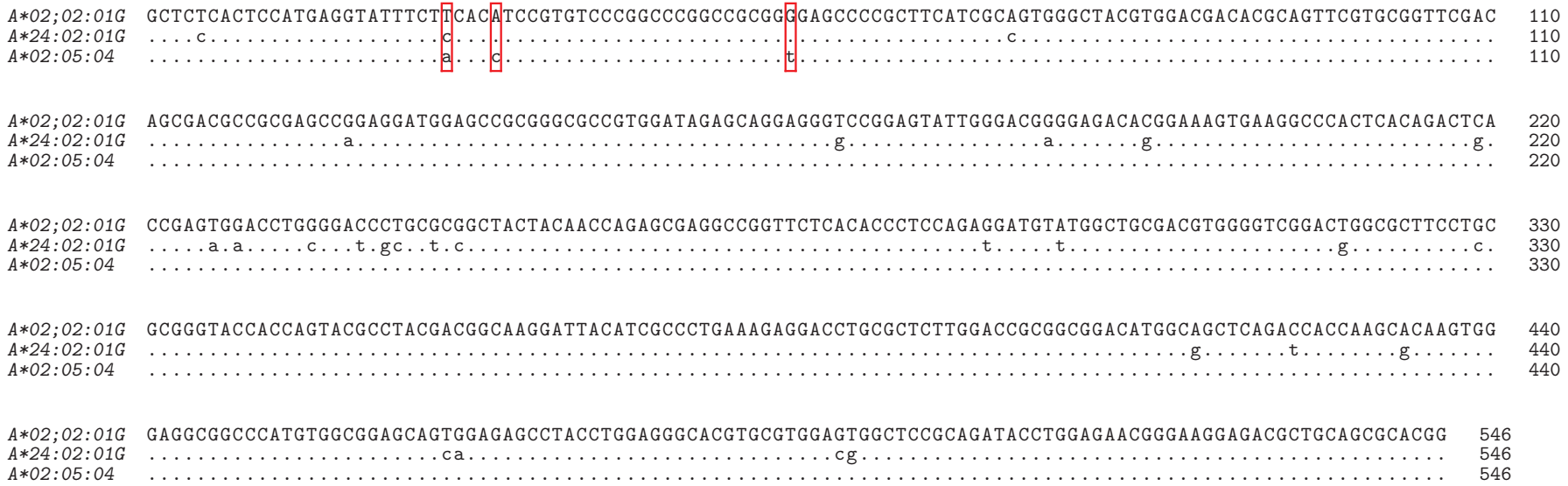
Supplementary Figure B7: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual HG00096. SBT Alleles: *C*07:01:01G* & *C*16:01:01G*. BWAkit Alleles: *C*07:01:01G* & *C*16:21*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B8: Multiple sequence alignment of *HLA-A* alleles incorrectly assigned by BWAkit for individual HG00268. SBT Alleles: *A*03:01:01G & A*25:01:01G*. BWAkit Alleles: *A*03:12 & A*25:01:01G*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B9: Multiple sequence alignment of *HLA-B* alleles incorrectly assigned by BWAkit for individual HG00268. SBT Alleles: *B*07:02:01G* & *B*18:01:01G*. BWAkit Alleles: *B*07:07* & *B*18:01:01G*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B10: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual HG00268. SBT Alleles: *C*07:02:01G* & *C*12:03:01G*. BWAkit Alleles: *C*07:31:01* & *C*12:03:01G*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B11: Multiple sequence alignment of *HLA-B* alleles incorrectly assigned by BWAkit for individual HG00419. SBT Alleles: *B*12:01:01G* & *B*40:01:01G*. BWAkit Alleles: *B*13:01:01G* & *B*40:49*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.



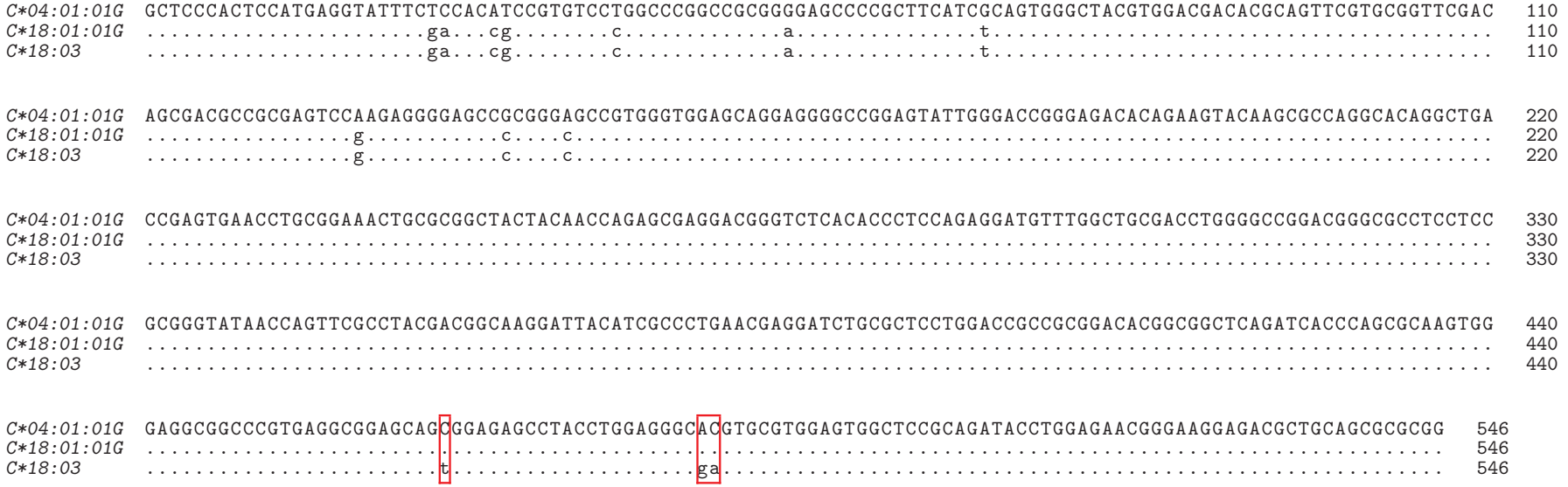
Supplementary Figure B12: Multiple sequence alignment of *HLA-A* alleles incorrectly assigned by BWAkit for individual HG01051. SBT Alleles: *A*02:02:01G & A*24:02:01G*. BWAkit Alleles: Homozygous *A*02:05:04*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B13: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual HG01051. SBT Alleles: *C*12:03:01G* & *C*14:02:01G*. BWAkit Alleles: Homozygous *C*12:03:04*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

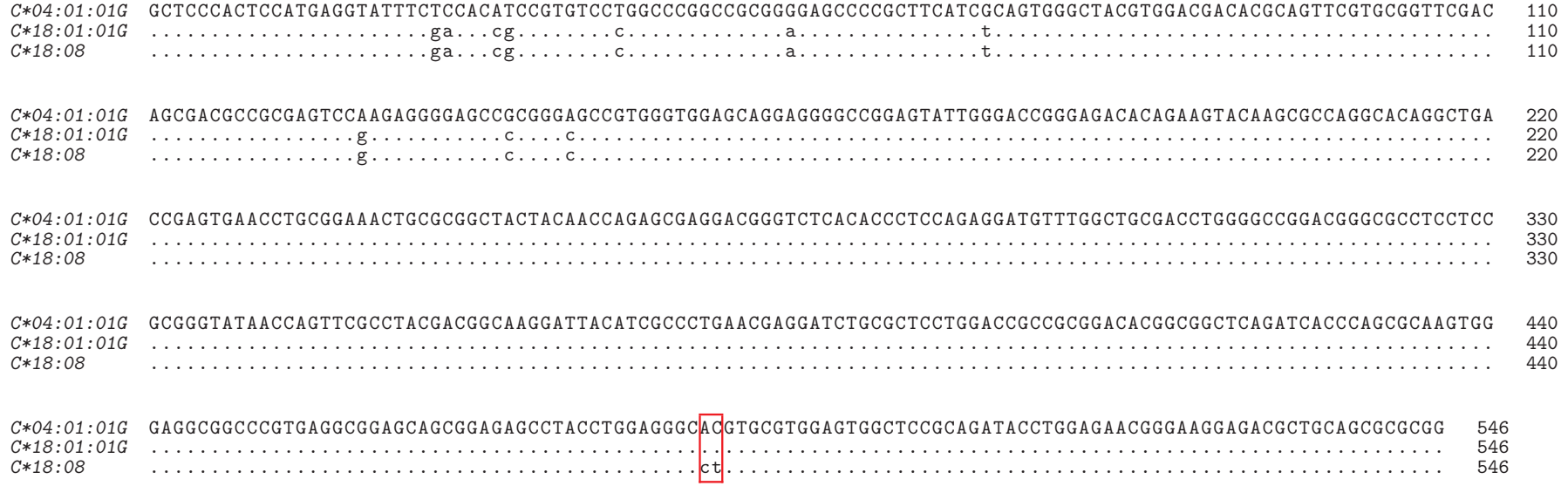
Supplementary Figure B14: Multiple sequence alignment of *HLA-A* alleles incorrectly assigned by BWAkit for individual HG01112. SBT Alleles: *A*02:01:01G* & *A*26:01:01G*. BWAkit Alleles: Homozygous *A*02:55*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B15: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual HG01112. SBT Alleles: *C*05:01:01G & C*12:03:01G*. BWAkit Alleles: *C*05:52 & C*08:25*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

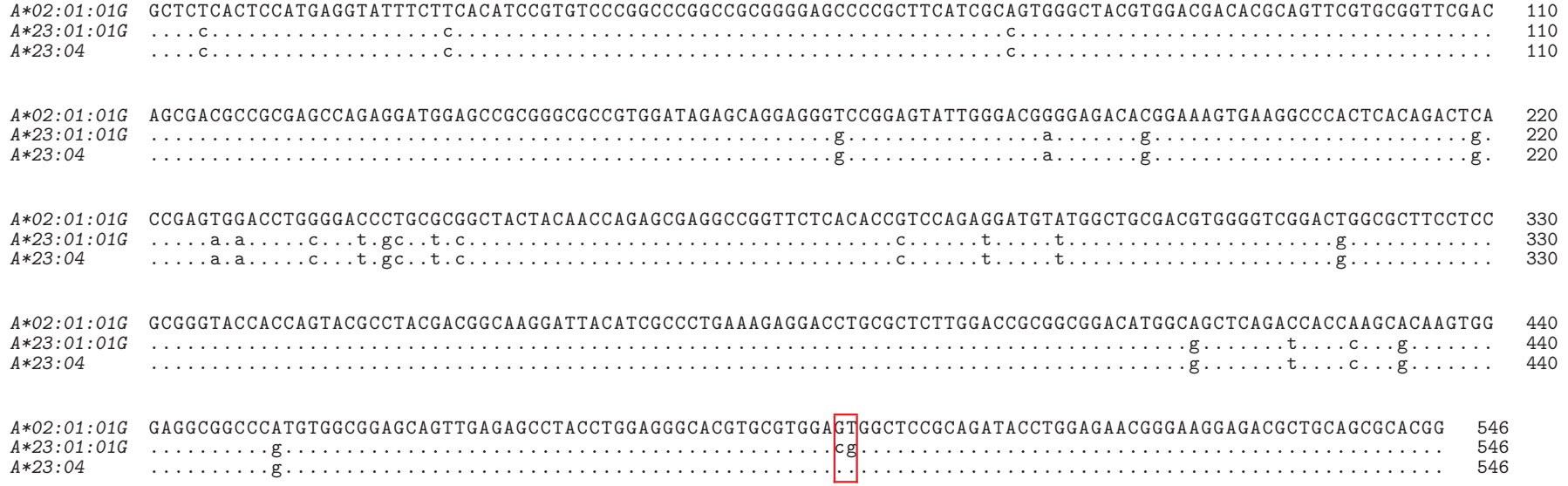
Supplementary Figure B16: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual NA18939. SBT Alleles: *C*07:02:01G* & *C*12:03:01G*. BWAkit Alleles: *C*07:314* & *C*12:03:01G*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.



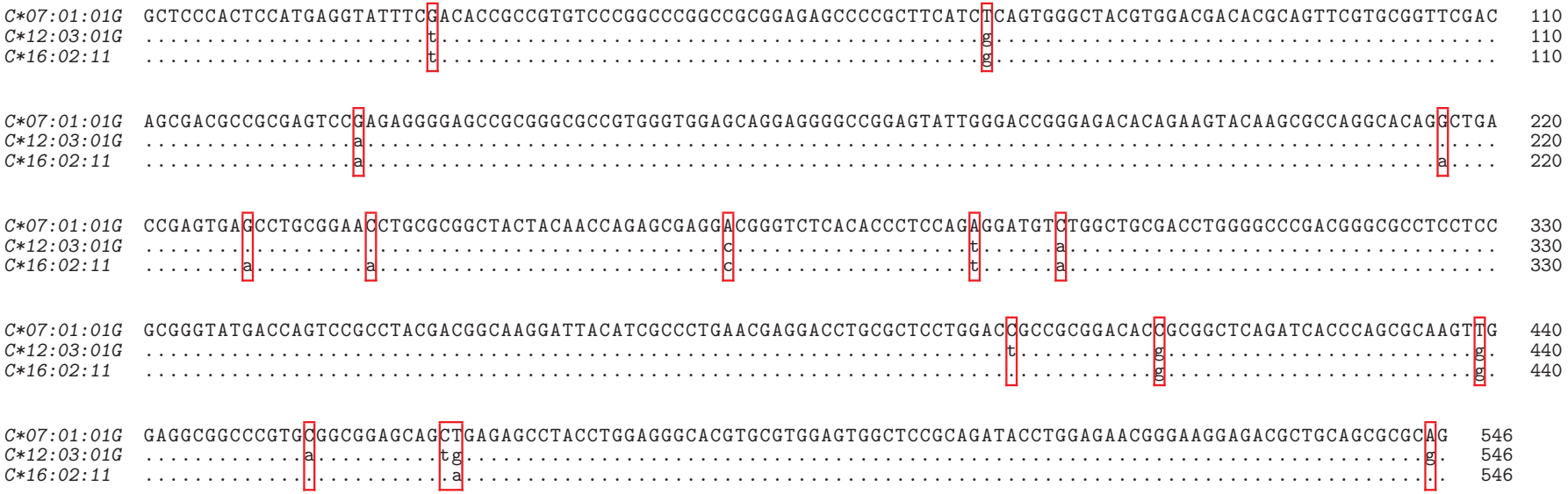
Supplementary Figure B17: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual NA19238. SBT Alleles: *C*04:01:01G* & *C*18:01:01G*. BWAkit Alleles: *C*04:01:01G* & *C*18:03*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.



Supplementary Figure B18: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual NA19240. SBT Alleles: *C*04:01:01G* & *C*18:01:01G*. BWAkit Alleles: *C*04:01:01G* & *C*18:08*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.



Supplementary Figure B19: Multiple sequence alignment of *HLA-A* alleles incorrectly assigned by BWAkit for individual NA19625. SBT Alleles: *A*02:01:01G & A*23:01:01G*. BWAkit Alleles: Homozygous *A*23:04:01*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

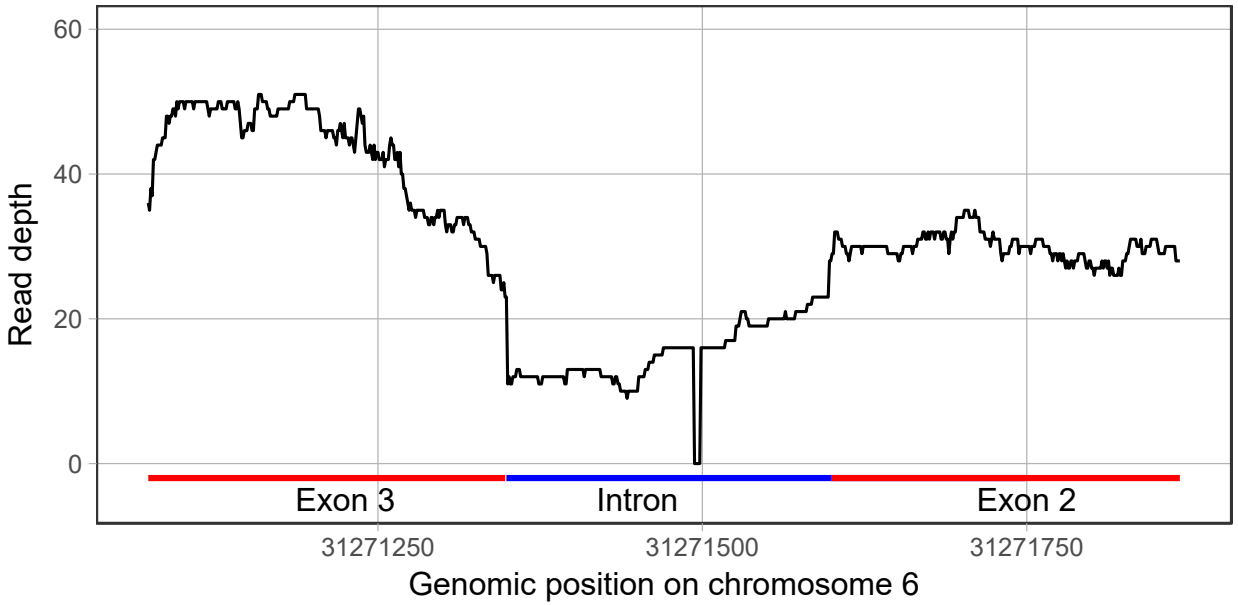


Supplementary Figure B20: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual NA19625. SBT Alleles: *C*07:01:01G* & *C*12:03:01G*. BWAkit Alleles: Homozygous *C*16:02:11*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

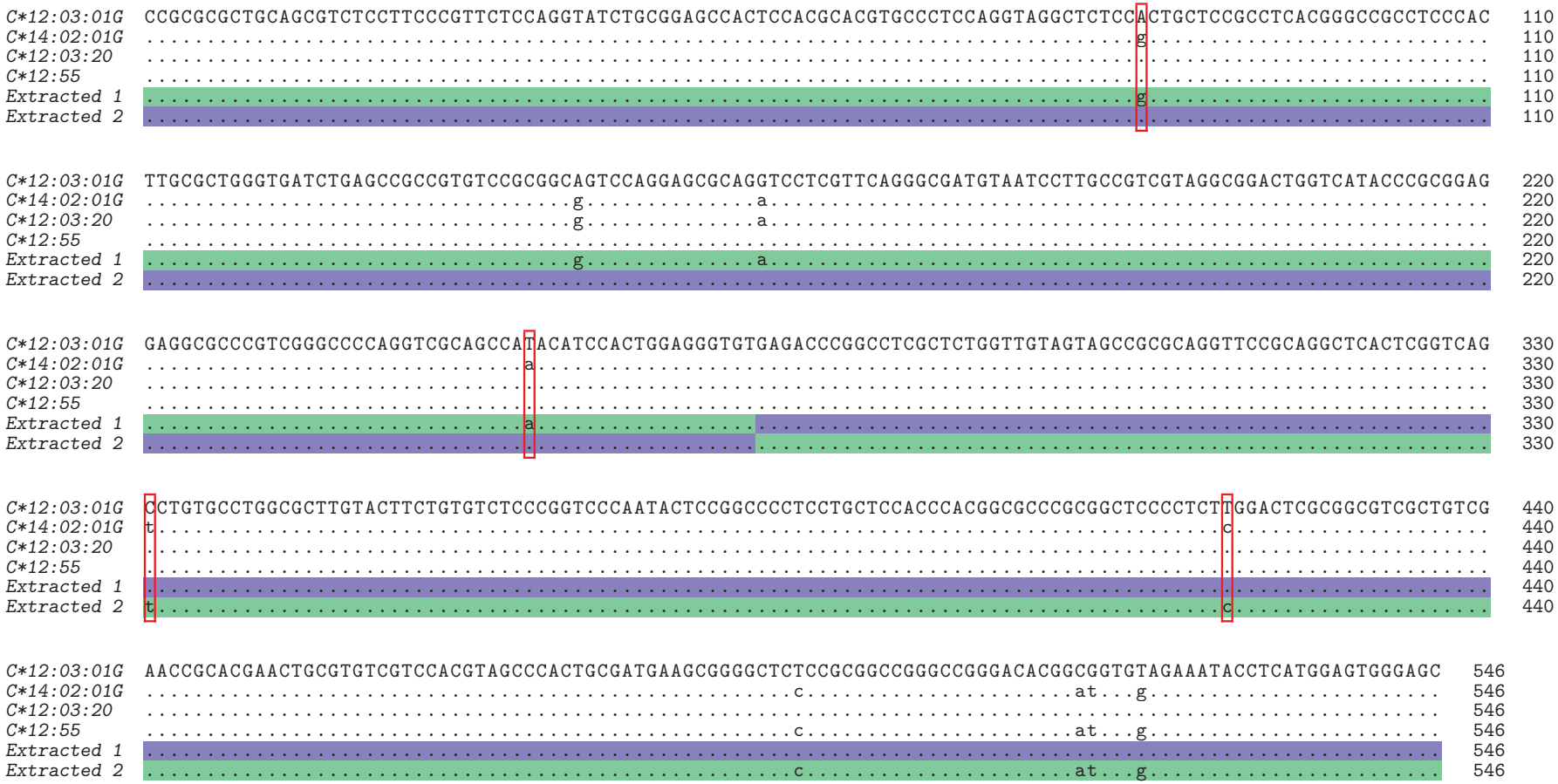
Supplementary Figure B21: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by BWAkit for individual NA19648. SBT Alleles: *C*01:02:01G* & *C*07:02:01G*. BWAkit Alleles: *C*01:02:01G* & *C*07:29*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B22: Multiple sequence alignment of *HLA-A* alleles incorrectly assigned by BWAkit for individual NA20502. SBT Alleles: *A*01:01:01G & A*31:01:02G*. BWAkit Alleles: *A*01:143 & A*31:01:02G*. Nucleotides enclosed in red indicate an incorrect homozygous call at that position, in which the tool represented a heterozygous position as homozygous. Sequence alignment consists of coding sequences for exons 2 and 3.

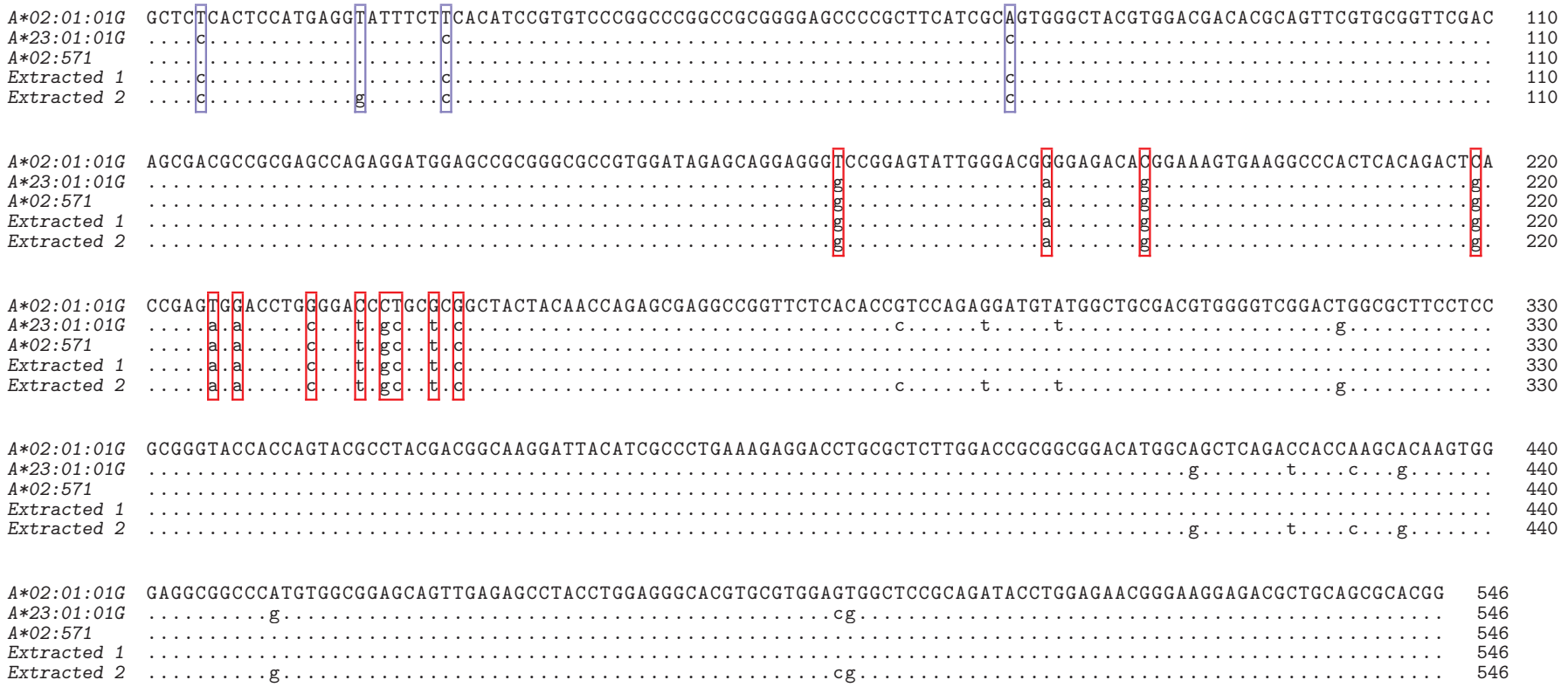
Kourami



Supplementary Figure B23: Read depth of HG01051 across exon 2 and exon 3 of *HLA-C*



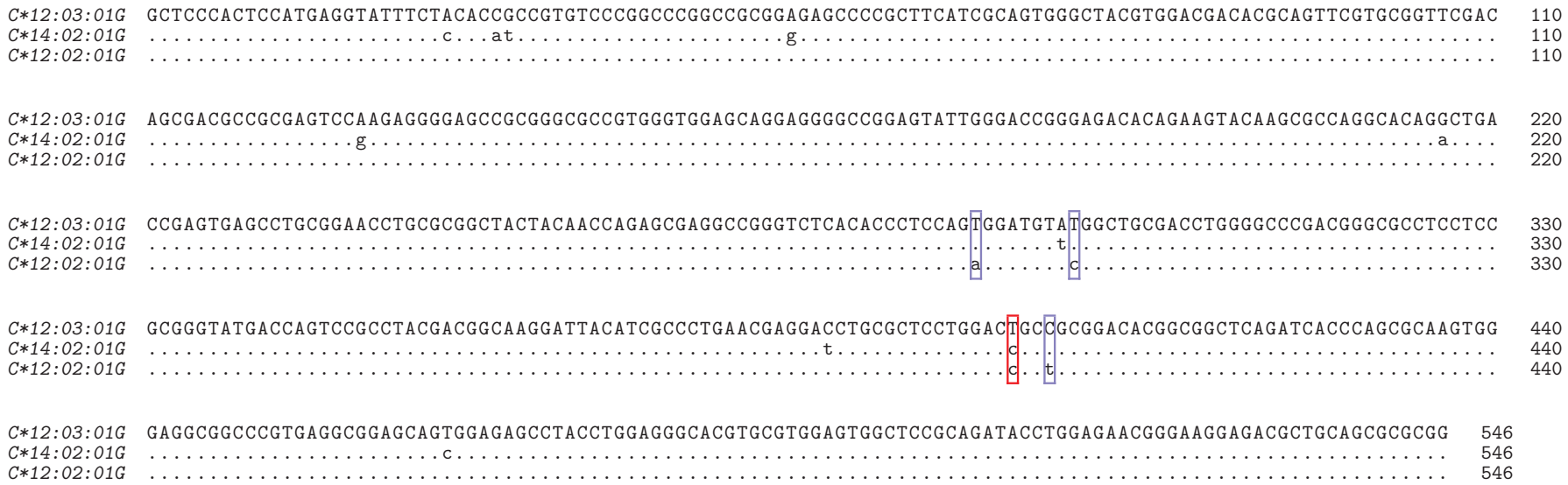
Supplementary Figure B24: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by Kourami for individual HG01051. SBT Alleles: *C*12:03:01G & C*14:02:01G*. Kourami Alleles: *C*12:03:20 & C*12:55*. Nucleotides shaded in blue correspond to SBT allele - *C*12:03:01G*. Nucleotides shaded in green correspond to the SBT allele - *C*14:02:01G*. Nucleotides enclosed in red indicate a variant contained in the extracted sequences but not the genotyped alleles. Sequence alignment consists of coding sequences for exons 2 and 3.



Supplementary Figure B25: Multiple sequence alignment of *HLA-A* alleles incorrectly assigned by Kourami for individual NA19625. SBT Alleles: *A*02:01:01G & A*23:01:01G*. Kourami Alleles: *A*02:571 & A*23:01:01G*. Nucleotides enclosed in red indicate incorrect homozygous nucleotide calls instead of a heterozygous call. Nucleotides enclosed in blue indicate incorrect homozygous nucleotide calls in the extracted sequences but not in the genotyped alleles. Sequence alignment consists of coding sequences for exons 2 and 3.

xHLA

Supplementary Figure B26: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by xHLA for individual HG00096. SBT Alleles: *C*07:01:01G* & *C*16:01:01G*. xHLA Alleles: *C*07:01:05* & *C*16:01:01G*. Nucleotides enclosed in red indicate positions which were incorrectly genotyped for both alleles called. Nucleotides shaded in grey indicate a homozygous genotype miscall instead of a heterozygous genotype call. Sequence alignment consists of coding sequences for exons 2 and 3.



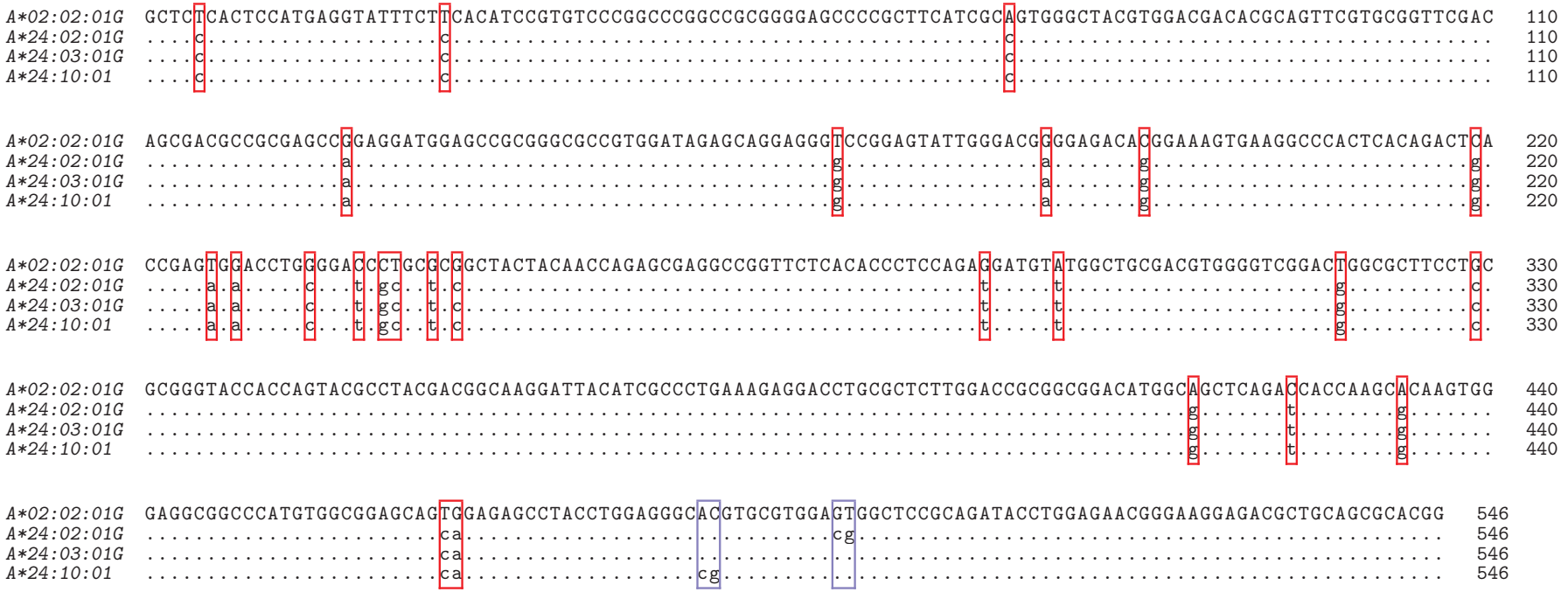
Supplementary Figure B27: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by xHLA for individual HG01051. SBT Alleles: *C*12:03:01G* & *C*14:02:01G*. xHLA Alleles: *C*12:02:01G* & *C*14:02:01G*. Nucleotides enclosed in red indicate an incorrect homozygous nucleotide call instead of a heterozygous nucleotide call. Nucleotides enclosed in blue indicate an incorrect heterozygous nucleotide call. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B28: Multiple sequence alignment of *HLA-C* alleles incorrectly assigned by xHLA for individual NA18939. SBT Alleles: *C*07:02:01G* & *C*12:02:01G*. xHLA Alleles: *C*07:02:04* & *C*12:02:01G*. Nucleotides enclosed in red indicate positions which were incorrectly genotyped for both alleles called. Sequence alignment consists of coding sequences for exons 2 and 3.

HISAT-Genotype

```
B*07:02:01G  CAGCGCGCTCCAGCTTGTCCTTCCCGTTCTCCAGGTATCTGCGGAGCCACTCCACGCACTCGCCCTCCAGGTAGGCTCTCCGCTGCTCCGCCTCACGGGCCGCCTCCCAC  110
B*18:01:01G  .c.......g....g.c..................g.....................gt..................a...........a..................  110
B*07:50      ..................................................................................................................  110

B*07:02:01G  TTGCGCTGGGTGATCTGAGCCGCCGTGTCCGCGGCGGTCCAGGAGCGCAGGTCCTCGTTCAGGGCGATGTAATCCTTGCCGTCGTAGGCGTACTGGTCATGCCCGCGGAG  220
B*18:01:01G  ........................g........c.............t.........................................g...................  220
B*07:50      ..................................................................................................................  220

B*07:02:01G  GAGGCGCCCGTCCGGCCCCACGTCGCAGCCGTACATGCTCTGGAGGGTGTGAGACCCGGCCTCGCTCTGGTTGTAGTAGCCGCGCAGGTTCCGCAGGCTCTCTCGGTCAG  330
B*18:01:01G  ....................................c.....................................................................a..  330
B*07:50      ..................................................................................................................  330

B*07:02:01G  TCTGTGCCTGGGCCTTGTAGATCTGTGTGTTCCGGTCCCAATACTCCGGCCCCTCCTGCTCTATCCACGGCGCCCGCGGCTCCTCTCTCGGACTCGCGGCGTCGCTGTCG  440
B*18:01:01G  ......tg.t..t....g.....................................t..................g......gtc.....................  440
B*07:50      ......tg.t..t....g..................................................................................  440

B*07:02:01G  AACCTCACGAACTGGGTGTCGTCCACGTAGCCCACTGAGATGAAGCGGGGCTCCCCGCGGCCGGGCCGGGACACGGAGGTGTAGAAATACCTCATGGAGTGGGAGCC  547
B*18:01:01G  ..................c.....................................................g.....................  547
B*07:50      ..................................................................................................  547
```

Supplementary Figure B29: Multiple sequence alignment of *HLA-B* alleles incorrectly assigned by HISAT-Genotype for individual HG00268. SBT Alleles: *B*07:02:01G* & *B*18:01:01G*. HISAT-Genotype Alleles: *B*07:50* & *B*18:01:01G*. Nucleotides enclosed in red indicate positions which were incorrectly genotyped for both alleles called. Sequence alignment consists of coding sequences for exons 2 and 3.

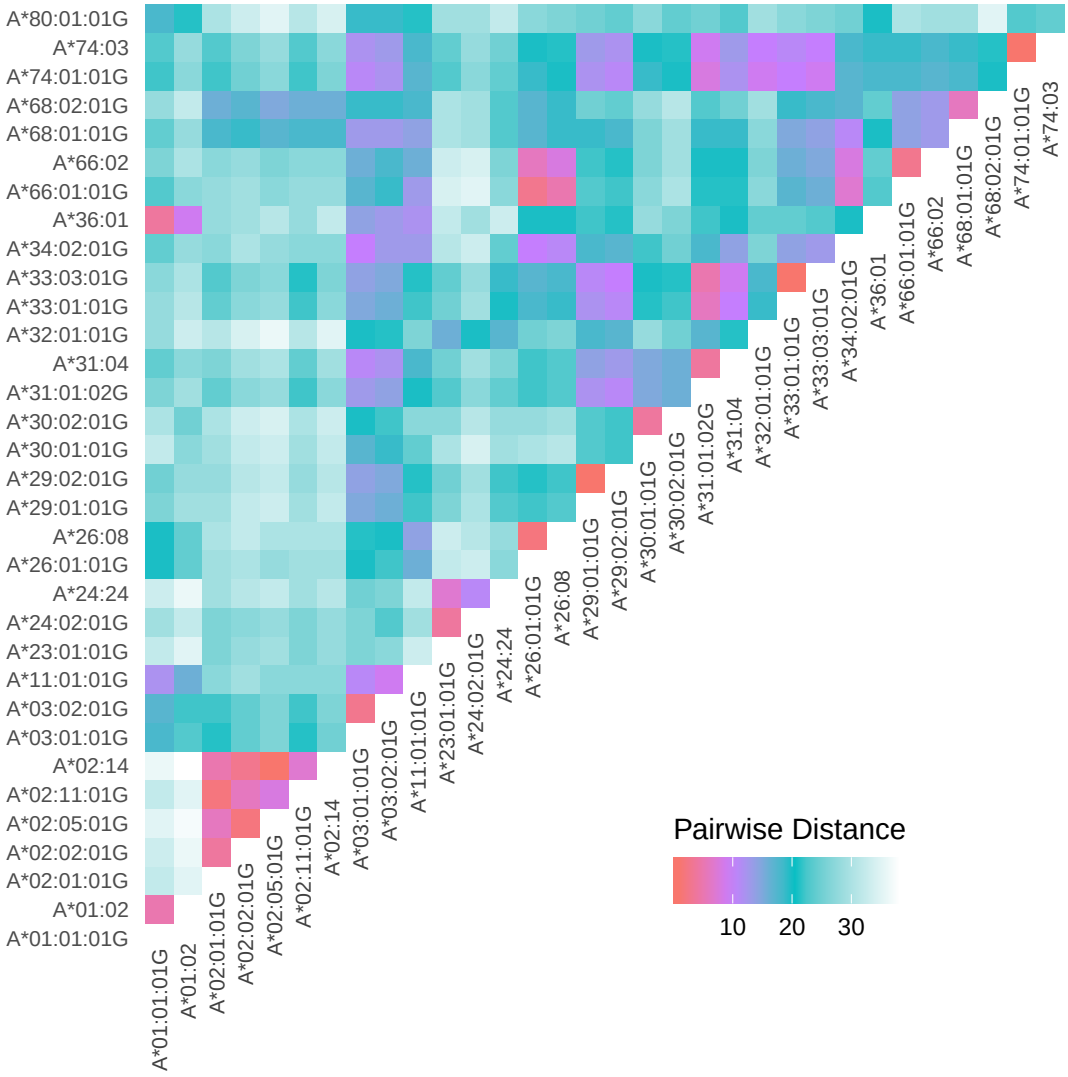


Supplementary Figure B30: Multiple sequence alignment of *HLA-A* alleles incorrectly assigned by HISAT-Genotype for individual HG01051. SBT Alleles: *A*02:02:01G* & *A*24:02:01G*. HISAT-Genotype Alleles: *A*24:03:01G* & *A*24:10:01*. Nucleotides enclosed in red indicate positions an incorrect homozygous nucleotide call instead of a heterozygous nucleotide call. Nucleotides enclosed in blue indicate an incorrect nucleotide call. Sequence alignment consists of coding sequences for exons 2 and 3.

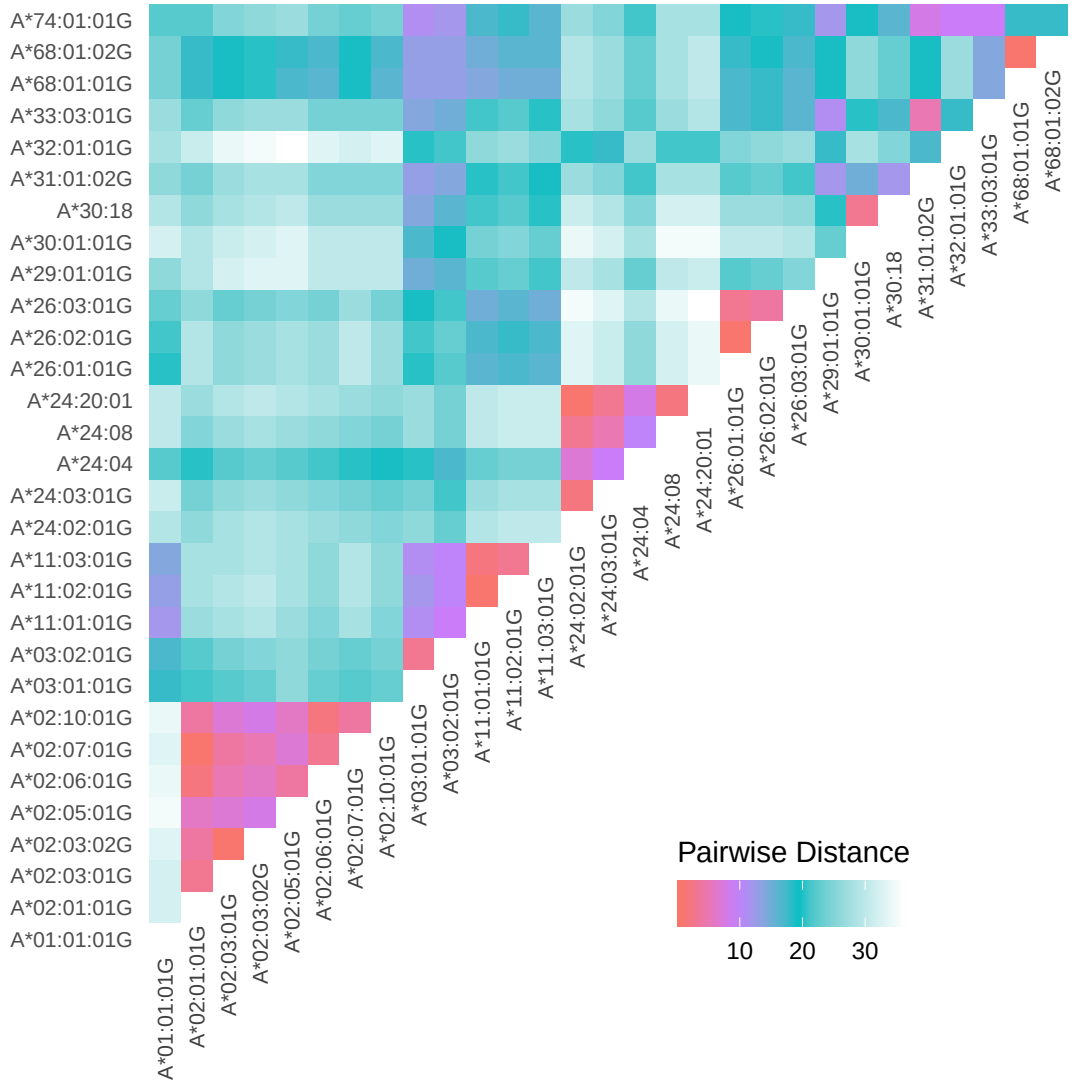
Supplementary Figure B31: Multiple sequence alignment of *HLA-B* alleles incorrectly assigned by HISAT-Genotype for individual HG01051. SBT Alleles: *B*15:16:01G* & *A*35:03:01G*. HISAT-Genotype Alleles: *A*15:16:01G* & *A*35:03:19*. Nucleotides enclosed in red indicate an incorrect nucleotide call. Sequence alignment consists of coding sequences for exons 2 and 3.

Supplementary Figure B32: Multiple sequence alignment of *HLA-B* alleles incorrectly assigned by HISAT-Genotype for individual NA19625. SBT Alleles: *B*07:02:02G* & *B*44:03:02G*. HISAT-Genotype Alleles: *B*07:36* & *B*44:03:33*. Nucleotides enclosed in red indicate an incorrect homozygous nucleotide call instead of a heterozygous nucleotide call. Nucleotides enclosed in blue indicate an incorrect nucleotide call. Sequence alignment consists of coding sequences for exons 2 and 3.

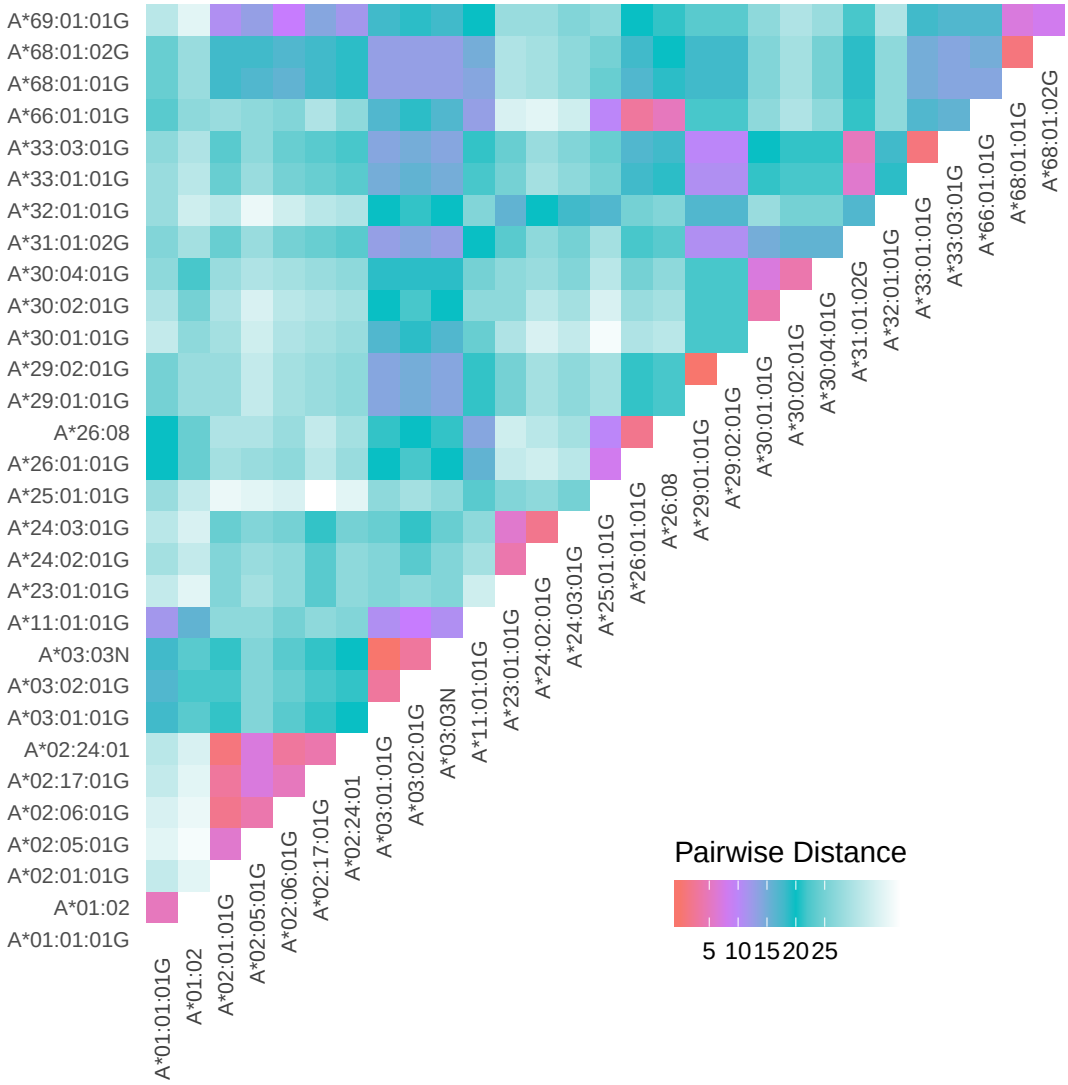
Intragenic Distances of *HLA-A* Between Four Different Super-populations



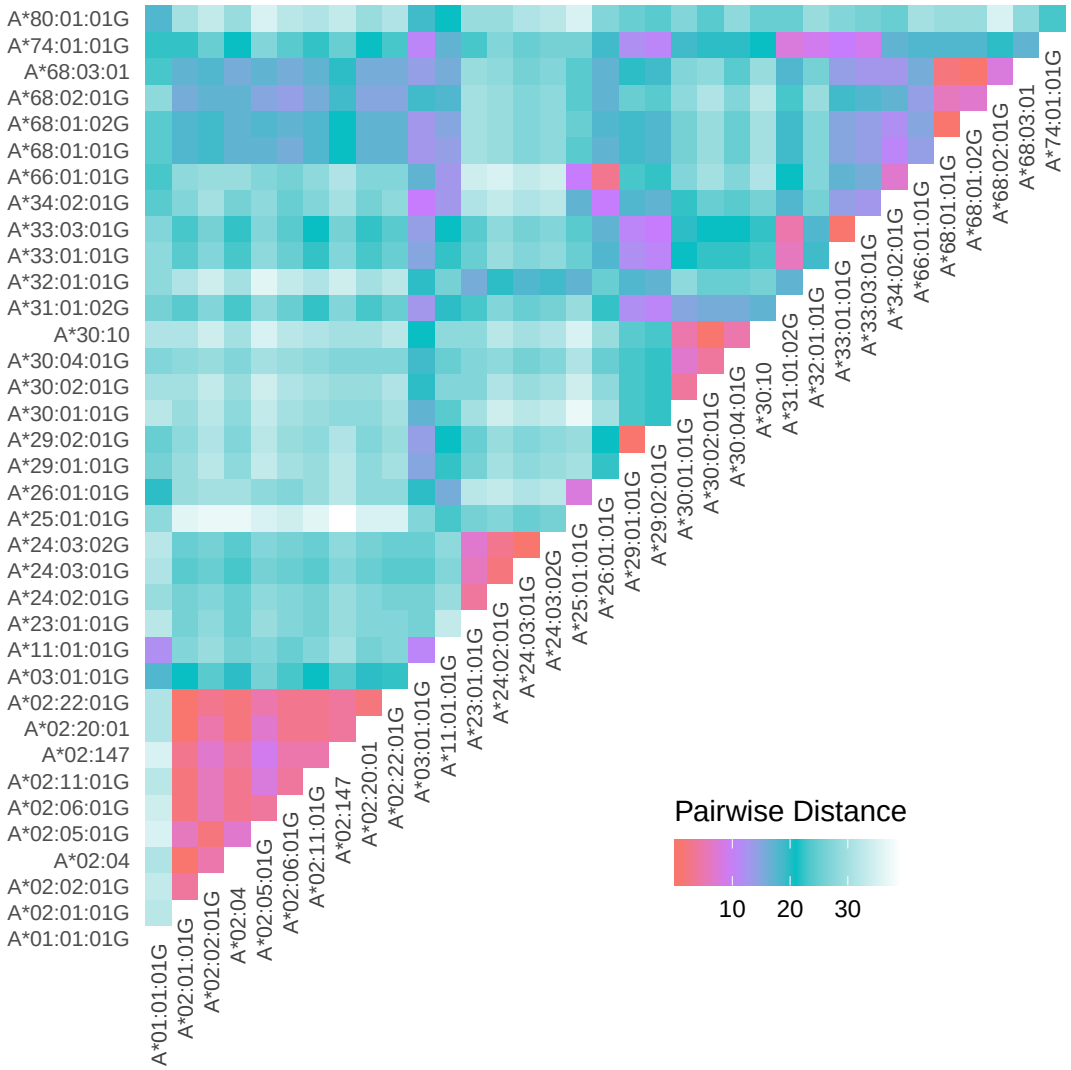
Supplementary Figure B33: Intragenic distances between the *HLA-A* alleles observed in the African super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.



Supplementary Figure B34: Intragenic distances between the *HLA-A* alleles observed in the Asian super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.

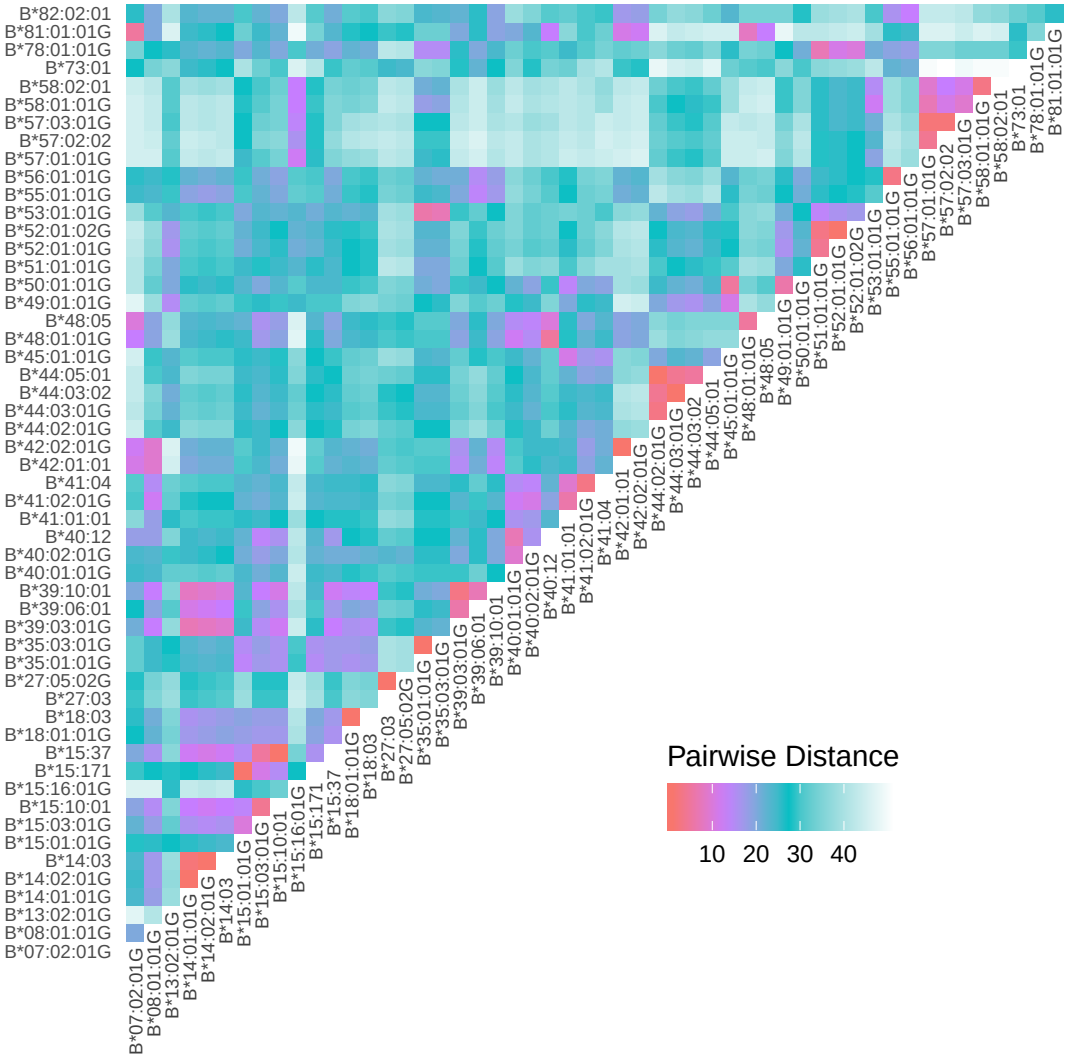


Supplementary Figure B35: Intragenic distances between the *HLA-A* alleles observed in the European super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.

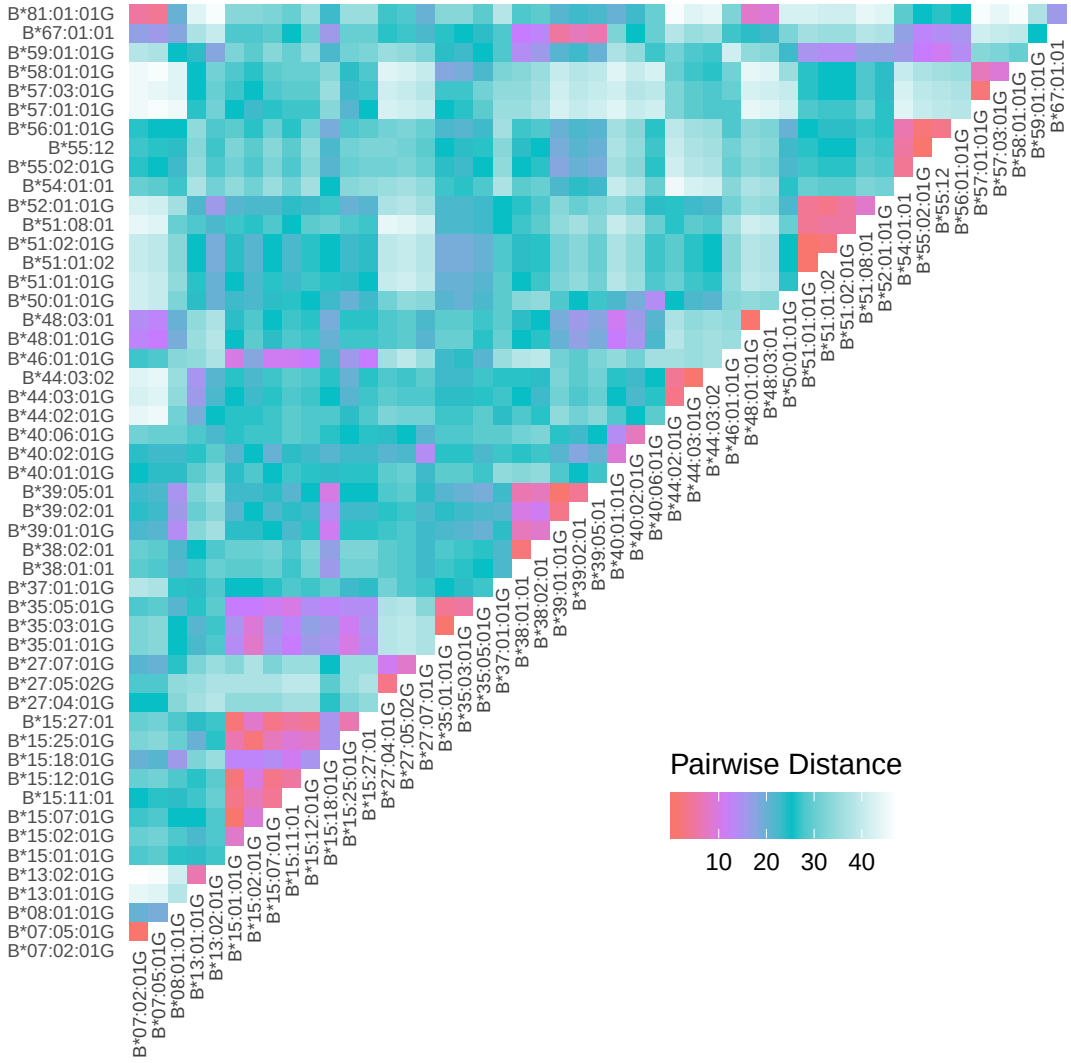


Supplementary Figure B36: Intragenic distances between the *HLA-A* alleles observed in the South American super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.

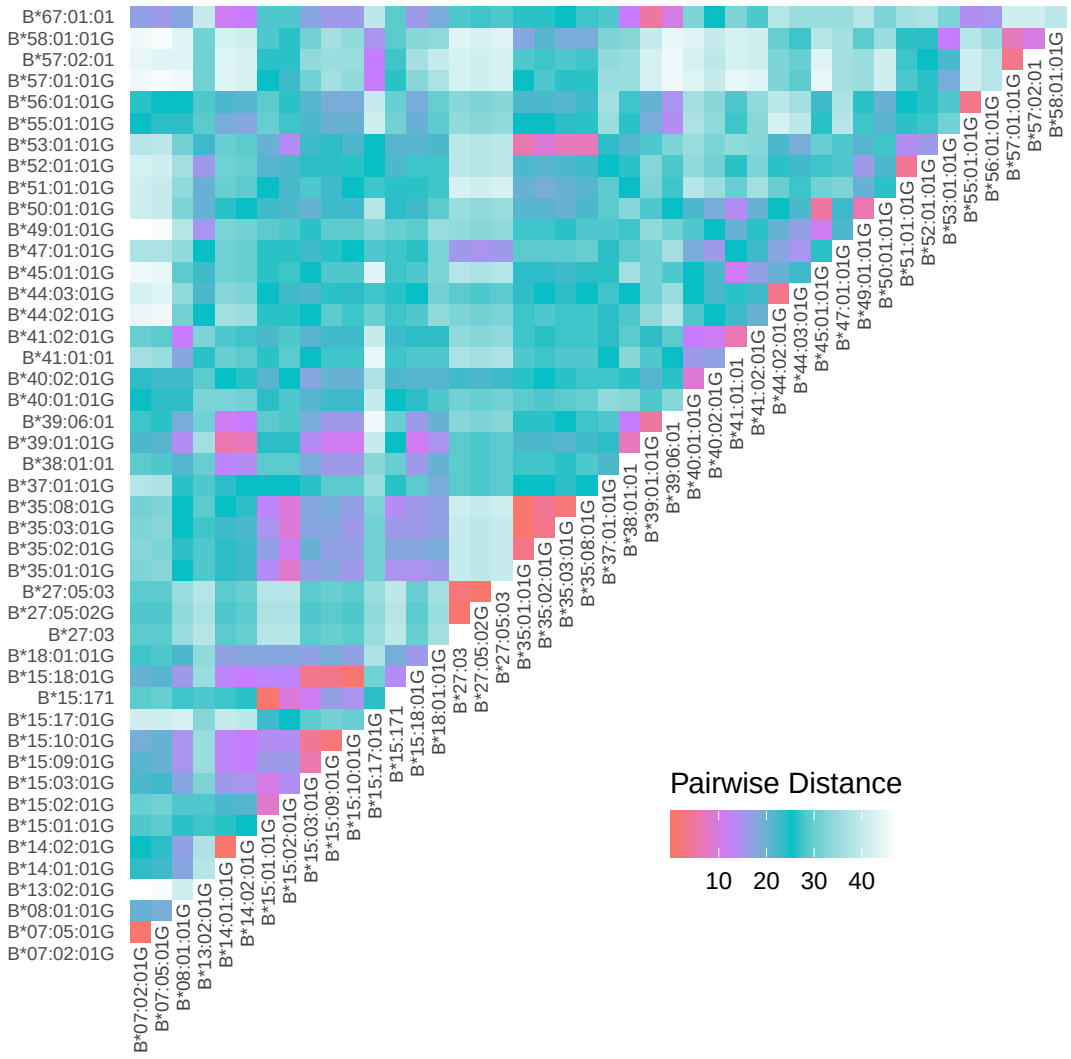
Intragenic Distances of *HLA-B* Between Four Different Super-populations



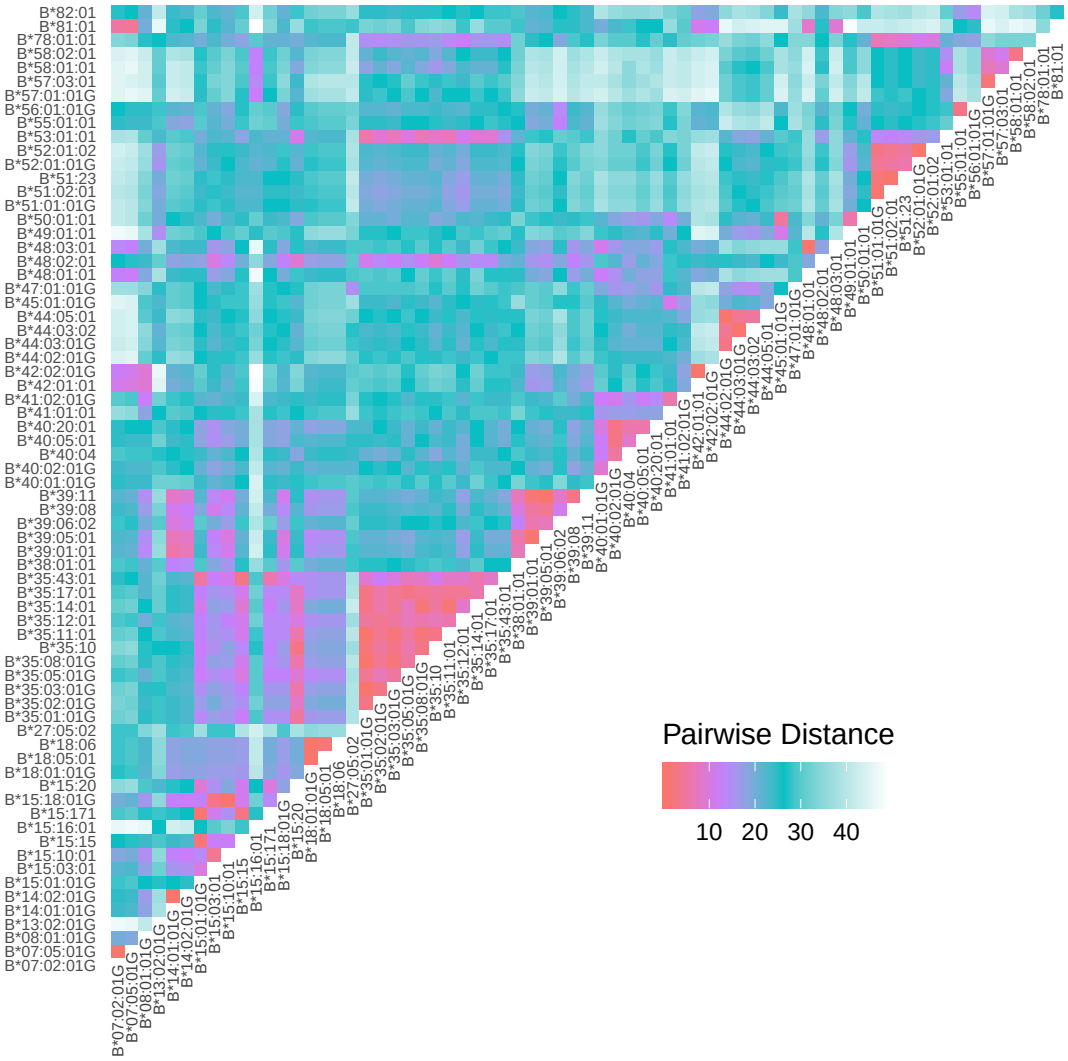
Supplementary Figure B37: Intragenic distances between the *HLA-B* alleles observed in the African super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.



Supplementary Figure B38: Intragenic distances between the *HLA-B* alleles observed in the Asian super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.

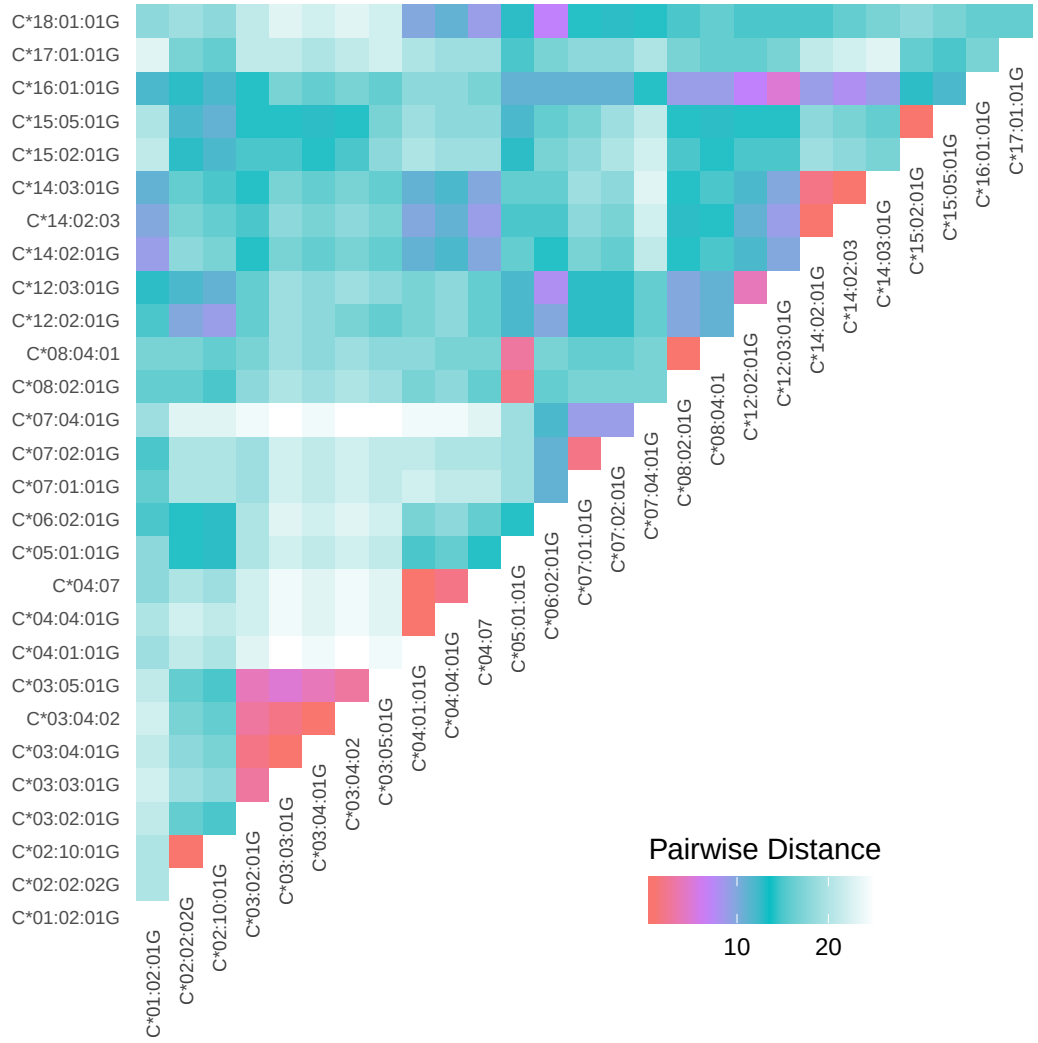


Supplementary Figure B39: Intragenic distances between the *HLA-B* alleles observed in the European super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.

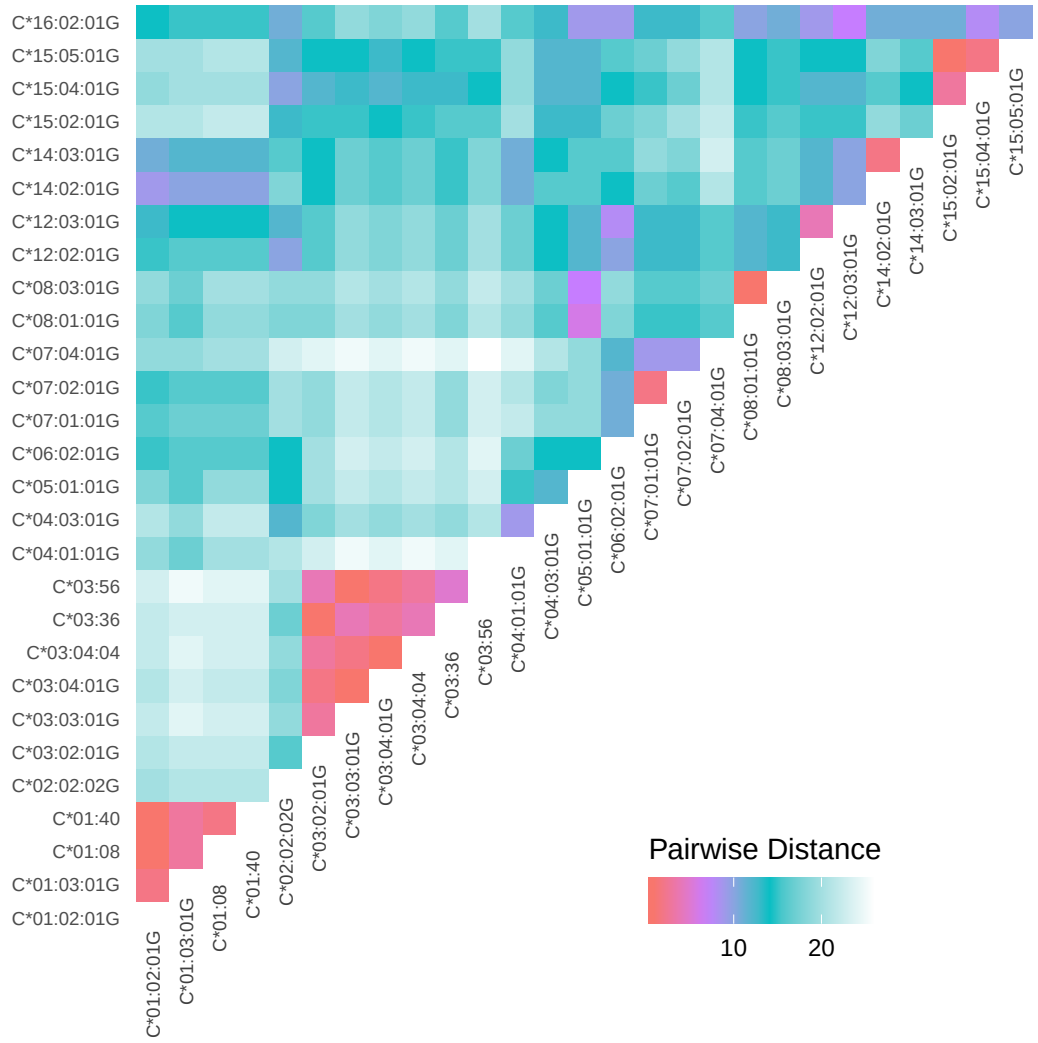


Supplementary Figure B40: Intragenic distances between the *HLA-B* alleles observed in the South American super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.

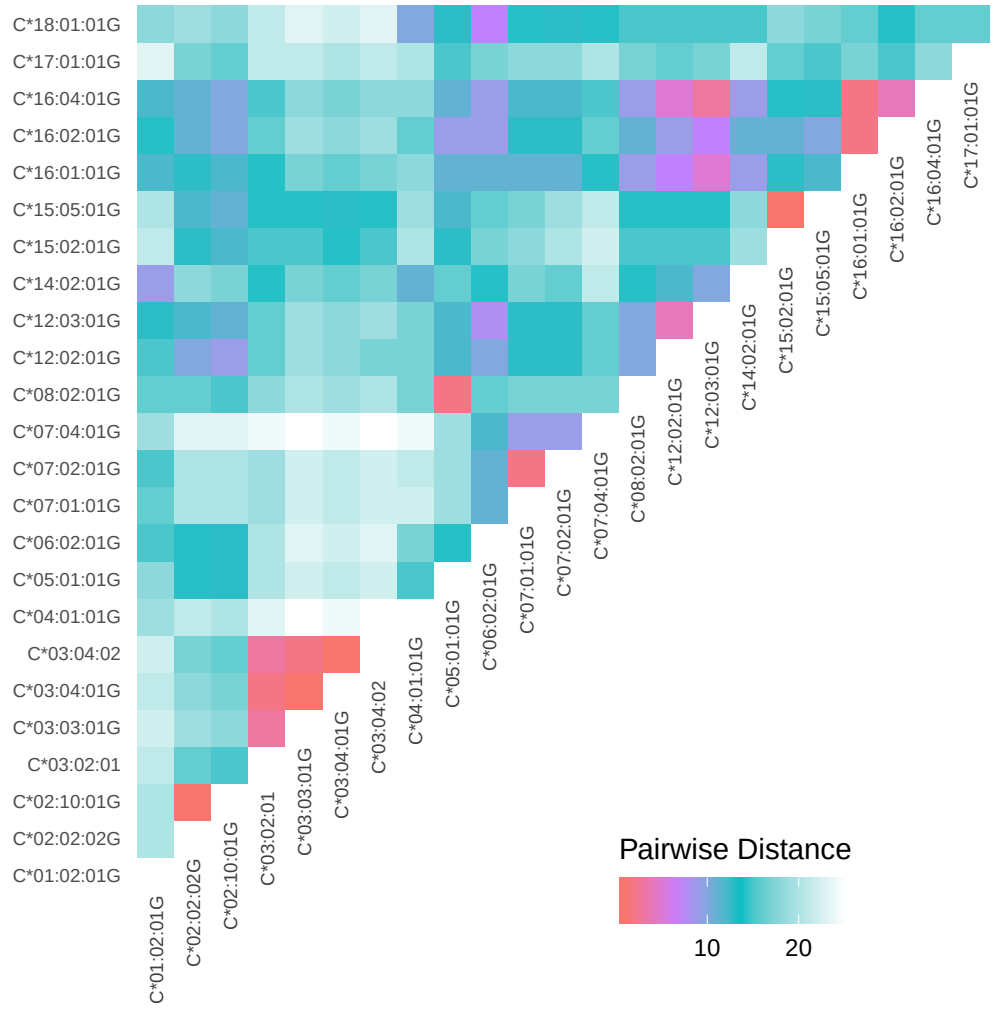
Intragenic Distances of *HLA-C* Between Four Different Super-populations



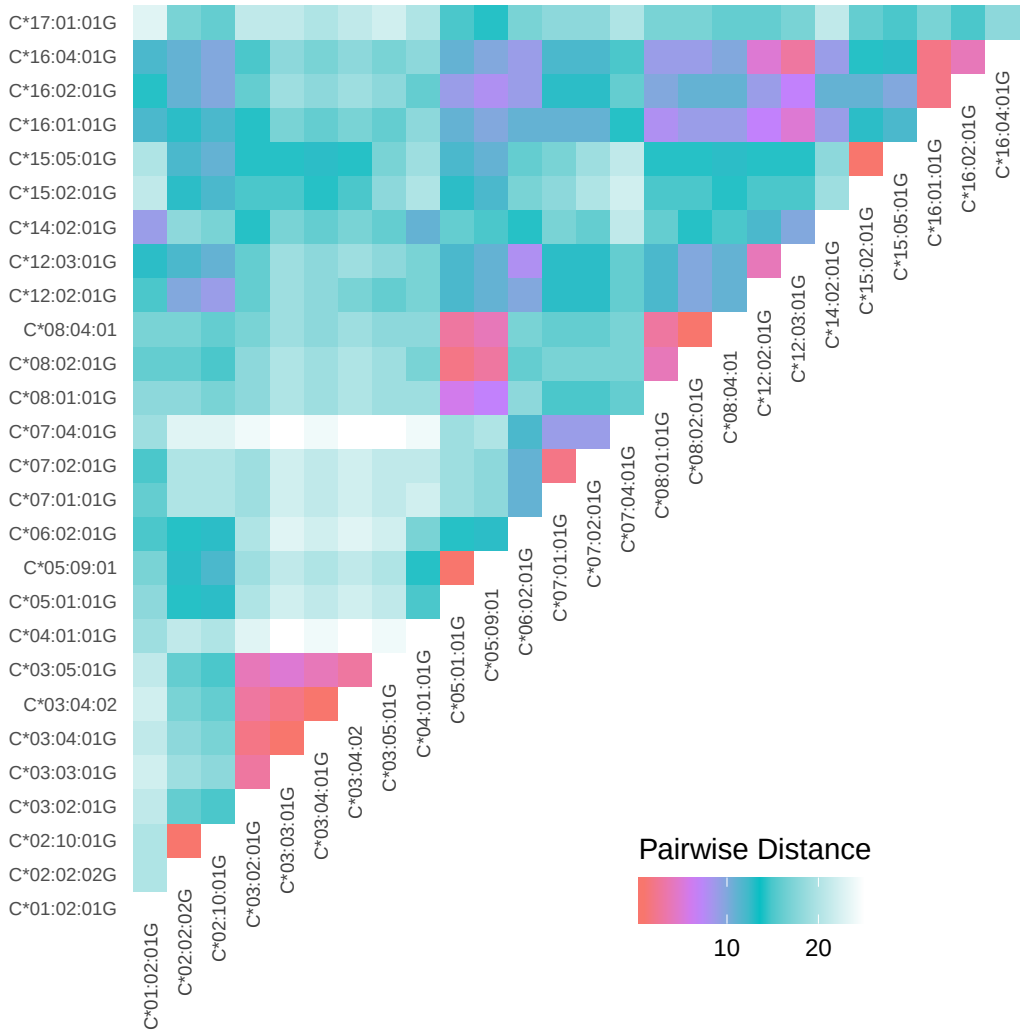
Supplementary Figure B41: Intragenic distances between the *HLA-C* alleles observed in the African super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.



Supplementary Figure B42: Intragenic distances between the *HLA-C* alleles observed in the Asian super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.



Supplementary Figure B43: Intragenic distances between the *HLA-C* alleles observed in the European super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.



Supplementary Figure B44: Intragenic distances between the *HLA-C* alleles observed in the South American super-population from the 1000 Genomes Project for whom high-resolution *HLA* SBT genotype data were available.

Appendix C - Code Listings

Data Acquisition

Listing C1: Data Acquisition

```
samtools view \
-h \ #Include the header in the output
-b \ #Output file in the BAM format
-o <SAMPLE>.highcoverage.bam  \ #Name of output file
<SAMPLE_URL>  \ #Input file
6:28000000-34000000 \ #HLA chromosomal region
GL000207.1 GL000226.1 GL000229.1 GL000231.1 \ #Alternate loci
GL000210.1 GL000239.1 GL000235.1 GL000201.1 \ #Alternate loci
GL000247.1 GL000245.1 GL000197.1 GL000203.1 \ #Alternate loci
GL000246.1 GL000249.1 GL000196.1 GL000248.1 \ #Alternate loci
GL000244.1 GL000238.1 GL000202.1 GL000234.1 \ #Alternate loci
GL000232.1 GL000206.1 GL000240.1 GL000236.1 \ #Alternate loci
GL000241.1 GL000243.1 GL000242.1 GL000230.1 \ #Alternate loci
GL000237.1 GL000233.1 GL000204.1 GL000198.1 \ #Alternate loci
GL000208.1 GL000191.1 GL000227.1 GL000228.1 \ #Alternate loci
GL000214.1 GL000221.1 GL000209.1 GL000218.1 \ #Alternate loci
GL000220.1 GL000213.1 GL000211.1 GL000199.1 \ #Alternate loci
GL000217.1 GL000216.1 GL000215.1 GL000205.1 \ #Alternate loci
GL000219.1 GL000224.1 GL000223.1 GL000195.1 \ #Alternate loci
GL000212.1 GL000222.1 GL000200.1 GL000193.1 \ #Alternate loci
GL000194.1 GL000225.1 GL000192.1 NC_007605 \ #Alternate loci
"*" #unmapped reads
```

Listing C2: RevertSam

```
java -jar -Xmx8G picard.jar RevertSam \ #Reverts BAM file to FASTQ
I=<SAMPLE>.highcoverage.bam \ #Input file
O=<SAMPLE>.unmapped.bam \ #Output file
SANITIZE=true \ #Remove shortened and unpaired reads
RESTORE_ORIGINAL_QUALITIES=true \ #Restore original Phred scores
```

```
ATTRIBUTE_TO_CLEAR=XT \ #read type (Unique, repeat, etc)
ATTRIBUTE_TO_CLEAR=XN \ #number of ambiguous bases in referenece
ATTRIBUTE_TO_CLEAR=AS \ #alignment score
ATTRIBUTE_TO_CLEAR=OC \ #original CIGAR string
ATTRIBUTE_TO_CLEAR=OP #original mapping location
```

Listing C3: SamToFastq

```
java -jar -Xmx8G picard.jar SamToFastq \ #Creates a FASTQ file/s
I=<SAMPLE>.unmapped.bam \ #Input file
FASTQ=<SAMPLE>.1.fq \ #Output file 1
SECOND_END_FASTQ=<SAMPLE>.2.fq #Output file 2
```

Listing C4: GRCh38 Read Alignment

```
bwa mem \
-t <threads> \ #Number of threads
-B 4 \ #Mismatch penalty
-O 6 \ #Gap open penalty
-E 1 \ #Gap extension penalty
-M \ #Mark shorter split hits as secondary
-R \ #Apply read group header
"@RG\tID:<SAMPLE>\tPL:illumina\\tSM:<SAMPLE>" \ #Read group header
<GRCh38> \ #Reference
<SAMPLE>.1.fq \ #Paired-end file 1
<SAMPLE>.2.fq | #Paired-end file 2
samtools sort \ #Sort output by position
-@ <threads>  \ #Number of threads
<SAMPLE>.sort.bam #Output

samtools index \ #Create index for BAM file
-@ <threads> \ #Number of threads
<SAMPLE>.sort.bam #Output

<GRCh38> - alt non-aware only included the linear GRCh38 sequence. The alt-aware
```

Listing C5: Read Count Data

```
samtools \
view \
-c \ #Produce read counts
<SAMPLE>.bam
<REGIONS>

# GRCh38 Regions: chr6: 28 510 120 - 33 480 577,
        Alternate loci located within the HLA
        region on chromosome 6,
        IMGT/HLA class I classical loci alleles
```

Listing C6: Depth of Coverage

```
java -jar GenomeAnalysisTK.jar \
-T DepthOfCoverage \
-R <GRCH38DH> \
-o <SAMPLE> \
-I <List of BAMS>.txt \
-geneList <LOCUS>.refseq \
-nt <THREADS>
-pt sample
```

Genotyping

BWAkit

Listing C7: BWAkit

```
run-bwamem \
-o <SAMPLE> \ #Output file
-t <threads> \ #Number of threads
-H \ #Apply HLA typing
-R \ #Apply read group header
"@RG\\tID:<SAMPLE>\\tPL:illumina\\tM:<SAMPLE>" \ #Read group header
GRCh38DH.fasta \ #Reference sequence
<SAMPLE>.1.fq \ #Paired-end file 1
<SAMPLE>.2.fq #Paired-end file 2
```

xHLA

Listing C8: xHLA

```
perl bin/typer.sh \ #xHLA genotyping script
<SAMPLE> \ #Input sample name
<SAMPLE>.sorted.bam \ #Input file
--full #Genotype to six-digit resolution
```

HISAT-Genotype

Listing C9: HISAT-Genotype

```
hisatgenotype_extract_reads.py \
--base genotype_genome \ #Reference assembly
-1 <SAMPLE>.1.fq \ #Paired-end file 1
-2 <SAMPLE>.2.fq \ #Paired-end file 2
--out-dir ./output \ #Output directory
-p <threads> #Number of threads

hisatgenotype_locus.py \ #Genotyping script
--base hla \ #Genotype HLA region
--locus-list A,B,C \ #Loci within HLA
-1 genomes.HLA/<SAMPLE>.extracted.1.fq.gz \ #Input file 1
-2 genomes.HLA/<SAMPLE>.extracted.2.fq.gz \ #Input file 2
-p <threads> #Number of threads
```

Kourami

Listing C10: Kourami

```
alignAndExtract_hs38DH.sh \ #Extract reads from HLA region
<SAMPLE> \ #Sample name
<SAMPLE>.sort.bam #Input file

java -jar -Xmx8G Kourami.jar \ #Genotyping program
-d ./db/ \ #HLA database location
```

```
-o ./output/<SAMPLE> \ #Output directory
<SAMPLE>_on_KouramiPanel.bam #Extracted reads location
```