Prospecting the MHC:

A Bioinformatic View of HLA Polymorphism and Gene Organization

Benedict Matern

PROSPECTING THE MHC:
A Bioinformatic View of HLA Polymorphism and Gene Organization

Benedict Mark Matern

© Benedict Mark Matern, 2020. Maastricht, The Netherlands.

ISBN: 978-94-6380-742-5
Printing: proefschriftmaken.nl
Cover Art & Layout: Fenna Schaap (www.fennaschaap.nl)

The studies presented in this thesis were conducted at GROW-School for Oncology and Developmental Biology, and the Department of Transplantation Immunology, Tissue Typing Laboratory at Maastricht University Medical Center.

PROSPECTING THE MHC:
A Bioinformatic View of HLA Polymorphism and Gene Organization

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public

on Wednesday, the 25th of March 2020 at 12:00

by

Benedict Mark Matern
Born on 28 April 1986 in Minneapolis, Minnesota, USA

Supervisor:
Em. Prof Dr. Marcel G.J. Tilanus

Co-supervisor:
Dr. Hans M. Groeneweg

Assessment committee:
Prof. Dr. Paul H.M. Savelkoul (Chairman) (Maastricht University Medical Center+)
Prof. Dr. Steven G.E. Marsh (Anthony Nolan Research Institute)
Prof. Dr. ir. Chris T.A. Evelo (Maastricht University)
Prof. Dr. Axel zur Hausen (Maastricht University Medical Center+)
Assoc. Prof. Dr. Eric Spierings (University Medical Center Utrecht)

Contents

CHAPTER 1

General Introduction

Introduction

Bioinformatics

This thesis is focused on the use of bioinformatics applied to molecular analysis to answer scientific questions in the field of HLA and immunology. Bioinformatics is a tool that helps develop a scientific vision; it informs the generation of queries, but also suggests methodology to solve the queries. Bioinformatics is a relatively new cross-disciplinary field of research, which has a broad and flexible definition due to its connections to concepts and techniques from many fields.[1] It combines concepts from computer science and biology into a platform for answering biological questions by computational methods. Bioinformatics would not be possible without ideas from database design, algorithm design, data standards, statistics, computer hardware, and computer programming, and it would not have a purpose without important background knowledge and scientific questions from the biological and biomedical fields. Bioinformatics represents an exciting and promising field with a variety of applications, such as molecular biology, metagenomics, evolutionary biology, protein modeling, and drug design. Perhaps most pertinently, bioinformatics has applications in the queries which apply the analysis of biological sequences, including peptide, RNA, and DNA sequences.

Deoxyribonucleic acid(DNA) is the long, double-helical polymer of nucleotides contained within nearly all cells of all living things. Although some methylation and epigenetic information can be passed to offspring,[2] DNA contains most of the inherited information for living organisms. The collection of DNA contained within human cells is referred to as the human genome, the bulk of which is made up of 23 chromosome pairs containing approximately 20,000 - 25,000 protein-encoding genes.[3] Humans, like nearly all mammals are diploid, which means that each chromosome contains a pair of homologous sequences. One copy of the genome is inherited from both the individual's father and the mother. The chromosome pairs are arranged in "H" shaped chromosomes, and paired chromosomes are joined near the center in a structure called a centromere. A single copy of the human genome contains approximately 3 billion nucleotide bases, so the entire genome including both paired copies consists of approximately 6 billion nucleotides. In addition to the 23 chromosomes, humans inherit 16.6kb[4] of mitochondrial DNA, contained within intracellular mitochondria and nearly always inherited from the mother.[5] Patterns within the DNA serve specific biological purposes, which continue to be studied and elucidated. Early analysis of the genome suggests that the genome contains tens of thousands of genes, but only 1% of Single Nucleotide Polymorphisms (SNPs) will affect protein function[6]. More recent studies suggest that the genome is rich with regions of DNA containing functional polymorphism, with estimates of the functional DNA ranging from 15%[7] to 80%[8]. It is not just expressed regions that are important; there is evidence that the non-coding regions[9] are full of polymorphism that suggest biological function.

The Central Dogma of Molecular Biology[10] is the widely accepted theory of how genes bring about protein function. Within the nucleus, RNA polymerase binds to specific promoter sequences nearby a gene and transcribes the DNA into precursor messenger RNA (pre-mRNA). While still in the nucleus, post-transcriptional modifications, including the splicing of intron sequences and polyadenylation, mature the pre-mRNA into a processed messenger RNA (mRNA). The mRNA leaves the nucleus and is subsequently translated into functional proteins by ribosomes. A cell's ribosomes bind to a messenger RNA, and decode the nucleotide sequence to create a very specific amino acid sequence, where every sequence of three nucleotides (codons) encodes a specific amino acid.

Genes are most commonly referred to as the regions of DNA that encode a protein. Genes are generally expressed as proteins, but gene expression can vary greatly among cell types, and is directly affected by environmental factors. The translation of messenger RNA into a protein can also be regulated by the actions of microRNAs, which complement the mRNA sequence and inhibit translation.[11] A single gene can encode multiple functioning protein isoforms by way of alternative splicing. Alternative splicing provides a great deal of flexibility in how organisms can adapt to different conditions, and how to express genes at different developmental stages, implying that nearly any estimations of gene counts or proportions of the genome that are functional are an underestimation of the variety of the functionality that is encoded in the human genome.[12]

Analysis of genes or regions is interesting and informative, but an individual gene does not act alone. The genome has pieces and components that interact with each other in unexpected ways, and the function of a gene must be considered in the context of neighboring genes, haplotypes, chromosomes, the entire genome, and the local cell microenvironment. The expression of one gene can change the behavior of another gene, possibly affecting expression levels. In different cell types, DNA structures might be methylated, which can activate or deactivate genes, providing flexibility in expression in cell types and stages of development. While it is convenient to view the genome as a static structure, in truth it is a dynamic system.

Although 99.9% of the genome is identical between humans,[13] there is still significant polymorphism between individuals. Genomes differ by many individual SNPs, and by sequence insertions or deletions, and in some cases major recombinations or genomic rearrangements. Patterns of individual SNPs can be studied to improve our understanding of historical evolutionary diversions, and analysis of these patterns allows comparison of individuals from varying ethnicities. More importantly, polymorphism between individuals can bring about changes in the biological function. Polymorphism can result in amino acid differences in proteins, or change the regulatory function of genes. Analysis

of the polymorphism, whether at gene level or on the level of the entire genome, helps to explain the biological differences that make an individual unique.

Molecular Analysis

The study of genetics, and the nature of DNA and its role in inheritance is an ever-evolving field. As new discoveries are made, it is critical to develop new techniques and systems of understanding. As techniques and technologies advance,[14] increasing amounts of data are being generated, and the amount of data being generated requires researchers to develop new ways to represent, analyse, and share data.

The most clear way to study genomics is by determining the actual nucleotide sequence of the DNA. Identifying DNA sequences extends our capability of analyzing not just specific polymorphism, but determining the composition of entire genes or regions. Sanger Sequencing[15] was invented in the late 1970s and quickly became the standard sequencing platform for several decades, and specific software tools were developed for accurately interpreting the sequencing data. Further platform developments led to Next-Generation-Sequencing (NGS), a general term encompasses a variety of sequencing technologies focused on higher throughput and efficiency improvements.[16,17] The efficiency improvements naturally lead to more rapid accumulation of data, and higher sample throughput. With NGS, more targeted regions could be analyzed, and data from individuals could be compared against each other, and thus it became necessary to develop more precise assembly and analysis tools. The usefulness of NGS technologies are limited by the short reads, which is partially resolved by technologies of third generation sequencing, including Pacific Biosystems SMRT sequencing,[18] and Oxford Nanopore single-molecule sequencing with the MinION.[19] Nanopore sequencing allows direct analysis of long stretches of DNA in individual reads, which provides more insights into the phasing and patterns of polymorphism in genes compared to short reads.

High-throughput gene sequencing technologies enable cheap and fast access to entire human genomes, and even facilitate comparisons of the patterns across populations and ethnicities. However, interpretation of sequencing data must be performed using specific algorithms. Sequence data can be interpreted by one of two major techniques: by aligning reads against a reference sequence in order to identify polymorphisms relative to the reference, or by using algorithms that perform de novo assembly, where read sequences are assembled into a continuous nucleotide sequence consensus without the need of a genomic reference. Reference-based analysis has the advantage of using previous knowledge of sequence patterns to quickly determine the most likely sequence of the sample, but de novo assembly, while more difficult and computationally intensive, may uncover sequence with novel genomic structure. As the tools for analysis of genetic data have advanced, the quantity of data has increased, creating a need to improve analytical

capabilities, both in how to store data, and how to accurately analyze it. This has led to improved algorithms for alignment and genome assembly, and new systems for storing and understanding the data.

One of the earliest alignment strategies was the Needleman-Wunsch (NW) algorithm, published in 1970,[20] which can find the optimum alignment of any two nucleotide or protein sequences. This algorithm is widely used, but has some limitations in accuracy and applicability. Smith-Waterman (SW) algorithm[21] is similar to the NW algorithm, but has the benefit of identifying accurate local alignments, which might have been hidden by the global alignment scores calculated by NW. These algorithms are created for the comparison of any two nucleotide sequences, and have been adopted for the purpose of aligning sequencing reads against a reference. The Basic Local Alignment Search Tool (BLAST) algorithm has a different approach, it was developed to quickly compare approximate local alignments of a query sequence against a database of references,[22] and remains a useful tool for identifying sequence similarities. Multiple sequence aligners, such as Clustal,[23] were designed to compare multiple sequences together, which also provides the capability to perform phylogenetic analysis. With the advancement of sequencing technologies, alignment algorithms such as LAST[24] or minimap2[25] were developed for specific platforms, and can accommodate read errors and sequence deletions. Algorithms were also developed for the purposes of identifying haplotype patterns.[26] Population genomics software such as Pypop[27] use multipurpose algorithms, especially expectation maximization, to identify haplotype frequency patterns in population samples.[28] Many alignment and analysis algorithms are provided with easy-to-use software application programming interfaces (APIs), such as those included in Biopython,[29] which promotes a community of sharing and reuse by enabling bioinformaticians to easily use common algorithms.

Although these advances are rapidly developing at sometimes unbelievable speed, It is important to consider all of the technology and analytical advances in a larger timeframe. It is without meaning that we refer to a set of technology as "Next"-Generation-Sequencing or "Third"-Generation-Sequencing. Tools are regularly superseded when new questions or applications require new approaches. New algorithms and data standards are developed continuously. What has become clear is that the genome is a complex, dynamic living structure, and analysis algorithms evolve to match that complexity. It is clear that new tools and scientific strategies will continue to be necessary for answering the new scientific questions.

Applications of Bioinformatics in Molecular Analysis

As technologies for molecular analysis advance, they can be applied to new projects. The classic example of bioinformatics analysis is the development of the Human Genome Project,[30] which took place approximately during the years 1990-2003. This project was an international collaboration, funded by the National Institutes of Health, Wellcome Trust Sanger Institute in Cambridge, UK, and was carried out in laboratories in countries around the world. The challenge of generating the first draft sequence of the human genome has been completed, and has provided a nearly-perfect map and sequence of the collection of human DNA. The completion of the project does not represent the conclusion of a scientific question, but represents more the beginning of the field of study of genomic analysis. The major functions of most of the content of genome was widely unknown, and questions were raised about the specific function of the DNA, especially in regions that do not encode proteins. This accomplishment marked the assembly of one individual genome, and naturally it raised questions about what makes individuals different, and how the DNA sequence in humans relate to each other. The goal of sequencing an individual genome progressed into the goals of sequencing multiple genomes, and the sequencing platforms and analysis tools also moved forward to match, bringing with them a reduction in the cost of sequencing. The concept of a $1000 genome has been commonly referenced as a benchmark for sequencing cost, but advances in sequencing technology have continuously reduced the cost of sequencing,[31] turning the concept of a $1000 genome from an idea into an achievable goal.

The 1000 Genomes Project,[13] enabled by the reduced difficulty and lower cost of sequencing a human's genome, is a natural extension of the Human Genome Project. It has the major aims of comparing sequences and structures within the genome in the context of populations. Individuals from 26 populations have been compared, and each genome discovered varied from the known references by 4.1 million - 5 million sites. The data obtained for this project has been extensively collected and curated. The 1000 genome data is available for the public; polymorphism can be explored easily in the genome browser, and summaries of polymorphism and SNPs for a region, including population frequencies, can be downloaded and used in bioinformatics analysis. Polymorphisms are mapped to locations within the genome assembly, and for many SNPs metadata are often available as well, such as references to published articles and results from disease association studies.

Polymorphism identified from whole genome sequencing can be applied to Genome-wide association studies (GWAS), a category of high-throughput bioinformatics techniques which are designed to detect the effect of single nucleotide polymorphisms on the causality or relationship with disease. DNA microarrays are designed with a variety of (500,000 - 1,000,000 or more) oligonucleotides that correspond to selected

polymorphisms across the genome. This enables the high-throughput detection of specific polymorphism across the genome, which can be statistically analyzed to identify correlations with a trait or disease. GWAS studies can be designed for a specific purpose, where groups of patients and controls are compared to determine general patterns of sequence polymorphism. Statistical tests can determine the significance of differences between the patient and control groups and suggest linkages of specific polymorphism with disease.

High throughput sequencing has applications outside of sequencing genomes. It also enables metagenomics analysis, where research strategies aim to quickly identify the presence of a variety of species, rather than to complete a full and accurate genome sequence. The ability of an investigator to identify the presence of diverse species applies to many real-world problems. One could investigate what organisms are present in a water sample to compare environments of different bodies of water,[32] or the presence of pathogens in soil[33] that can indicate the health of land a farmer wants to plant in. Analysis of an individual's gut microflora can also give clues to their health and inform medical decisions,[34] demonstrating that clinical decisions can be made without determining an entire genome sequence.

Molecular analysis can also be used to track evolutionary patterns of an organism. Nick Loman used the MinION to sequence the Zika Virus[35] as the disease migrated across South America. Samples were collected at various locations which were affected by the Zika infection at different time points. Genetic differences can be mapped to geographical locations, which provide a real-world mapping of how the virus evolves, spreads, and infects individuals. Nanopore sequencing was also applied to monitor the spread of the Ebola virus across West Africa,[36] demonstrating that molecular analysis and high throughput sequencing provide important clues in the spread of infectious disease.

Each of these applications require new ways of thinking, and new algorithms and techniques to interpret the data. Specifically built bioinformatics tools can be created to address specific questions, including the prediction of functional sequence patterns. Some software packages will search for the target sites for restriction enzymes within an assembled DNA sequence, and generate visualizations of the restriction site locations. This can indicate where enzymes will act on a sequence, and gives indications about the genomic context of the sequenced region. Likewise, tools have been created to use known patterns to identify putative alternative splice sites,[37] or the presence of Alu elements in a given nucleotide sequence,[38] which can indicate sequence that affects splicing or expression of a gene.

HLA and the MHC

In the human genome, Chromosome 6 stands out as unique and exceptional compared to the other chromosomes. The short arm of Chromosome 6 contains the human Major Histocompatibility Complex (MHC). Polymorphism is present across the genome in similar patterns, due to evolutionary patterns and random mutations, but Leffler *et al*. have found in their comparisons of human and non-human primate polymorphism, that polymorphisms on chromosome 6 are on average much more densely packed compared with any other chromosome.[39] When the MHC region is excluded from analysis, chromosome 6 shows a nearly identical pattern of polymorphism to the other chromosomes, clearly illustrating the differences in the MHC polymorphism compared to the other loci.

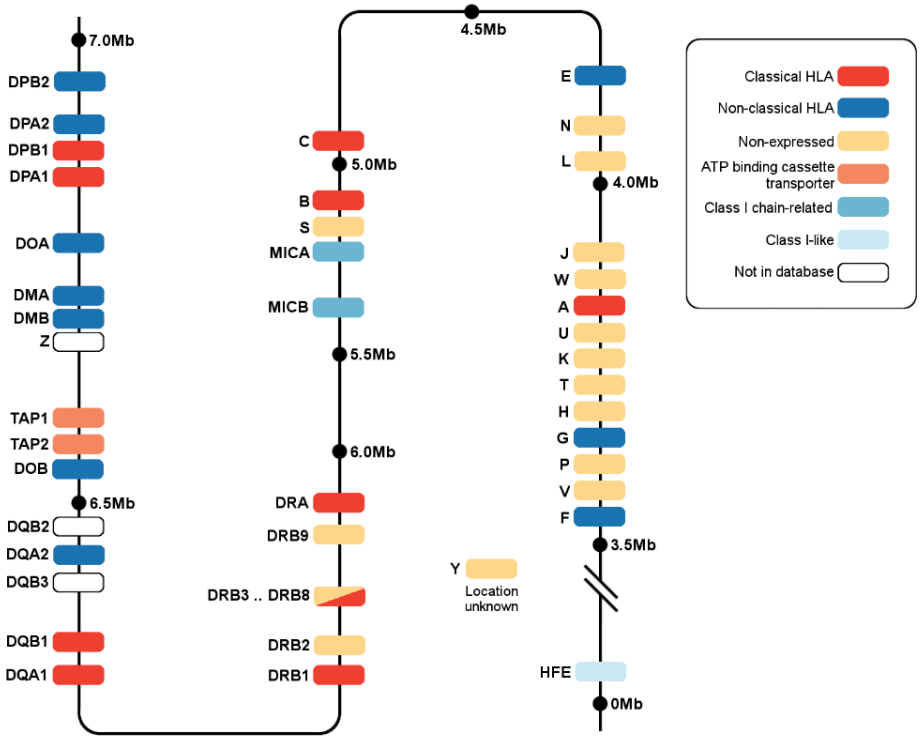


Figure 1. A map of the human MHC region located on the short arm of Chromosome 6. The MHC is rich with genes; it contains HLA class I and class II genes, non-expressed HLA pseudogenes, several genes related to peptide loading and immune function, as well as many genes without known immune function, such as olfactory receptors (not shown). HLA-DPB1 and -DPA1 (Chapter 8) can be seen near the far left (centromeric) end, and mark the beginning of the classical class II region. HLA-DQB1, -DRB1, -DRB3/4/5, and -DRA (Chapters 6 & 7) are also located within the class II region. HLA-A, -B, and -C, as well as many class I pseudogenes are located in the HLA class I region (Chapters 1, 5, & 9), which lies closer to the telomeric end of the chromosome. Reproduced with kind permission of Prof. Steven GE Marsh, Anthony Nolan Research Institute, London, UK [http://hla.alleles.org/alleles/index.html][40]

The MHC region developed by a series of recombinations and duplications.[41,42] The recombinations can occur by several different mechanisms. Recombinations may be focused at specific positions within the genome, which may vary among populations.[43] Sometimes these recombinations can result in copy number changes, where the same gene is encoded at multiple loci. Recombinations are relatively rare events, around one in every 100 meioses for any stretch of 1 million bases,[44] and with the exception of these rare recombinations, the MHC is inherited as a single inheritable haplotype. The concept of a "haplotype" was developed originally around 1967 by Ruggero Ceppellini to describe the idea of HLA antigens being segregated to each of the two chromosomes,[45] but outside of the HLA community it can be used to describe any linked genetic determinants that both lie on the same chromosome which are inherited together. In analysis of HLA and the MHC, the phrase "haplotype" might be used in reference to two SNPs on the same chromosome, or the combined genotyping of multiple complete HLA alleles. "Haplotype" can be, perhaps more informatively, used to refer to the complete collection of HLA and non-HLA genes within one of an individual's two copies of the MHC. Many of the genes within the MHC in linkage disequilibrium (LD), which means that alleles at two loci are inherited together at higher rates than they would if they were randomly inherited. LD can allow researchers and clinicians to make inferences about sequence at different loci without actually sequencing that region.[46]

The MHC region is rich with genes, including both HLA and non-HLA genes (Figure 1). The Human Leukocyte Antigen (HLA) genes are the most relevant and heavily-studied genes in this region. The HLA gene system was originally referred to as HL-A,[47] referring to the human version of the MHC molecules, but in 1968 upon determining that multiple genes contribute to the system, the genes were renamed as HLA-A, HLA-B, HLA-C, etc. The HLA genes are divided into two major classes (HLA class I and class II) which differ in function and morphology. The MHC also has a class III region,[48] but the genes in this region, unlike class I and II molecules, do not seem to be primarily responsible for antigen presentation. The HLA genes are the most polymorphic in the human genome,[49] to the point that HLA is commonly referred to as hyperpolymorphic. Each HLA locus has distinctions which define the genes, but alleles at each loci contain their own polymorphism. The variation in HLA alleles is vast, and this diversity is critical to maintain variability in immune function for the survival of mankind.

HLA-A, -B, and -C are the classical HLA class I genes, which are expressed on nearly all nucleated cells. The HLA class I molecule is formed by a heterodimer of a class I subunit, encoded by a HLA class I gene, and the nearly non-polymorphic beta-2 microglobulin, which is encoded on chromosome 15. The class I HLAs have the primary function of binding intracellular peptides and presenting them on the cell surface. These intracellular peptide antigens, which have been degraded by the proteasome, are presented to CD8+

(cytotoxic) T-cells, which are able to distinguish between self vs non-self antigens and trigger a cytotoxic immune response, thereby destroying the cell with an unrecognized antigen. In this way, HLA class I serves as a marker of cell health, and can be used to communicate that a cell is damaged by *e.g.* a virus.

The HLA class II molecule, in contrast to the HLA class I, is formed of two heterodimer proteins, an alpha and beta subunit. The alpha and beta subunits are encoded by two HLA genes (*e.g.* HLA-DPA1 and HLA-DPB1 respectively) which are encoded in the MHC. The resulting proteins form a heterodimer which is expressed on the cell surface. Another major distinction from HLA class I is that class II is expressed mainly on antigen presenting cells, such as macrophages and dendritic cells, and B-cells. An antigen presentation cell can endocytose extracellular particles, load the peptide antigens into the peptide binding groove of the class II heterodimer, and present them on the cell surface. HLA class II interacts with CD4+ (helper) T-cells, and when helper T-cells are activated by a foreign antigen, they can begin an immune response by proliferating and by secreting cytokines.

The cluster of genes within the MHC perform a variety of interrelated functions.[50] The HLA genes are the most notable within this region, but the non-HLA genes are important to note as well. Many of the MHC genes, such as HLA and the Tumor Necrosis Factor (TNF) genes, are core to the functions of immunity and inflammation, but some genes are responsible for a different set of functions. Genes responsible for olfactory receptors also lie within the MHC, and these loci may be linked to preferences for mate selection.[51] The MHC has also been linked to an individual's preference for food or drink,[52] although further studies are necessary to identify if the preferences correlation is with an HLA gene directly, or by the behavior of olfactory genes in linkage disequilibrium with the correlated sequence.

Although it is convenient to think of haplotypes as having the same organization of genes in a single pattern, there are several common haplotype patterns within the human MHC.[53] Specific alleles at one HLA locus are often inherited alongside alleles at a different locus, which defines haplotypes that are more frequently observed than others.[54] Some haplotype organizations have been conserved throughout a long evolutionary history.[55] These haplotypes generally remain conserved over time, likely due to beneficial interactions between the genes on the same haplotype, or a protective effect. Common haplotypes are useful for defining general haplotype patterns, but more insight can be found by analyzing the HLA genes and MHC of the wider population.

Analysis of the MHC

Early DNA-based analysis methods like RFLP[56] use restriction enzymes to digest pieces of DNA, and the lengths of the resulting fragments allow comparison of the restriction sites and genomic context between samples. RFLP has the benefit of detecting some changes in sequence without knowledge of the specific difference, at the expense of very low-resolution analysis. Even so, comparisons of the fragments generate biological data which can be analyzed and compared among individuals. RFLP techniques have led to a more advanced understanding of chromosome arrangement and DNA sequence structure. Techniques with higher resolution, however, can lead to improved understanding of the underlying DNA sequence and the structure of protein-coding genes.

The introduction of Sanger sequencing in 1977[15] and PCR in the mid 1980s[57] led to the development of Sequence-Specific Oligonucleotide (SSO) probes, which hybridize specific sequences. SSO has enabled the detection of specific HLA alleles,[58] and these techniques were further developed into Sequence Specific Primers (SSP). SSO and SSP are both considered to be low-intermediate resolution, as multiple alleles might be detected by a single probe. SSO, SSP, and Sanger Sequencing require more complex analysis software, but since these techniques use probes that target individual DNA sequences, they allow analysis of localized polymorphisms, and typing of individualized SNPs. With Sanger sequencing, combined with clever primer design and PCR strategies, it is also feasible to analyze targeted genes for specific analysis.[59]

The IPD-IMGT/HLA[40] database is the standard repository for reference sequences of HLA alleles. As sequencing technologies have improved, submissions of allele data to the repository has increased at high rates. Contrary to other nucleotide databases, HLA alleles in IPD-IMGT/HLA are given official names according to the World Health Organization (WHO) nomenclature committee.[60] HLA allele names are divided into four fields, which are assigned to represent polymorphism that determines the encoded protein, as well as polymorphism in the intron and UTR regions. After the HLA gene name, the first field represents an allele group, which often correlates with serological typing. The second field distinguishes polymorphism that encodes a unique protein, while the third field distinguishes synonymous exon-level polymorphism which does not encode an amino acid difference. The fourth field encodes non-coding differences, i.e. intron and UTR polymorphism. As determined by the WHO nomenclature committee, class I sequences must, at a minimum, be represented by unique nucleotide sequences for exons 2 and 3, which form the antigen presentation domain of the HLA molecule. Likewise, class II alleles must be represented by at minimum an exon 2 sequence. This represents a minimum requirement for the amount of data that must be included to represent the polymorphism within an HLA allele.

The HLA Dictionary[61] is a collection of HLA alleles with their corresponding serological typing. The dictionary is a manually constructed set of data generated in a combined effort from laboratories across the world. Initial versions began as a book-based dictionary, but it has advanced into software-based tools. It is now feasible and common to download a digital PDF containing the known serological subtypes of a large number of HLA alleles, and a digital copy of the data dictionary is available at the IPD-IMGT/HLA website. Although the HLA dictionary is not complete, as the full set of mappings of HLA alleles to their serotypes are not available, some attempts have been made in predicting serotypes based on amino acid sequences using machine learning methods.[62] Efforts like this are becoming more difficult, because the variety in HLA sequences and lack of known serological equivalents provide challenges in identifying the serological function of unknown HLA molecules.

The National Marrow Donor Program has performed extensive studies on how population diversity is related to HLA. This is of critical importance, because certain ethnicities have much higher probabilities of finding an HLA-compatible donor.[63] The NMDP provides publicly accessible data on HLA allele and Haplotype Frequencies,[54,64] and much of the available data shows differences between ethnicities. The NMDP has put a lot of effort into the generation of data standards and microservices,[65-68] that are intended to ensure the transmission of high-quality genotypes, and enable better analysis of HLA data.

HLA Matching for Transplantation

Both stem cell transplantation, as well as solid organ transplantation involve at their core the introduction of non-self cells to an immune system, which is trained to respond to such foreign material. This creates complications that must be overcome, in order for the transplantation to be successful. Immunosuppressive drugs have been developed to suppress the immune system's natural response, but these drugs do not act specifically. Weakening an immune system leaves an individual susceptible to threats that are encountered in daily life, where exposure to normally trivial pathogens presents an increased risk of disease. A better strategy is to disguise the non-self tissue and trick the immune system into identifying non-self tissue as self.

Matching of HLA alleles in a transplantation setting allows an immune system to recognize non-self tissue as self, thereby reducing the effects of an immune response. Serological methods can distinguish between HLA antigens, but matching based on these methodologies have widely been replaced by DNA-based methods. DNA-based matching can be performed at varying resolutions based on fields of HLA nomenclature. Matching of the first field, the allele group, is effectively the same as antigen level matching in many cases. Matching on the second field, which represents coding exon polymorphism, ensures that the complete protein sequence is identical. The third and fourth fields

represent polymorphism that are not reflected in the protein sequence, but matching on these fields provides a higher resolution patient-donor comparison.

For solid organ transplantations, the identification of donor-specific HLA antibodies in the recipient provides the most important contra-indicator for transplantation, and high resolution typing of HLA is useful for defining the HLA epitopes. Since epitopes act as a target for Donor Specific Antibodies (DSAs),[69] and epitopes are defined by DNA sequence, molecular analysis gives insights into the actual immunogenic components of the HLA molecule. This indicates that typing and matching of HLA alleles in high resolution is an important consideration in solid organ transplantations.

In the Stem Cell Transplantation setting, typing of HLA based on the DNA sequence is a primary consideration.[70] Matching of HLA alleles provides vast improvements to outcomes. Due to the difficulty and cost in sequencing extended regions of DNA, especially the complex regions within introns, many typing methods are not based on full length sequences. They are likely focused on exons 2 and 3 in class I loci, and exon 2 for HLA class II. These regions encode the antigen presentation domain of the HLA molecule, and are also the most polymorphic regions, and therefore typing of these regions provides the most information about the behavior of the HLA antigen and subsequent biological behavior.

It has also been shown, however, that in addition to exon matching, high resolution matching of HLA types results in improved SCT outcomes.[71] There are many potential reasons for this. Full-length HLA sequences provides more complete information about the entire gene, rather than just the peptide sequences of the antigen presentation domain. The exons outside of the antigen presentation domain do affect the behavior of the protein, and polymorphism within the trans-membrane region or the leader peptide can affect the function as a cell surface protein.[72] Sequence within the introns and UTRs may affect the behavior of promoter sequences, or may interact with microRNAs,[73] either as an encoding or target sequence. In addition to high resolution typing of full-length HLA sequences, there are further improvements to outcomes when extended HLA haplotypes are considered. Petersdorf *et. al.* have shown that matching of extended HLA haplotypes in addition to HLA alleles provides improved outcomes in SCT matching.[74,75] The exact mechanisms of the improvements from the haplotype effect are not fully elucidated. It may be related to cooperative interactions between protein heterodimers encoded on the same haplotype, or due to linkage disequilibrium, where matching of HLA alleles results in the matching of sequence that was not explicitly sequenced and typed.

There are biological and bioinformatic approaches that can improve or ease analysis of haplotypes. Haplotypes can be identified by physically linking polymorphism using

laboratory techniques,[76] and population genetics software such as PyPop[27] can identify the most likely haplotype patterns. The most promising approaches may be capture techniques, which can be designed for sequencing and analysis of the entire MHC region at once. Region Specific Extraction,[77] a technique that uses oligonucleotide probes and biotinylated nucleotides to enrich and extract sequence across the MHC, has had some success in identifying haplotype differences between individuals. It has also been useful in identifying HLA allele differences,[78] which could give clues to evolutionary lineages. There are also attempts to analyze HLA and MHC data from whole-genome sequencing data.[79] Every attempt to identify and match polymorphism in the HLA is performed with the goal of identifying the most relevant polymorphism linked to transplantation outcome, or correlations with disease.

Immunogenetics

A human is constantly under attack from outside threats, whether it be viruses, bacteria, parasites, or other pathogens. In the course of human evolution, the immune system has evolved to detect and respond to the presence of non-self pathogens. Due to the role the immune system plays, it is not surprising that it has been linked to a variety of human diseases. A critical component of the immune system is the HLA molecule, which is responsible for antigen presentation. While class I is primarily responsible for presenting intracellular peptides, class II often presents peptides derived from extracellular pathogens, which have been phagocytosed and digested by immune cells. HLA class I and II are recognized by immune cells, *e.g.* T cells and NK cells, resulting in a cell-mediated immune response,[80] but HLA can also be a target of antibodies and B cells, which can trigger a humoral response.[81] The immune system has evolved to trigger these responses against non-self antigens, and reactions against self may indicate an autoimmune disease.

Determining immunogenicity can be performed using high-resolution sequencing techniques. High resolution sequencing defines the amino acid sequence of the HLA protein, and even small differences in amino acid sequences may result in an immunogenic response.[82] It is even likely that the immunogenic epitopes may be a combination of amino acids on both of the two proteins in a heterodimer.[83] Techniques to match HLA alleles based on direct analysis of epitopes is a promising approach to determine transplant compatibility.[84] Algorithms such as PIRCHE[85] can predict epitopes which are presented through indirect antigen presentation pathways. These epitopes are also likely to trigger an immune response, and has been shown to be important in determining what epitopes may affect transplantation outcome.[86]

Defining HLA polymorphism is important for reasons besides transplantation as well. Polymorphism within the MHC has been linked to a variety of diseases, drug sensitivity, or antibiotic allergies. These linkages are often identified through GWAS studies or statistical

methods, but it can be initially unclear if the polymorphism is directly linked to disease-causing biological differences, or if the association is in linkage disequilibrium with a functional polymorphism elsewhere.

Multiple Sclerosis and its links with the MHC are well established.[87] There are well-known haplotype patterns that are linked with the progression[88,89] of or protection[90] from the disease. The correlation of multiple sclerosis are originally linked to specific SNPs by statistical methods, but the link between these polymorphisms and the causality of the disease is still being explored. Structural differences HLA-DRB1*15:01 molecule may indicate differences that lead to disease,[91] but some evidence suggests that a SNP within HLA-DRA (rs8084), which can create an alternative splice variant, may also affect the HLA-DR heterodimer and play a role in disease causality.

Molecular analysis is also important to define SNPs that can affect protein expression. Most analysis of gene expression has been correlated with the 5′ UTR,[92] which contains gene promoter sequences, but there is growing evidence that the 3′ UTR can also affect expression.[93-95] The rs9277534 SNP within the 3′ UTR of HLA-DRB1 has been linked to higher or lower expressed HLA-DPB1 alleles.[96] The mechanism is not completely clear, but there is some evidence that microRNAs may interact with the region containing this SNP.[97] All of these correlations indicate that high-resolution molecular analysis is critical to understanding HLA, the MHC and how polymorphism defines immunogenicity.

The Future

Molecular analysis of HLA and their role in the MHC is an ever-evolving process. Categorizing the function of HLA began with serological methods, which identify the expressed molecules and reactivity of anti-HLA antibodies at low resolution. The transition to molecular analysis, focused on exon analysis provides more detailed polymorphism data, and a new nomenclature which defines polymorphism based on serological equivalents, but also creates more challenges in analysis. As focus switches from exon analysis to full-length gene analysis, more discoveries are made about the function of noncoding DNA and its role in transplantation and diagnostics. Further enlightenment is provided by increasing resolution from individual genes to studies of linkage disequilibrium and MHC haplotypes, which are extended to the relationship of the MHC with the entire human genome. These advances create new ways of thinking, create new scientific questions, and require new techniques for analysis. Bioinformatics plays a critical role in the creation and answering of these questions, and will continue to be present in the field of molecular analysis and immunogenetics in the future.

References

1. Bayat A. Science, medicine, and the future: Bioinformatics. *BMJ (Clinical research ed).* 2002;324(7344):1018-1022.
2. Heard E, Martienssen Robert A. Transgenerational Epigenetic Inheritance: Myths and Mechanisms. *Cell.* 2014;157(1):95-109.
3. Consortium IHGS. Finishing the euchromatic sequence of the human genome. *Nature.* 2004;431(7011):931-945.
4. Anderson S, Bankier AT, Barrell BG, *et al*. Sequence and organization of the human mitochondrial genome. *Nature.* 1981;290(5806):457-465.
5. Luo S, Valencia CA, Zhang J, *et al*. Biparental Inheritance of Mitochondrial DNA in Humans. *Proceedings of the National Academy of Sciences.* 2018;115(51):13039-13044.
6. Venter JC, Adams MD, Myers EW, *et al*. The sequence of the human genome. *Science.* 2001;291(5507):1304-1351.
7. Graur D. An Upper Limit on the Functional Fraction of the Human Genome. *Genome Biology and Evolution.* 2017;9(7):1880-1885.
8. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57-74.
9. di Iulio J, Bartha I, Wong EHM, *et al*. The human noncoding genome defined by genetic diversity. *Nat Genet.* 2018;50(3):333-337.
10. Crick FH. On protein synthesis. *Symp Soc Exp Biol.* 1958;12:138-163.
11. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell.* 2009;136(2):215-233.
12. Schalk A, Greff G, Drouot N, *et al*. Deep intronic variation in splicing regulatory element of the ERCC8 gene associated with severe but long-term survival Cockayne syndrome. *Eur J Hum Genet.* 2018;26(4):527-536.
13. The Genomes Project C, Auton A, Abecasis GR, *et al*. A global reference for human genetic variation. *Nature.* 2015;526:68.
14. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian J Microbiol.* 2016;56(4):394-404.
15. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America.* 1977;74(12):5463-5467.
16. Bentley DR, Balasubramanian S, Swerdlow HP, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 2008;456(7218):53-59.
17. Merriman B, Rothberg JM. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis.* 2012;33(23):3397-3417.
18. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics.* 2015;13(5):278-289.
19. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology.* 2016;17(1):239.
20. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology.* 1970;48(3):443-453.

21. Smith TF, Waterman MS. Identification of common molecular subsequences. *Journal of Molecular Biology.* 1981;147(1):195-197.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410.
23. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research.* 1994;22(22):4673-4680.
24. Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience.* 2014;3:22-22.
25. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094-3100.
26. Niu T. Algorithms for inferring haplotypes. *Genet Epidemiol.* 2004;27(4):334-347.
27. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update – a software pipeline for large-scale multilocus population genomics. *Tissue Antigens.* 2007;69(s1):192-197.
28. Do CB, Batzoglou S. What is the expectation maximization algorithm? *Nature Biotechnology.* 2008;26(8):897-899.
29. Cock PJA, Antao T, Chang JT, *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 2009;25(11):1422-1423.
30. Lander ES, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860-921.
31. KA W. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcostsdata. Accessed Sept 12, 2019.
32. Venter JC, Remington K, Heidelberg JF, *et al*. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science.* 2004;304(5667):66-74.
33. Leininger S, Urich T, Schloter M, *et al*. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature.* 2006;442(7104):806-809.
34. Wang W-L, Xu S-Y, Ren Z-G, Tao L, Jiang J-W, Zheng S-S. Application of metagenomics in the human gut microbiome. *World journal of gastroenterology.* 2015;21(3):803-814.
35. Worobey M. Molecular mapping of Zika spread. *Nature.* 2017;546(7658):355-356.
36. Quick J, Loman NJ, Duraffour S, *et al*. Real-time, portable genome sequencing for Ebola surveillance. *Nature.* 2016;530(7589):228-232.
37. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 2009;37(9):e67.
38. David M, Mustafa H, Brudno M. Detecting Alu insertions from high-throughput sequencing data. *Nucleic acids research.* 2013;41(17):e169-e169.
39. Leffler EM, Gao Z, Pfeifer S, *et al*. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science.* 2013;339(6127):1578.
40. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. *Nucleic Acids Research.* 2019;48(D1):D948-D955.

41. Cullen M, Noble J, Erlich H, *et al*. Characterization of recombination in the HLA class II region. *Am J Hum Genet.* 1997;60(2):397-407.
42. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. High-Resolution Patterns of Meiotic Recombination across the Human Major Histocompatibility Complex. *The American Journal of Human Genetics.* 2002;71(4):759-776.
43. Lam TH, Shen M, Chia JM, Chan SH, Ren EC. Population-specific recombination sites within the human MHC region. *Heredity (Edinb).* 2013;111(2):131-138.
44. Dumont BL, Payseur BA. EVOLUTION OF THE GENOMIC RATE OF RECOMBINATION IN MAMMALS. *Evolution.* 2008;62(2):276-294.
45. Petersdorf EW. In celebration of Ruggero Ceppellini: HLA in transplantation. *HLA.* 2017;89(2):71-76.
46. Petersdorf EW, Stevenson P, Malkki M, *et al*. Patient HLA Germline Variation and Transplant Survivorship. *J Clin Oncol.* 2018;36(24):2524-2531.
47. Committee WN. Nomenclature for factors of the HL-a system. *Bull World Health Organ.* 1968;39(3):483-486.
48. Xie T, Rowen L, Aguado B, *et al*. Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse. *Genome research.* 2003;13(12):2621-2636.
49. Robinson J, Guethlein LA, Cereb N, *et al*. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genet.* 2017;13(6):e1006862.
50. Gruen JR, Weissman SM. Evolving Views of the Major Histocompatibility Complex. *Blood.* 1997;90(11):4252-4265.
51. Chaix R, Cao C, Donnelly P. Is Mate Choice in Humans MHC-Dependent? *PLOS Genetics.* 2008;4(9):e1000184.
52. Pirastu N, Kooyman M, Traglia M, *et al*. Genome-wide association analysis on five isolated populations identifies variants of the HLA-DOA gene associated with white wine liking. *Eur J Hum Genet.* 2015;23(12):1717-1722.
53. Marsh SGE, Parham P, Barber LD. 3 - The Organization of HLA Genes Within the HLA Complex. In: Marsh SGE, Parham P, Barber LD, eds. *The HLA FactsBook.* London: Academic Press; 2000:7-13.
54. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human Immunology.* 2013;74(10):1313-1320.
55. Degli-Esposti MA, Leaver AL, Christiansen FT, Witt CS, Abraham LJ, Dawkins RL. Ancestral haplotypes: conserved population MHC haplotypes. *Human Immunology.* 1992;34(4):242-252.
56. Tilanus MGJ, van Eggermond MCJA, van der Bijl M, *et al*. HLA class II DNA analysis by RFLP reveals novel class II polymorphism. *Human Immunology.* 1987;18(4):265-276.
57. Saiki R, Scharf S, Faloona F, *et al*. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science.* 1985;230(4732):1350-1354.
58. Erlich H. HLA DNA typing: past, present, and future. *Tissue Antigens.* 2012;80(1):1-11.

59. Voorter CEM, Palusci F, Tilanus MGJ. Sequence-Based Typing of HLA: An Improved Group-Specific Full-Length Gene Sequencing Approach. In: Beksaç M, ed. *Bone Marrow and Stem Cell Transplantation*. New York, NY: Springer New York; 2014:101-114.
60. Marsh SGE, Albert ED, Bodmer WF, *et al*. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens.* 2010;75(4):291-455.
61. Holdsworth R, Hurley CK, Marsh SGE, *et al*. The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. *Tissue Antigens.* 2009;73(2):95-170.
62. Maiers M, Schreuder GMT, Lau M, *et al*. Use of a neural network to assign serologic specificities to HLA-A, -B and -DRB1 allelic products. *Tissue Antigens.* 2003;62(1):21-47.
63. Gragert L, Eapen M, Williams E, *et al*. HLA match likelihoods for hematopoietic stem-cell grafts in the U.S. registry. *N Engl J Med.* 2014;371(4):339-348.
64. Gonzalez-Galarza FF, Takeshita LY, Santos EJ, *et al*. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.* 2015;43(Database issue):D784-788.
65. Mack SJ. A GENE FEATURE ENUMERATION APPROACH FOR DESCRIBING HLA ALLELE POLYMORPHISM. *Human immunology.* 2015;76(12):975-981.
66. Milius RP, Mack SJ, Hollenbach JA, *et al*. Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string. *Tissue Antigens.* 2013;82(2):106-112.
67. Milius RP, Heuer M, Valiga D, *et al*. Histoimmunogenetics Markup Language 1.0: Reporting next generation sequencing-based HLA and KIR genotyping. *Human immunology.* 2015;76(12):963-974.
68. Mack SJ, Milius RP, Gifford BD, *et al*. Minimum information for reporting next generation sequence genotyping (MIRING): Guidelines for reporting HLA and KIR genotyping via next generation sequencing. *Human immunology.* 2015;76(12):954-962.
69. El-Awar N, Jucaud V, Nguyen A. HLA Epitopes: The Targets of Monoclonal and Alloantibodies Defined. *Journal of Immunology Research.* 2017;2017:3406230.
70. Fürst D, Müller C, Vucinic V, *et al*. High-resolution HLA matching in hematopoietic stem cell transplantation: a retrospective collaborative analysis. *Blood.* 2013;122(18):3220-3229.
71. Mayor NP, Hayhurst JD, Turner TR, *et al*. Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. *Biology of Blood and Marrow Transplantation.* 2019;25(3):443-450.
72. Gerritsen K. *HLA-C exon 5 skipping by alternative splicing.* 2016.
73. Clark PM, Chitnis N, Shieh M, Kamoun M, Johnson FB, Monos D. Novel and Haplotype Specific MicroRNAs Encoded by the Major Histocompatibility Complex. *Scientific reports.* 2018;8(1):3832-3832.
74. Petersdorf EW, Malkki M, Horowitz MM, Spellman SR, Haagenson MD, Wang T. Mapping MHC haplotype effects in unrelated donor hematopoietic cell transplantation. *Blood.* 2013;121(10):1896-1905.

75. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Med.* 2007;4(1):e8.
76. Guo Z, Hood L, Malkki M, Petersdorf EW. Long-range multilocus haplotype phasing of the MHC. *Proc Natl Acad Sci U S A.* 2006;103(18):6964-6969.
77. Dapprich J, Ferriola D, Mackiewicz K, *et al*. The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC genomics.* 2016;17:486-486.
78. Steiner NK, Hou L, Hurley CK. Characterizing alleles with large deletions using region specific extraction. *Hum Immunol.* 2018;79(6):491-493.
79. Hayashi S, Yamaguchi R, Mizuno S, *et al*. ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data. *BMC Genomics.* 2018;19(1):790.
80. Zino E, Frumento G, Marktel S, *et al*. A T-cell epitope encoded by a subset of HLA-DPB1 alleles determines nonpermissive mismatches for hematologic stem cell transplantation. *Blood.* 2004;103(4):1417-1424.
81. Duquesnoy RJ. A Structurally Based Approach to Determine HLA Compatibility at the Humoral Immune Level. *Human Immunology.* 2006;67(11):847-862.
82. Hurley CK, Steiner N. Differences in peptide binding of DR11 and DR13 microvariants demonstrate the power of minor variation in generating DR functional diversity. *Human Immunology.* 1995;43(2):101-112.
83. Hollenbach JA, Madbouly A, Gragert L, *et al*. A combined DPA1~DPB1 amino acid epitope is the primary unit of selection on the HLA-DP heterodimer. *Immunogenetics.* 2012;64(8):559-569.
84. Tambur AR. HLA-Epitope Matching or Eplet Risk Stratification: The Devil Is in the Details. *Frontiers in Immunology.* 2018;9(2010).
85. Otten HG, Calis JJ, Kesmir C, van Zuilen AD, Spierings E. Predicted indirectly recognizable HLA epitopes presented by HLA-DR correlate with the de novo development of donor-specific HLA IgG antibodies after kidney transplantation. *Hum Immunol.* 2013;74(3):290-296.
86. Lachmann N, Niemann M, Reinke P, *et al*. Donor–Recipient Matching Based on Predicted Indirectly Recognizable HLA Epitopes Independently Predicts the Incidence of De Novo Donor-Specific HLA Antibodies Following Renal Transplantation. *American Journal of Transplantation.* 2017;17(12):3076-3086.
87. Hollenbach JA, Oksenberg JR. The immunogenetics of multiple sclerosis: A comprehensive review. *J Autoimmun.* 2015;64:13-25.
88. Oksenberg JR, Barcellos LF, Cree BAC, *et al*. Mapping Multiple Sclerosis Susceptibility to the HLA-DR Locus in African Americans. *American Journal of Human Genetics.* 2004;74(1):160-167.
89. Morrison BA, Ucisik-Akkaya E, Flores H, Alaez C, Gorodezky C, Dorak MT. Multiple sclerosis risk markers in HLA-DRA, HLA-C, and IFNG genes are associated with sex-specific childhood leukemia risk. *Autoimmunity.* 2010;43(8):690-697.
90. Mack SJ, Udell J, Cohen F, *et al*. High resolution HLA analysis reveals independent class I haplotypes and amino-acid motifs protective for multiple sclerosis. *Genes Immun.* 2018;20:308-326.

91. Misra MK, Damotte V, Hollenbach JA. Structure-based selection of human metabolite binding P4 pocket of DRB1*15:01 and DRB1*15:03, with implications for multiple sclerosis. *Genes & Immunity*. 2018.
92. Thomas R, Apps R, Qi Y, *et al*. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet*. 2009;41(12):1290-1294.
93. Kulkarni S, Savan R, Qi Y, *et al*. Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature*. 2011;472(7344):495-498.
94. Svendsen SG, Nilsson LL, Djurisic S, *et al*. Extended HLA-G haplotypes in patients with age-related macular degeneration. *HLA*. 2018.
95. Craenmehr MHC, Haasnoot GW, Drabbels JJM, *et al*. Soluble HLA-G levels in seminal plasma are associated with HLA-G 3'UTR genotypes and haplotypes. *HLA*. 2019;94(4):339-346.
96. Petersdorf EW, Malkki M, O'hUigin C, *et al*. High HLA-DP Expression and Graft-versus-Host Disease. *New England Journal of Medicine*. 2015;373(7):599-609.
97. Shieh M, Chitnis N, Clark P, Johnson FB, Kamoun M, Monos D. Computational assessment of miRNA binding to low and high expression HLA-DPB1 allelic sequences. *Hum Immunol*. 2019;80(1):53-61.

Outline of the Thesis

The applications of bioinformatics to answer real-world biological questions requires a good foundation in supportive techniques and technologies. Generating meaningful and high-quality data enables high quality analysis. This thesis is divided into two main themes. **Chapters 2-5** describe techniques related to the generation and analysis of HLA sequencing data. **Chapters 6-9** describe answers to scientific questions that are obtained by using bioinformatic approaches to evaluate and interpret data obtained by using available sequencing techniques.

For accurate and meaningful analysis of HLA polymorphism, it is necessary to have the availability of high quality full-length reference sequences. **Chapter 2** discusses Saddlebags, the software tool for the submission of full-length allele sequences to the EMBL/ENA database. This step simplifies the process of including an HLA allele in the IPD-IMGT/HLA database, the standard database for HLA reference sequences.

For laboratories to unambiguously communicate an allele typing with each other, it is necessary to have a community-wide consensus on reference sequences and nomenclature. The International HLA and Immunogenetics Workshop is an opportunity for members of the community to come together and cooperate to improve the inter-laboratory communication. Comparison and analysis of HLA sequences employ the IPD-IMGT/HLA database as a standard reference. **Chapter 3** is a report on the full-length sequencing and genotyping of HLA allele sequences, and submission to IPD-IMGT/HLA.

Chapter 4 is focused on the validation of the HLA genotyping method using the Oxford Nanopore MinION single-molecule sequencer. The use of long read sequencing to unambiguously type HLA alleles has a lot of potential, but since nanopore sequencing is a new technology with a unique read quality profile, the HLA typing using this platform must be validated. This chapter describes the validation procedure and how it was implemented into clinical diagnostics.

In order to type HLA alleles, it is first necessary to accurately amplify the genes, without losing or misrepresenting the data. **Chapter 5** describes a multiplexed HLA amplification protocol for 11 HLA loci, and how it was validated for specificity and robustness. The resulting amplicons can be used in a variety of NGS typing protocols, and therefore represents an important step in the process of generating unambiguous HLA typing.

A common scientific theme in this thesis is HLA haplotypes. Individual HLA genes do not act alone, but are part of an intricate system. **Chapter 6** is about the HLA-DRA gene, which has been commonly thought to be monomorphic due to its limited protein polymorphism.

This study demonstrates that the identification of full length allele polymorphism within this gene forms distinct patterns with HLA-DRB1, HLA-DRB3,4,5, and HLA-DQB1. These patterns extend and redefine the previous assumptions of HLA haplotype patterns, and indicates a need to change from a static view of haplotypes to a more flexible MHC gene organization.

The way that HLA-DRA relates to HLA haplotypes is most apparent in haplotypes that contain HLA-DRB1*13 alleles. These alleles have been shown to have reduced immunogenicity, which cannot be explained by the HLA-DR13 antigens alone. **Chapter 7** shows that including HLA-DRA in HLA-DR~HLA-DQ haplotypes extends our current understanding of these sequences, and the realization that HLA-DRB1*13 alleles do not have defining epitopes may help to explain why HLA-DR13 has such enigmatic immunogenicity. This leads to a conclusion that haplotype patterns in the MHC in general may be more flexible than what is commonly understood.

The HLA-DP region is unique compared to the rest of the HLA loci. HLA-DPB1 and HLA-DPA1 are situated in a head-to-head orientation, with 2.5 kb of intergenic sequence which contains the promoter regions of both genes. HLA-DPB1 nomenclature is unique in that allele groups are not linked with serological subtypes, and instead alleles are assigned their own allele group and named based on the order of discovery. **Chapter 8** describes the sequencing a panel of full-length HLA-DPA1~promoter~HLA-DPB2 sequences. Clustering of these haplotypes by the intergenic 5′ HLA-DP sequence shows distinct patterns within the hypervariable regions of HLA-DPB1. This redefines our understanding of HLA-DPB1 haplotypes, and could lead to a more functional HLA-DP nomenclature.

Early typing of HLA was performed by serological methods, which are defined by how antibodies or immune cells interact with epitopes on an HLA antigen. These methods have widely been surpassed by sequence-based methodology, but determining the relationship between sequence and serology is an important question that is not fully understood. **Chapter 9** describes an approach to predict the serological subtype of HLA-B15 alleles which have not been tested by serological methods.

Each of these studies have an important place in the HLA field, as well as general scientific and biological studies. Together, they form a story about how bioinformatics can be used to create scientific queries, and how it can be applied to HLA research questions. This sheds some light on how HLA sequence defines the HLA antigen, and how HLA alleles relate to the full MHC haplotypes.

Part 1.
Rules and Tools of HLA Analysis

CHAPTER 2

2

Saddlebags: A software interface for submitting full-length HLA allele sequences to the EMBL-ENA nucleotide database.

B.M. Matern, M. Groeneweg, C.E.M. Voorter, M.G.J. Tilanus

Transplantation Immunology, Tissue Typing Laboratory, Maastricht University Medical Center, Maastricht, The Netherlands

HLA. 2018 Jan;91(1):29-35. doi: 10.1111/tan.13179.

Abstract

Abstract

Submission of full-length HLA allele sequences presents a unique challenge, both for high-throughput sequencing laboratories and smaller diagnostic laboratories. HLA's extensive polymorphism means that accurate representation and annotation of allele sequence is of critical importance, and curators of nucleotide databases must establish submission formats to ensure high-quality data and prevent ambiguities. The IPD-IMGT/HLA database is established as the standard repository for HLA sequences, and it is a major goal of the 17th International HLA and Immunogenetics Workshop to fill the IPD-IMGT/HLA database with full-length HLA sequences. The process of preparing sequence annotation and metadata is cumbersome and error prone, and it is desirable to create a straightforward and concise method of preparing sequence submissions. We introduce Saddlebags, a software tool for rapid generation of HLA (novel) full-length allele sequence submissions. HLA allele sequences are submitted first to EMBL European Nucleotide Archive (EMBL-ENA), and metadata is gathered for subsequent preparation of an IPDIMGT/HLA formatted submission. Combining these steps into a pipeline reduces effort and minimizes errors for submitting laboratories. This software has been used by Maastricht University Medical Center Transplantation Immunology Laboratory to submit 79 novel alleles to EMBL-ENA, and the tool is freely available for the HLA community.

Introduction

Full-Length HLA Sequences

Historically, full-gene sequencing of HLA genes has been difficult and expensive[1]. Modern NGS[2,3] and single molecule sequencing technologies[4], such as MinION[5], allow laboratories to more easily sequence a full-length gene, rather than just exon regions[6]. One of the major goals of the 17th International HLA and Immunogenetics Workshop is to fill the IPD-IMGT/HLA allele database with full-length sequences, while replacing or confirming incompletely defined sequences. This collection of full-length HLA genomic sequences is critical for full-length sequence analysis and HLA allele assignment. Simplifying the submission process encourages laboratories to submit high-quality novel and confirmatory sequences, when otherwise they might not provide their sequences due to the human effort required.

Submitting sequences to IPD-IMGT/HLA is a process of several steps, outlined in figure 1. This pipeline is non-trivial, and effort is required in gathering the sequence data and metadata. Specifically designed software, such as Saddlebags, can alleviate these efforts. Saddlebags performs the necessary intermediate step of submission to EMBL-ENA, to ease the process of inclusion in IPD-IMGT/HLA.

IPD-IMGT/HLA Database

The IPD-IMGT/HLA database[7] is a centralized repository for immunological data, especially HLA nucleotide sequences. IPD-IMGT/HLA maintains a curated database of HLA allele sequences. Each HLA allele is assigned an informative name using standard HLA nomenclature[8], which has been designed to be unambiguous and informative. IPD-IMGT/HLA represents a centralized and standardized repository for HLA knowledge, and it is a resource for acquiring standard reference sequences. A variety of software analysis tools are dependent on the sequences they provide.

Centralizing HLA sequence information allows for analysis techniques that are unavailable when data is scattered and inconsistent. IPD-IMGT/HLA can be leveraged by researchers to analyze alignments, find patterns, and understand the behavior of the MHC system. This enables more specific studies on whole-gene polymorphism in HLA that would have otherwise been impossible.

IPD-IMGT/HLA has a more strict set of submission guidelines when compared with general nucleotide repositories. For an HLA sequence to be included in the IPD-IMGT/HLA database, it must satisfy sequence quality standards, and provide all the required additional information related to the sample. The most tangible requirement is that the sequence must be included in another publically available database. This requirement is satisfied by EMBL-ENA, or any of the other partner archive in the International Nucleotide

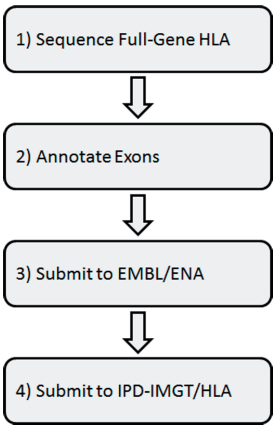


Figure 1. **Allele submission to major databases requires several steps.** Full length HLA sequences are provided by the submitter. Saddlebags facilitates submission at EMBL/ENA and prepares for subsequent submission at IPD-IMGT/HLA

Sequence Database Collaboration (http://www.insdc.org/), an international initiative to share data amongst major databases. This software uses EMBL-ENA as the intermediate sequence repository, where submissions are uploaded to acquire an accession number, which is necessary for the sequence to be submitted to IPD-IMGT/HLA.

EMBL-ENA Database

The European Molecular Biology Laboratory (EMBL, http://www.ebi.ac.uk/) is responsible for maintaining the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena). EMBL-ENA is a repository for a variety of biological nucleotide sequences, including HLA[9]. The database is used by a variety of research and diagnostics laboratories, and they receive sequence submissions from a range of sources.

EMBL-ENA accepts sequence submissions via REpresentational State Transfer (REST, http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm). REST is a software architecture style, which is used to transfer and process data over standard HTTP protocols. REST enables laboratories to submit formatted sequences to EMBL-ENA, and receive validation feedback quickly. This allows laboratories to submit data using custom software tools, such as Saddlebags, as an alternative to standard web interfaces.

Material and Methods

Saddlebags: The Software

Saddlebags was designed to streamline the process of preparing allele submissions for IPD-IMGT/HLA. The Saddlebags software is open source, and available at our department's Github page (https://github.com/transplantation-immunology-maastricht/saddle-bags). It was released under the GNU General Purpose License 3.0 (https://www.gnu.org/licenses/gpl-3.0.en.html), which allows users to modify the software for their own purposes, while promoting attribution. The software is written in Python 2.7, which is commonly used for bioinformatics tools, and is able to run on a variety of platforms (Windows, Mac, or Linux) with minimal setup. Protein translation is handled by Biopython (Version 1.6.8, http://biopython.org). The software includes Pyinstaller (Version 3.2, http://www.pyinstaller.org/) specification files, which allow users or developers to generate new executables for Mac and Windows.

Saddlebags presents a simple user interface, which is intuitive to users. The interface provides Saddlebags' main functionality, flat file generation and sequence submission to EMBL-ENA. An executable for windows is provided for download, so submitting laboratories will not require any scripting knowledge.

Results

Saddlebags: The User Interface

Submitters can download a Saddlebags executable for Windows, which provides the main functionality of allele submission at EMBL-ENA (Figure 1) to support downstream submission at IPD-IMGT/HLA. The executable will launch a user interface, which is easy to use, and includes only a few interactive elements for simplicity. It is designed to be used from top to bottom, for a logical flow of data.

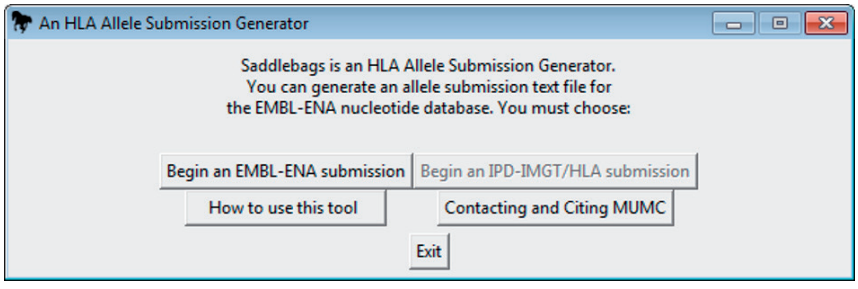


Figure 2. Saddlebags User Interface.

When the program is launched, the first window provides initial instructions to the user (Figure 2). The window provides buttons, which guide the user, and begin the process of generating a text submission for EMBL-ENA. The [How to use this tool] or [Contacting and Citing MUMC] buttons are used to gather more information about the submission process. The software was designed anticipating that in the future we will automatically generate and submit sequences to IPD-IMGT/HLA.

The process of creating an EMBL-ENA sequence submission is straightforward. Formatted sequences can be provided in the submission form (Figure 3), and the required sequence metadata is provided in a Submission Options subform (Figure 4). The automatically

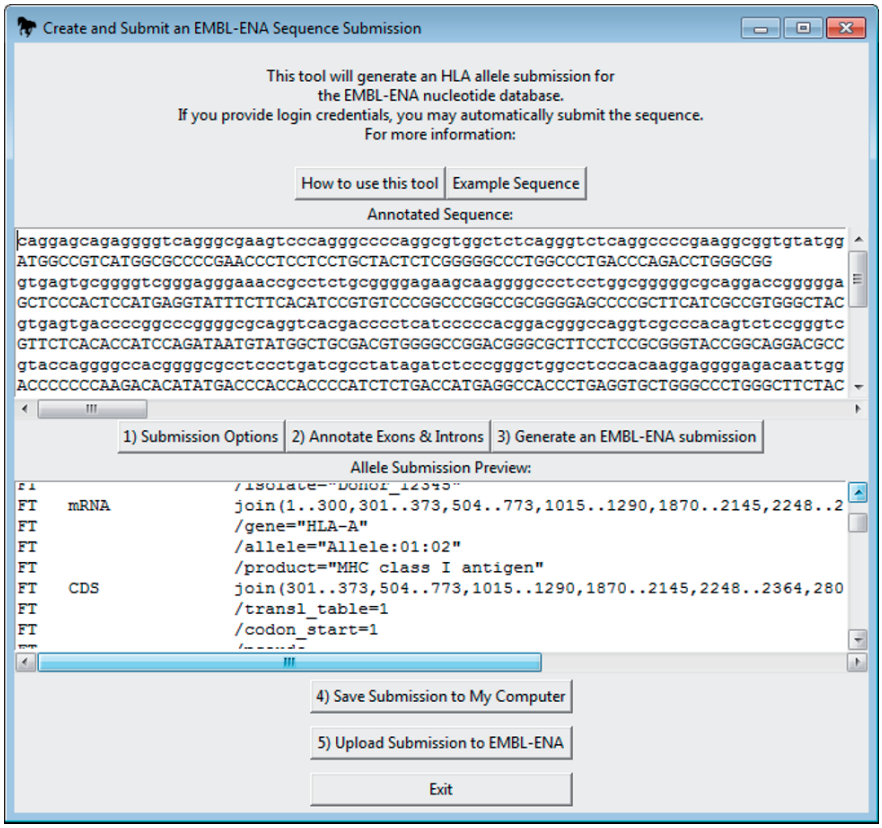


Figure 3. User Interface for generating an EMBL-ENA sequence submission. The nucleotide sequence demonstrates the correct format of a "complete" genomic HLA sequence using lowercase and uppercase nucleotides. The genomic features of the sequence were identified and formatted using NMDP BeTheMatch Allele Calling Tool / Gene Feature Enumeration services. The submission is performed over the internet via REST, and after EMBL verifies the sequence, the EMBL sequence accession number is
used for subsequent IPD-IMGT/HLA submission.

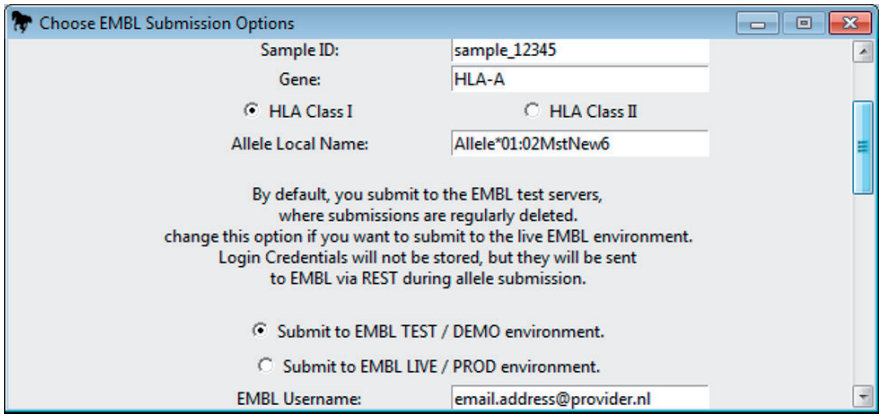


Figure 4. EMBL-ENA Sequence Submission Options. The metadata required for an EMBL/ENA submission is gathered and validated using a simple desktop input form. Configuration values (except login information) are retained for subsequent submissions.

generated flat files created for EMBL-ENA have a specific text format, established by the database curators, and metadata from the options form are used to populate the flatfile text.

An EMBL-ENA accession number is required to complete the IPD-IMGT/HLA submission, so the EMBL-ENA submission represents a necessary intermediate step. Validating the sequence is not instantaneous, a confirmation including a sequence accession number is provided by EMBL in subsequent emails.

Saddlebags stores a configuration file on your local computer to store data from recent submissions. Any relevant sequence data that is provided or calculated during EMBL-ENA submission is maintained for downstream submission to IPD-IMGT/HLA. Sequence metadata, but not user passwords, will be stored in the configuration file, to ease the process of submitting multiple sequences.

Saddlebags: Quick Start Instructions

The process of submitting an HLA allele to the IPD-IMGT/HLA database can be described in four main steps (Figure 1). Submitting laboratories will provide full-length HLA sequences for Step 1. Steps 2-3 are facilitated by the Saddlebags software. We provide here a more detailed set of instructions, so users can rapidly evaluate the tool for use in their own laboratories. These instructions assume the user has downloaded a Saddlebags executable for Windows from our department's Github page.

Step 1 - Sequence a full-length HLA gene

HLA sequences which lack intron regions may be acceptable for inclusion in the EMBL-ENA and IPD-IMGT/HLA databases, but this tool is intended for full-length HLA nucleotide

sequences, from the 5' to 3' UTRs. Sequencing methodology is the choice of the submitting laboratory, and details of the sequencing method can be included in the IPD-IMGT/HLA submission metadata.

Step 2 - Identify Sequence Features

The HLA protein function and structure depends on the exon boundaries and splicing behavior of the gene's nucleotide sequence. The exon boundaries of these regions implies biological function, and identifying the feature boundaries is required for submissions of HLA.

Saddlebags is able to identify the sequence features in the provided HLA sequence. The software will connect to the Allele Calling Tool (ACT) provided by the National Marrow Donor Program (http://act.b12x.org/ui/). The service uses an implementation of Gene Feature Enumeration[10] to provide information about intron/exon boundaries, which is interpreted by Saddlebags to provide an annotated genomic sequence. The service is able to annotate full-length HLA sequences. In cases where the ACT is unable to provide reliable exon information, the sequences can be annotated manually by the submitter.

Saddlebags: Sequence Input Format

To facilitate rapid use by a submitting laboratory, a simple sequence input format was devised. This format is automatically generated by Saddlebags' feature annotation tool, and in most cases it is not necessary to format sequences manually. Submitters can use the "Example Sequence" button on the Saddlebags interface to see a representative example of the sequence formatting style. Nucleotide sequences are entered using standard unambiguous IUPAC nucleotide characters[11]. If the sequence contains ambiguous or nonstandard nucleotide characters, Saddlebags will warn the user before submission. Saddlebags will attempt to interpret the ambiguous nucleotides, but the submission is subject to the standards of the receiving database, which do not generally allow ambiguous characters in HLA allele sequences.

The user must provide a full-length HLA allele, sequenced in the 5' to 3' direction. For consistency, the sequences are annotated and arranged in a standard input format, which assumes that an HLA allele holds a consistent general structure. The sequence must have a UTR, at both the 5' and 3' ends. Between the 5' and 3' UTRs, the gene contains alternating exons, with introns distributed between them. After the automatic annotation, genomic features are indicated by both lowercase and uppercase nucleotide characters. Exon sequences are specified using uppercase characters, while introns and UTRs are lowercase characters. Space, tab, and newline characters are completely disregarded by Saddlebags. These characters are removed before the exon boundary indexes are calculated, and are used to visually organize the allele's genomic features. Each exon or intron is placed on a separate line, to clearly distinguish the features.

The exons represent coding regions, and these sequences are translated to a peptide sequence to represent the predicted HLA protein. Saddlebags does not require a specific count of exons, but will respond if the required genomic features are missing. If a UTR or exon is missing, the interface will present a descriptive warning message with the nature and location of the problem, to allow the submitter to respond.

Manually assigning exon boundaries requires nontrivial effort and knowledge of common HLA sequences. To alleviate the efforts required in annotation, a table of the sequences surrounding the most common intron/exon boundaries was created (Table 1). This table can be accessed from the software's Github page. The submitter can find the common boundary sequences using this table, or by combining this information with the relevant alignments from the IPD-IMGT/HLA database.

Step 3 - Prepare and submit an EMBL-ENA sequence submission

To prepare the EMBL-ENA submission, the user must provide the allele sequence, critical sequence metadata, and valid login credentials for the EMBL website.

1) **Launch the Saddlebags software, and press [Begin an EMBL-ENA submission]**
   Executables for Windows are provided on our department's Github page.
2) **(Optional) Press [Example Sequence] to see an example of formatted sequence data and metadata.**
   This will discard and replace any sequence data in the form with a sample nucleotide sequence, containing a few sample exons and introns. This sample represents a formatted HLA allele, and demonstrates the correct sequence format.
3) **Provide a well-formatted HLA allele sequence.**
   This tool requires a full-length sequence, including a 5' and 3' UTR and all exons and introns. Paste this sequence into the "Annotated Sequence" field. You may annotate the genomic features automatically in subsequent steps, or annotate manually.
4) **Press [Submission Options] to provide the necessary sequence metadata.**
   Metadata, such as sample ID, HLA locus, and locally-assigned allele name are specified on this options window. You must also provide EMBL-specific information, including valid login credentials, and a reference to the Project / Study to which this sequence is to be assigned.
5) **Press [Annotate Exons & Introns] to format your nucleotide sequence.**
   This will use the NMDP / BeTheMatch ACT service to align against the relevant HLA allele, identify your genomic features, and format your input sequence to a structure that can be interpreted by Saddlebags.
6) **Press the [Generate an EMBL-ENA submission] button**
   Some sequence validation is performed, and a text submission is generated based on the input provided by the user. You may inspect the document visually, and you have

an opportunity now to verify that all your sequence data and metadata are correct. The user is able to save a local copy of the submission, if desired.

7) **Press the [Upload Submission to EMBL-ENA] button to upload your submission to EMBL.**

Saddlebags is configured to point at the "test" submission servers by default. You must use [Submission Options] to instruct Saddlebags to use the live servers, but it is recommended that you test your sequence submissions in the test environment first. If Saddlebags encounters problems in the submission step, attempts will be made to inform the user about the nature of the error. Although sequences are submitted nearly instantly, EMBL-ENA must validate the submission. If there are no problems, the EMBL sequence accession number will be provided via email. The sequence accession number is necessary for subsequent submission to IPD-IMGT/HLA.

Step 4 - Prepare and submit an IPD-IMGT/HLA sequence submission

The IPD-IMGT/HLA database requires submissions in a standard flatfile format which is similar in appearance to the EMBL-ENA flatfile. Submissions can be made using their web interface (https://www.ebi.ac.uk/ipd/imgt/hla/subs/submit.html). Future versions of Saddlebags will allow these flatfiles to be prepared and submitted directly to IPD-IMGT/HLA.

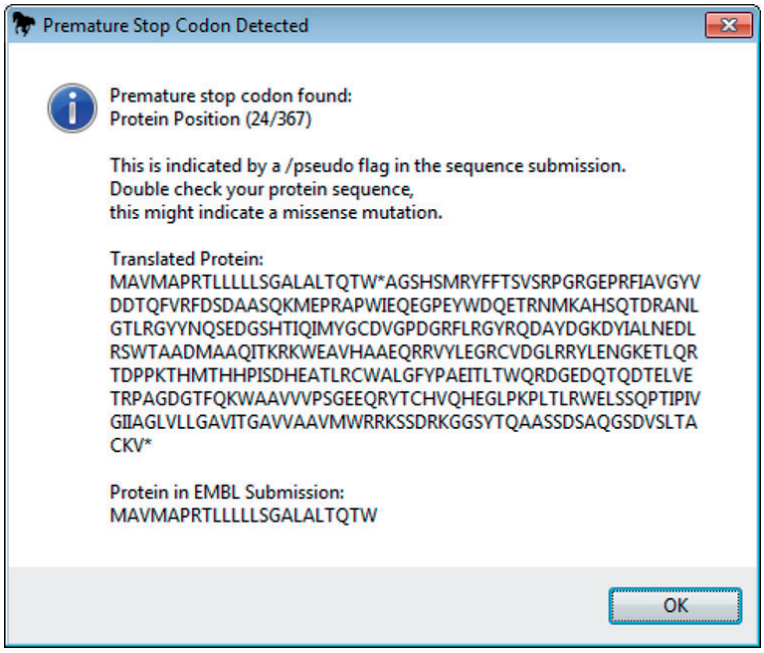


Figure 5. Example warning message for an invalid protein sequence. It identifies the location of potential issues with sequence interpretation to allow the submitter to respond.

	HLA-A	HLA-B	HLA-C
5′ UTR | Start Exon 1	CCGAGG|ATGGCC	GCCGAG|ATGCTG	GCCGAG|ATGCGG
		ACCGAG|ATGCGG	
		GCCAAG|ATGCTG	
		GCGGAG|ATGCTG	
End Exon 1 | Start Intron 1	GGGCGG|GTGAGT	GGGCCG|GTGAGT	GGGCCT|GTGAGT
	GGGCAG|GTGAGT	GGGCTG|GTGAGT	GGGCCG|GTGAGT
			GGACCG|GTGAGT
End Intron 1 | Start Exon 2	CCCCAG|GCTCCC	CCCCAG|GCTCCC	CCCCAG|GCTCCC
	CCCCAG|GCTCTC	CCC.AG|GCTCCC	
End Exon 2 | Start Intron 2	AGGACG|GTGAGT	AGGCCG|GTGAGT	AGGCCG|GTGAGTG
	AGGCCG|GTGAGT		AGGACG|GTGAGTG
			AGGCCA|GTGAGTG
End Intron 2 | Start Exon 3	GGCCAG|GTTCTC	GGCCAG|GGTCTC	GGCCAG|GGTCTC
			GGCCAG|GTTCTC
End Exon 3 | Start Intron3	GCACGG|GTACCA	GCGCTG|GTACCA	GCGCGG|GTACCA
	GCACGG|GTACCG	GCGCGG|GTACCA	GCGCAG|GTACCA
End Intron 3 | Start Exon 4	CGTCAG|ACCCCC	CATCAG|ACCCCC	CGTCAG|AACACC
	TGACAG|ACGCCC		CGTCAG|AACCCC
			CGTCAG|AACGCC
End Exon 4 | Start Intron 4	GATGGG|GTAAGG	GGAGGG|GTAAGG	GATGGG|GTAAGG
			GATGGA|GTAAGG
			GCTGGG|GTAAGG
End Intron 4 | Start Exon 5	TCCCAG|AGCTGT	TCCCAG|AGCCGT	TCCCAG|AGCCGT
	TCCCAG|AGCCGT	TCCCAG|AGCCAT	TCCCAG|AGCCAT
	TCCCAG|AGCCAT		TCCCAG|GGCCAT
End Exon 5 | Start Intron 5	GCTCAG|GTGGAG	GTTCAG|GTAGGG	GCTCAG|GTAGGG
	GCTCAG|GTGGGG	GCTCAG|GTAGGG	
End Intron 5 | Start Exon 6	CCACAG|ATAGAA	CCACAG|GTGGAA	CCACAG|GTGGAA
	CCACAG|TTAGAA		
	TCACAG|ATAGAA		
End Exon 6 | Start Intron 6	CTGCAA|GTAAGT	CTGCGT|GTAAGT	CTGCGT|GTAAGT
			TTGCGT|GTAAGT
End Intron 6 | Start Exon 7	CCCCAG|GCAGTG	CTCCAG|GCAGCG	CCCCAG|CCAGCA
		CTCCAG|CCAGCG	CCCCAG|GCAGCA
End Exon 7 | Start Intron 7	GTAAAG|GTGAGA		GTAAAG|GTGAGA
End Exon 7 | Start 3′ UTR		GCTTGA|AAAGGT	
End Intron 7 | Exon 8 | 3′ UTR	CTATAG|TGTGA|GACAGC		CTGTAG|CCTGA|GACAGC

TABLE 1. Common splice-site boundaries for class I HLA alleles

Protein Translations

As we were submitting sequences, we encountered an issue regarding stop codons. In biology, premature or atypical stop codons will change the protein expression of the gene, likely resulting in null alleles[12]. When such sequences were encountered in Saddlebags, the translated peptide sequence was invalid, and EMBL-ENA was unable to handle the submission. As a workaround, logic was added to Saddlebags to better interpret the exonic sequences. When coding sequences contain a premature stop codon, or the sequence length is not a multiple of three nucleotides, the software will provide the most correct translation available, and warn the submitter that their protein may not be what is expected (Figure 5). The warning messages clearly specify the nature and location of the unexpected sequence, allowing submitters to identify and respond to problems with their sequence.

Discussion

We have created a software tool to be used for rapid submission of full-length HLA sequences to EMBL-ENA for subsequent submission at IPD-IMGT/HLA. Submitting novel sequences, or submitting full-length sequences for previously incomplete alleles, will help to centralize HLA knowledge, allowing for more focused research. Saddlebags requires full-length sequences, and therefore promotes HLA submissions of high-quality, full-length allele sequences to the IPD-IMGT/HLA database, which is a major goal of the 17th International HLA and Immunogenetics Workshop.

Some tools have been developed for the purpose of submitting HLA sequences to standard databases. Sequin (https://www.ncbi.nlm.nih.gov/Sequin/) is available from NCBI, and can be used for submission to the Genbank database. Sequin has the advantage of providing graphical alignment and sequence feature editing functions, but does not support online submission of sequences. Sequin is intended for Genbank submission, and although Genbank accession numbers are acceptable for inclusion in IPD-IMGT/HLA submissions, EMBL-ENA is widely used as the intermediate nucleotide database. Typeloader[13] is a web-based tool developed at DKMS Life Science Lab. Although Typeloader is able to generate EMBL-formatted flatfiles, it does not perform the online REST submission as required by EMBL-ENA. The software that is available for download requires submissions in the GenDX XML format, which precludes its use in many submitting laboratories. Saddlebags provides the added value of automated submission specifically to EMBL-ENA, while not being constrained to a single sequencing or analysis platform

When we transfer genomic data, data standards are an important consideration. The format required for Saddlebags is understandable by the HLA community, but more

strictly defined data standards are desirable. To remain relevant, it is important to use standards decided by the community, to promote research and shareability of data. XML is a common data transmission format, which is extended by the well-defined HML standard[14]. HML is a structured blueprint for data, promoted by NMDP, which provides placeholders for immunogenetic sequence data, as well as relevant metadata. NMDP regularly provides tools to promote the use of standards by the HLA community, and HML is already supported by several Genomic Software vendors and HLA laboratories. In addition to Saddlebags current functionality of automated feature annotation of full-length HLA sequences, future versions of Saddlebags will support HML input, where feature annotations are either provided, or generated automatically.

The Saddlebags software is in regular use in the Maastricht University Medical Center Transplantation Immunology Laboratory, and has been used for 79 novel allele submissions[15] to date. Users are encouraged to download an executable for Windows, evaluate the software, and use for their own (novel) HLA allele submissions. Users of the Saddlebags Submission Tool are encouraged to cite this publication in their manuscripts.

Acknowledgements

The authors would like to thank Christel Meertens, Sophie Onclin and Dr. Burcu Duygu for their detailed analysis and evaluation of the Saddlebags tool, and thanks to Diana van Bakel for her contributions to manuscript submission. Thanks to the developers at EMBL for their assistance in implementing REST-based submission. Thanks to Mike Halagan at NMDP for his excellent work on the ACT and GFE tools. And finally, thanks to James Robinson for insights into IPD-IMGT/HLA submission format.

References

1. Mayor NP, Robinson J, McWhinnie AJM, Ranade S, Eng K, et al. (2015) HLA Typing for the Next Generation. PLOS ONE 10(5): e0127153.https://doi.org/10.1371/journal.pone.0127153
2. Goodwin, Sara, McPherson, John D., McCombie, W. Richard. (2016) Coming of age: ten years of next-generation sequencing technologies. Nature Reviews (2016) 17, 333. http://dx.doi.org/10.1038/nrg.2016.49
3. Duke, J. L., Lind, C., Mackiewicz, K., Ferriola, D., Papazoglou, A., Gasiewski, A., Heron, S., Huynh, A., McLaughlin, L., Rogers, M., Slavich, L., Walker, R. and Monos, D. S. (2016), Determining performance characteristics of an NGS-based HLA typing method for clinical applications. HLA, 87: 141–152. doi:10.1111/tan.12736
4. Breton Hornblower, Amy Coombs, Richard D Whitaker, Anatoly Kolomeisky, Stephen J Picone, Amit Meller & Mark Akeson. Single-molecule analysis of DNA-protein complexes using nanopores. Nature Methods (2007) 4, 315 - 317
5. Nicholas J Loman, Mick Watson. Successful test launch for nanopore sequencing. Nature Methods (2015) **12**, 303–304
6. Marcel G. J. Tilanus. The power of Oxford Nanopore MinION in human leukocyte antigen immunogenetics. Annals of Blood [Online], 2.6 (2017): doi:10.21037/aob.2017.08.02
7. Robinson J, Halliwell JA, Hayhurst JH, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA Database: allele variant databases. Nucleic Acids Res 2015;43:D423-431
8. SGE Marsh, ED Albert, WF Bodmer, RE Bontrop, B Dupont, HA Erlich, M Fernández-Vina, DE Geraghty, R Holdsworth, CK Hurley, M Lau, KW Lee, B Mach, WR Mayr, M Maiers, CR Müller, P Parham, EW Petersdorf, T Sasazuki, JL Strominger, A Svejgaard, PI Terasaki, JM Tiercy, J Trowsdale: Nomenclature for factors of the HLA system, 2010. Tissue Antigens 2010 75:291-455
9. Kanz C, Aldebert P, Althorpe N, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 2005;33(Database Issue):D29-D33.
10. Mack SJ. A GENE FEATURE ENUMERATION APPROACH FOR DESCRIBING HLA ALLELE POLYMORPHISM. Human immunology. 2015;76(12):975-981. doi:10.1016/j.humimm.2015.09.016.
11. IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970. Eur J Biochem. 1970;15(2):203-208.
12. Judith Reinders, Erik H. Rozemuller, Henny G. Otten, et. al. Identification of HLA-A*0111N: A Synonymous Substitution, Introducing an Alternative Splice Site in Exon 3, Silenced the Expression of an HLA-A Allele. Hum Immunol. 2005;66(8):912-920
13. Surendranath V , Albrecht V , Hayhurst JD , Schöne B , Robinson J , Marsh SGE , Schmidt AH and Lange V . TypeLoader: A fast and efficient automated workflow for the annotation and submission of novel full-length HLA alleles. HLA. 2017;90:25–31.
14. Milius RP, Heuer M, Valiga D, et al. Histoimmunogenetics Markup Language 1.0: Reporting Next Generation Sequencing-based HLA and KIR Genotyping. Hum Immunol. 2015;76(12):963-974.

15. Duygu B, Matern BM,Groeneweg M, Voorter CEM and Tilanus MGJ. Polymorphism at residue 156 of the new HLA-A*02:683 allele suggests immunological relevance. HLA.2017;90:107–109. https://doi.org/10.1111/tan.13059NEW ALLELE ALERTS 109

CHAPTER 3

3

Full-length extension of HLA allele sequences by HLA allele-specific hemizygous Sanger sequencing (SSBT)

C.E.M. Voorter, B.M. Matern, T.H. Tran, A. Fink, B. Vidan-Jeras, S. Montanic, G. Fischer, I. Fae, D. De Santis, R. Whidborne, M. Andreani, M. Testi, M. Groeneweg, M.G.J. Tilanus

Hum Immunol. 2018 Nov;79(11):763-772. doi: 10.1016/j.humimm.2018.08.004

Abstract

Abstract

The gold standard for typing at the allele level of the highly polymorphic Human Leukocyte Antigen (HLA) gene system is sequence based typing. Since sequencing strategies have mainly focused on identification of the peptide binding groove, full-length sequence information is lacking for >90% of the HLA alleles. One of the goals of the 17*th* IHIWS workshop is to establish full-length sequences for as many HLA alleles as possible. In our component "Extension of HLA sequences by full-length HLA allele-specific hemizygous Sanger sequencing" we have used full-length hemizygous Sanger Sequence Based Typing to achieve this goal. We selected samples of which full length sequences were not available in the IPD-IMGT/HLA database. In total we have generated the full-length sequences of 48 HLA-A, 45 -B and 31 -C alleles. For HLA-A extended alleles, 39/48 showed no intron differences compared to the first allele of the corresponding allele group, for HLA-B this was 26/45 and for HLA-C 20/31. Comparing the intron sequences to other alleles of the same allele group revealed that in 5/48 HLA-A, 16/45 HLA-B and 8/31 HLA-C alleles the intron sequence was identical to another allele of the same allele group. In the remaining 10 cases, the sequence either showed polymorphism at a conserved nucleotide or was the result of a gene conversion event. Elucidation of the full-length sequence gives insight in the polymorphic content of the alleles and facilitates the identification of its evolutionary origin.

1. Introduction

The human leukocyte antigen (HLA) loci are located on the short arm of chromosome 6 and belong to the most polymorphic gene system described in the human genome [1]. The HLA molecules play a pivotal role in the immune response, and the huge variability of these molecules has ensured an adequate immune defense against a variety of invaders [2]. Due to the critical role in organ and stem cell transplantation and associations with diseases, HLA gene polymorphism has been studied extensively by molecular typing methods [3]. HLA gene sequencing has been performed since the early 1990s [4], with the focus on the exons which encode the peptide binding groove [5]. This has resulted in identification of thousands of HLA alleles, but generated limited data on full-length gene sequences [6]: the major reason that the IPD-IMGT/HLA database [7] has high ratios of incomplete alleles ranging from 98% for HLA-DRB1 to 84% for HLA-C in the database release 3.27.0 [8].

With increasing knowledge of HLA and new developments in sequencing strategies and approaches, laboratories are moving to the full length amplification and sequencing of HLA genes [3, 9, 10]. However, accurate analysis of the complete gene is impaired by the lack of intron data in the IPD-IMGT/HLA database. Most HLA allele assignment software programs circumvent the problem by either disregarding the introns completely, or by comparing them with the intron sequences of the first allele of the allele group that is fully known. One of the major goals of the 17*th* IHIWS was to perform full-gene typing of the classical class I and class II genes through the application of next generation sequencing (NGS) technologies. Not all HLA laboratories have adopted NGS technologies, but have interesting samples and Sanger sequencing available. Our full-length hemizygous Sanger Sequencing approach results in unambiguous full length typing results by separation of the alleles by group-specific amplification for class I and class II genes [11]. Therefore, we have organized the component "Extension of HLA allele sequences by full-length HLA allele-specific hemizygous Sanger sequencing (SSBT)" to facilitate active participation in the workshop of all laboratories, regardless of the sequencing approach they are using.

The specific goal of this workshop component was to extend the sequences of as many incompletely covered HLA alleles as possible by full-length unambiguous (hemizygous) Sanger Sequencing. The Maastricht SSBT approach [11, 12] enables full-length sequencing and separates the alleles based on available low-resolution typing data. The NGS methods currently in use are based on short-read shotgun approaches, which might result in phasing problems of polymorphic sites/SNPs. Until now, single molecule long read sequencing (SMS) technology has had limitations due to the relatively high error rate at the individual read level, resulting in a high rate of insertion or deletion errors. Both NGS and SMS impair accurate analysis of homopolymer stretches and repetitive sequences

[8]. Due to these shortcomings the Sanger Sequence Based Typing method of separated alleles remains the best reference sequence method for HLA high resolution typing. This SSBT method can be applied to identify and confirm sequence polymorphism in novel alleles, alleles with unknown intron and exon sequences, or sequences generated from NGS/SMS data that are difficult to interpret due to homopolymer stretches or phasing issues. Previous approaches with Sanger sequencing revealed three kinds of ambiguities: (a) genotype ambiguity due to cis-trans polymorphism, (b) allele ambiguity due to polymorphism outside the peptide-binding groove and (c) ambiguities due to incomplete sequences. By using the SSBT Maastricht approach of group-specific amplification and Sanger sequencing both ambiguity types (a) and (b) are resolved by hemizygous and full-length sequencing, respectively, whereas the third ambiguity type (c) is addressed by the workshop efforts to complete the sequences of as many HLA alleles as possible. Furthermore, the group-specific priming minimizes the problem of jumping PCR that could result in in vitro artifactual hybrid sequences, which is a potential problem with long amplicons [13-15].

During the workshop component "Extension of HLA allele sequences by full-length HLA allele-specific hemizygous Sanger sequencing (SSBT)", participants submitted samples based on the selection that the full-length sequence of the submitted HLA alleles was unknown. Initial selection was focused on HLA class I alleles. Participants had the choice of performing the sequencing in their own laboratory with provided SSBT reagents from Maastricht, or sending DNA to the Maastricht laboratory for sequencing. A total of 155 samples were submitted, resulting in identification of the extended full-length sequences of 145 different class I alleles in this workshop component. All sequences were compared with the genomic sequences of the first allele of the same allele group, with other alleles of the same allele group and/or with alleles from other allele groups, depending on the differences found.

2. Material and methods

2.1 DNA isolation and regular HLA typing

Genomic DNA was isolated either manually with QIA-AMP kits following the supplier's protocol (Qiagen, Westburg, Leusden, The Netherlands) or with the salting out method [16], or automated with EZ1 (Qiagen) or Maxwell kits (Promega, Madison, Wisconsin, USA). Concentration and purity of DNA samples were measured at 260 nm and 260/280 nm. The HLA alleles included from samples submitted by the participants were previously typed by the laboratory of origin, based upon sequencing the exons encoding the peptide binding groove minimally.

2.2 SSBT method

The SSBT method was performed as previously described [11, 12]. In brief: a group-specific amplification product of the full-length allele of interest was generated using group-specific primers in the 5′ and 3′ UTRs. Sanger sequencing was performed with generic sequencing primers in both forward and reverse direction by means of cycle sequencing, using the 3730 DNA-analyzer for electrophoresis. Sequencing analysis was performed with the SeqPilot software (JSI medical systems, Kippenheim, Germany), intron analysis was done manually or by using Lasergene (DNA Star, Madison, Wisconsin USA). The full-length sequence was confirmed by sequencing 2 different polymerase chain reaction products for each allele. In cases where two alleles belong to the same allele group, sequencing was performed with Oxford Nanopore MinION $1D^2$ or 2D sequencing [17] to obtain the allele sequence in isolation for correct phasing. In these cases, the MinION sequence was confirmed with heterozygous Sanger sequencing, approving all polymorphic positions.

2.3 NGS typing

For 14 samples the SSBT results performed in this workshop were used to confirm NGS data previously obtained by the Vienna laboratory. This NGS method was based on parallel sequencing of small fragments. DNA was amplified by long range PCR using locus specific primers. The amplification products were purified using ExoSap (Biotechrabbit, Biozym, Berlin, Germany). They served as template for library preparation: fragmentation was achieved enzymatically and adapters were subsequently attached to the fragments (NebNext, New England Biolabs, Ipswich, Massachusetts, USA). After a further purification step using magnetic beads (HighPrep PCR Cleanup System, MAGBIO Genomics, Gaithersburg, Maryland, USA) size selection of 400 bp long fragments was accomplished using E-gel (Invitrogen, Life Technologies, Carlsbad, California, USA). These fragments were re-amplified (NebNext New England Biolabs, Ipswich, Massachusetts, USA); the concentration was adjusted, the fragments were attached to IonSpheres and emulsion PCR was performed (OneTouch, Life Technologies, Carlsbad, California, USA). The products of the emulsion PCR were enriched and loaded onto an Ion Chip (Life Technologies, Carlsbad, California, USA). Finally, the sequence was determined on an IonTorrent Personal Genome Machine (Life Technologies, Carlsbad, California, USA). Data analysis was done using two independent software (HLA TypeStream Life Technologies, Carlsbad, California, USA, NGSengine, GenDx, Utrecht, The Netherlands).

2.4 Submission to EMBL (ENA) and IPD-IMGT/HLA

All extended sequences have been submitted to the EMBL ENA database with the aid of the Saddlebags software, an ENA submission tool developed by Matern *et al* [18]. In short, full length sequences, alongside EMBL credentials and relevant sample data, are converted into an annotated EMBL allele submission, which is uploaded to EMBL ENA, using standard web protocols. Accession numbers were obtained within 48 hours, and

the retrieved annotated sequences were used for manual submission to the IPD-IMGT/HLA database, providing all additional information manually. Both the EMBL accession numbers as well as the obtained IPD-IMGT/HLA submission numbers are indicated in tables 1-4.

2.5 Comparison of genomic sequences

The obtained genomic sequences were compared with the genomic sequence of the first allele of the same allele group with the lowest numbered allele name for which a full length sequence was known e.g. the sequence of HLA-A*01:17 is compared to the genomic sequence of HLA-A*01:01:01:01, whereas the sequence of C*02:29 is compared to the genomic sequence of C*02:02:02:01, since the full-length genomic sequence of C*02:02:01 is not yet known. The only exception is the B*40 group. Within this group two different evolutionary lineages are recognized, identified by numerous nucleotide differences in both exons and introns. In this group comparison was based on exon 1: when exon 1 was identical to B*40:01, the genomic sequence was compared with B*40:01:01 as first allele of the same allele group, if exon 1 was identical to B*40:02, the genomic sequence was compared with B*40:02:01:01, the first allele of the same allele group. Only differences in the intron sequences were considered. When differences were observed, the genomic sequence of the allele was compared to other alleles of the same allele group to determine whether the intron sequences were already known for that allele group. If the sequence was not yet known for the allele group, the sequence was compared to the genomic sequences of all other allele groups to enable identification of previously conserved and variable positions, and possible recombination or gene conversion events between different allele groups.

HLA allele	EMBL accession number	IPD-IMGT/HLA submission number
A*01:03:01:02	LT618820	HWS10029815
A*02:22:01:01	LT618825	HWS10029781
A*03:01:03	LT976502	HWS10051195
A*26:08	LT934331	HWS10029821
A*29:10:01	LT971013	HWS10051163
A*29:95	LT618817	HWS10029811
A*32:01:01:04	LT934313	HWS10029823
A*30:10	LT604089	HWS10029783
B*08:09	LT618798	HWS10029785
B*27:12	LT604095	HWS10029553
B*27:19	LT969621	HWS10051011
B*35:12:01	LT604097	HWS10029789
B*35:43:01	LT934312	HWS10029829
B*39:01:24	LT934333	HWS10029827
B*39:31	LT604098	HWS10029791
B*40:23	LT632312	HWS10029793
B*42:02:01:02	LT934330	HWS10029825
B*57:04:01	LT618823	HWS10029795
C*07:01:01:14Q	LT898183	HWS10029799
C*03:05	LT934335	HWS10029831
C*04:07	LT618809	HWS10029797

Table 1. Confirmatory extended full length sequenced HLA alleles with EMBL accession numbers and IPD-IMGT/HLA submission numbers.

HLA-A allele	compared to	EMBL accession number	IPD-IMGT/HLA submission number	Number of samples
A*01:01:03	A*01:01:01:01	LT604086	HWS10029539	1
A*01:17	A*01:01:01:01	LT618839	HWS10029541	1
A*01:25	A*01:01:01:01	LT618801	HWS10029543	1
A*02:35:01	A*02:01:01:01	LT618840, LT934427, LT969580	HWS10029643, HWS10029877, HWS10050963	3
A*02:67	A*02:01:01:01	LT985838	HWS10051867	1
A*02:113:01N	A*02:01:01:01	LT965076	HWS10051015	1
A*02:140	A*02:01:01:01	LT960366	HWS10050135	1
A*02:161	A*02:01:01:01	LT618802	HWS10029527	1
A*02:241	A*02:01:01:01	LT969582	HWS10050961	1
A*02:242	A*02:01:01:01	LT969601	HWS10050959	1
A*03:05:01	A*03:01:01:01	LT618838	HWS10029525	1
A*03:10	A*03:01:01:01	LT618837	HWS10029523	1
A*03:53	A*03:01:01:01	LT960360	HWS10050137	1
A*03:123:01	A*03:01:01:01	LT969578	HWS10050965	1
A*03:143	A*03:01:01:01	LT960367	HWS10050139	1
A*11:03	A*11:01:01:01	LT618803	HWS10029547	1
A*11:04	A*11:01:01:01	LT618824	HWS10029535	1
A*11:29	A*11:01:01:01	LT618804, LT969581	HWS10029537, HWS10050967	2
A*11:126	A*11:01:01:01	LT935656	HWS10029899	1
A*23:18	A*23:01:01:01	LT960361	HWS10050143	1
A*24:02:25	A*24:02:01:01	LT618805	HWS10029489	1
A*24:02:26	A*24:02:01:01	LT985839	HWS10051869	1
A*24:02:65	A*24:02:01:01	LT960354	HWS10050145	1
A*24:25	A*24:02:01:01	LT969583	HWS10050973	1
A*24:31	A*24:02:01:01	LT969600	HWS10050981	1
A*24:32	A*24:02:01:01	LT934429	HWS10029879	1
A*24:54	A*24:02:01:01	LT604085	HWS10029491	1
A*24:135:01	A*24:02:01:01	LT960356	HWS10050147	1
A*25:11	A*25:01:01:01	LT970871	HWS10051171	1
A*26:09	A*26:01:01:01	LT970916	HWS10051169	1
A*26:26	A*26:01:01:01	LT960364	HWS10050149	1
A*26:27	A*26:01:01:01	LT970911	HWS10051167	1
A*26:47	A*26:01:01:01	LT934428	HWS10029875	1
A*26:52	A*26:01:01:01	LT604087	HWS10029495	1
A*30:11:01	A*30:01:01	LT604090	HWS10029507	1
A*31:12	A*31:01:02:01	LT934430	HWS10029881	1
A*31:13	A*31:01:02:01	LT976488	HWS10051193	1
A*32:11Q	A*32:01:01:01	LT960355	HWS10050151	1
A*32:15	A*32:01:01:01	LT971377	HWS10051177	1

HLA-B allele	compared to	EMBL accession number	IMGT submission number	Number of samples
B*07:10	B*07:02:01:01	LT962901, LT971385	HWS10050491, HWS10051161	2
B*07:15	B*07:02:01:01	LT604084	HWS10029515	1
B*07:20	B*07:02:01:01	LT985837	HWS10051871	1
B*07:22:01	B*07:02:01:01	LT969547	HWS10050983	1
B*07:26	B*07:02:01:01	LT906668	HWS10029645	1
B*07:31	B*07:02:01:01	LT969542	HWS10050985	1
B*07:69	B*07:02:01:01	LT960369	HWS10050153	1
B*07:92	B*07:02:01:01	LT971378	HWS10051175	1
B*07:104	B*07:02:01:01	LT935653	HWS10029901	1
B*07:129	B*07:02:01:01	LT618806, LT618816	HWS10029519, HWS10029517	2
B*07:161N	B*07:02:01:01	LT898199	HWS10029475	1
B*15:01:06	B*15:01:01:01	LT960357	HWS10050157	1
B*15:33	B*15:01:01:01	LT969540	HWS10050987	1
B*18:12:01	B*18:01:01:01	LT969541	HWS10050989	1
B*27:09	B*27:02:01:01	LT934434	HWS10029885	1
B*27:10	B*27:02:01:01	LT604094	HWS10029549	1
B*27:17	B*27:02:01:01	LT604096	HWS10029551	1
B*27:30	B*27:02:01:01	LT962903	HWS10050495	1
B*40:08	B*40:02:01:01	LT934329	HWS10029873	1
B*40:229	B*40:02:01:01	LT934309	HWS10029817	1
B*41:21	B*41:01:01	LT969543	HWS10050993	1
B*44:21	B*44:02:01:01	LT934433	HWS10029887	1
B*44:37:01	B*44:02:01:01	LT965073	HWS10050767	1
B*44:55	B*44:02:01:01	LT969539	HWS10050991	1
B*53:07	B*53:01:01	LT972227	HWS10051191	1
B*55:11	B*55:01:01	LT618841	HWS10029571	1

HLA-C allele	compared to	EMBL accession number	IMGT submission number	Number of samples
C*01:44	C*01:02:01:01	LT934432	HWS10029895	1
C*02:29	C*02:02:02:01	LT934435	HWS10029893	1
C*02:44	C*02:02:02:01	LT604099	HWS10029573	1
C*02:106	C*02:02:02:01	LT960362	HWS10050163	1
C*03:08	C*03:02:01	LT898186	HWS10029529	1
C*03:16	C*03:02:01	LT575485	HWS10029599	1
C*03:35:01	C*03:02:01	LT632317, LT934332	HWS10029581, HWS10029833	2
C*03:44	C*03:02:01	LT934436,	HWS10029897,	2
C*03:74	C*03:02:01	LT962900	HWS10050489	1
C*04:01:22	C*04:01:01:01	LT971249	HWS10051165	1
C*04:41	C*04:01:01:01	LT965072	HWS10050765	1
C*04:108	C*04:01:01:01	LT604100	HWS10029597	1
C*05:01:05	C*05:01:01:01	LT969623	HWS10051017	1
C*05:13	C*05:01:01:01	LT985859	HWS10051881	1
C*07:01:08	C*07:01:01:01	LT970872	HWS10051173	1
C*07:55N	C*07:01:01:01	LT906666	HWS10029647	1
C*12:10:02	C*12:02:02:01	LT960359	HWS10050171	1
C*14:11	C*14:02:01:01	LT969548	HWS10050997	1
C*16:07:02	C*16:01:01:01	LT960353	HWS10050173	1
C*16:74	C*16:01:01:01	LT618815	HWS10029595	1

Table 2 (a-c). New extended full length sequenced HLA alleles with identical intron sequences compared to the first allele of the same allele group for (A) HLA-A alleles, (B) HLA-B alleles and (C) HLA-C alleles.

3. Results

3.1 Confirmation of full-length sequences

In total, 145 alleles were studied during the workshop. The genomic sequence of 124 of them were not known in the IPD-IMGT/HLA database v3.29, and 21 sequences were already submitted in 2017 by other laboratories, but were not yet confirmed. Therefore, we confirmed these alleles by full-length sequencing and submitted them to EMBL and IPD-IMGT/HLA. The confirmed alleles and their EMBL and IPD-IMGT/HLA accession numbers are listed in table 1. The full-length sequences were compared with the known genomic sequence in the database, and no differences were found, with the exception of HLA-B*27:12. The genomic sequence of the database (v 3.29) showed 5 G nucleotides in intron 2 from position 684 to 688. However, we identified in a B*27:12 sequence (cell id 44302 IPD-IMGT/HLA) 6 G nucleotides at these positions, as is the case for all other B*27 genomic sequences known. Furthermore, sequencing B*27:12 from three other unrelated individuals revealed in all cases also 6 G nucleotides at these positions. In the more recent database version 3.30, an additional allele has been assigned, B*27:12:01:02, that has 6 G nucleotides at position 684 in intron 2, whereas the original B*27:12 was renamed to B*27:12:01:01, with 5 G nucleotides. The B*27:12:01:01 was identified by sequencing with NGS (PacBio), whereas B*27:12:01:02 was sequenced with hemizygous Sanger based sequencing. It is tempting to speculate that the 5G variant might be a sequencing artefact, considering the difficulty NGS has with identifying homopolymers correctly [19].

3.2 Extension of sequences not yet known

During the workshop, full-length sequences were determined for a total of 134 samples with 124 different alleles, of which the full length genomic sequence was not yet known (IPD-IMGT/HLA v3.29). All sequences were compared with the genomic sequences of either the first allele of the same allele group with the lowest numbered allele name, with other alleles of the same allele group, and/or with alleles from other allele groups, selected based on the polymorphisms found.

3.2.1 Alleles with no differences in the intron sequences compared to first allele of the same allele group

The majority of the alleles studied showed no differences in their intron sequences with the intron sequences of the first allele of the same allele group. These alleles are listed for HLA-A, -B and -C in tables 2 A-C, respectively, including the EMBL accession number and IMGT submission number. The first allele of the same allele group to which the intron sequences were compared is included. In total 39/48 HLA-A, 26/45 HLA-B and 20/31 HLA-C genomic sequences showed no differences with the first allele of the same allele group in their intron sequences.

Allele	compared to	Differences *a*	no differences with	EMBL accession number	IPD-IMGT/HLA submission number
HLA-A					
A*02:209	A*02:01:01:01	I3 T1356C	A*02:02:01:01	LT962390	HWS10050333
A*24:122	A*24:02:01:01	I3 A1384G	A*24:02:01:05	LT969579	HWS10050969
A*24:314	A*24:02:01:01	I6 C2678T	A*24:02:01:04	LT969584	HWS10050975
A*30:04:02	A*30:01:01	I5 G2269A, I5 T2284A	A*30:04:01	LT970914	HWS10051181
A*68:12	A*68:01:01:01	I3 G1567C	A*68:01:01:02	LT962902	HWS10050493
HLA-B					
B*14:06:01	B*14:01:01:01	I2 T665G	B*14:02:01:01	LT604093	HWS10029577
B*14:31	B*14:01:01:01	I2 T665G	B*14:02:01:01	LT985858	HWS10051873
B*18:07:01	B*18:01:01:01	I3 T1126C, I5 A2170G, 3ut T3014C	B*18:01:01:02	LT898198, LT960351	HWS10029471, HWS10050159
B*18:18	B*18:01:01:01	I3 T1126C, I5 A2170G, 3ut T3014C	B*18:01:01:02	LT618813, LT898180, LT934426	HWS10029555, HWS10029531, HWS10029883
B*18:33	B*18:01:01:01	I3 T1126C, I5 A2170G, 3ut T3014C	B*18:01:01:02	LT934431	HWS10029889
B*18:73	B*18:01:01:01	I3 T1126C, I5 A2170G, 3ut T3014C	B*18:01:01:02	LT985841	HWS10051877
B*35:20:01	B*35:01:01:01	3UT G2932A	B*35:01:01:02	LT972238	HWS10051239
B*35:38	B*35:01:01:01	3'UT C2707T, ins 11 nucl at 2711, G2932A	B*35:03:01:03	LT970917	HWS10051179
B*35:231	B*35:01:01:01	3UT G2932A	B*35:01:01:02	LT985852	HWS10051879
B*35:240	B*35:01:01:01	3UT G2932A	B*35:01:01:02	LT960368	HWS10050161
B*39:15	B*39:01:01:01	I5 T2407C	B*39:01:01:03	LT604103	HWS10029569
B*39:25N	B*39:01:01:01	I5 T2407C	B*39:01:01:03	LT934311	HWS10029819
B*56:04	B*56:01:01:01	I5 G2466C	B*56:01:01:04	LT969616	HWS10051027
B*56:07	B*56:01:01:01	I5 G2466C	B*56:01:01:04	LT632315	HWS10029567
B*57:02:01	B*57:01:01:01	I5 C2246T	B*57:03:01:01	LT618810	HWS10029641
B*57:02:02	B*57:01:01:01	I5 C2246T	B*57:03:01:01	LT618811	HWS10029557

HLA-C					
C*07:02:10	C*07:01:01:01	I1 G127T, I2 C543A, I3 G1468C, I7 A2843G, 3UT G2987A, T2988C, T3005C	C*07:02:01:10	LT618814	HWS10029561
C*07:02:16	C*07:01:01:01	I1 G127T, I2 C543A, I3 G1468C, I7 A2843G, 3UT G2987A, T2988C	C*07:02:01:03	LT965075	HWS10050769
C*07:02:60	C*07:01:01:01	I1 G127T, I2 C543A, I3 G1468C, I7 A2843G, 3UT G2987A, T2988C	C*07:02:01:03	LT960363	HWS10050165
C*07:13	C*07:01:01:01	I1 G127T, I2 C543A, I7 A2843G, 3UT G2987A, T2988C	C*07:02:01:01	LT604102	HWS10029579
C*07:46	C*07:01:01:01	I1 G127T, I2 C543A, I3 G1468C, I7 A2843G, 3UT G2987A, T2988C	C*07:02:01:03	LT985840	HWS10051883
C*07:195	C*07:01:01:01	I1 G127T, I2 C543A, I3 G1468C, I7 A2843G, 3UT G2987A, T2988C	C*07:02:01:03	LT960352	HWS10050167
C*07:294	C*07:01:01:01	I1 G127T, I2 C543A, G710C, I3 G1468C, I7 A2843G, 3UT G2987A, T2988C	C*07:02:01:13	LT969549	HWS10050995
C*08:33:01	C*08:01:01:01	I2 C503A, I3 C1137 T, G1338A	C*08:02:01:01	LT960358	HWS10050169

Table 3: Newly extended full length sequenced HLA alleles with differences in the intron sequences compared to the first allele of the same allele group, but identical to another allele of the same allele group.

a I = intron, 5UT = 5′ untranslated region, 3UT = 3′ untranslated region, number indicates the position of the genomic sequence according to IPD-IMGT/HLA database

Allele	Compared to	Differences *a*	EMBL accession number	IPD-IMGT/HLA submission number
A*01:37	A*01:01:01:01	I3 C1032T	LT618800	HWS10029545
A*36:03	A*36:01	I2 G670C	LT604091	HWS10029533
A*74:03	A*74:01:01	I1 C125T	LT976505	HWS10051241
A*74:06	A*74:01:01	I2 G490C, G504C, C509T	LT604092	HWS10029513
B*35:42:01	B*35:01:01:01	5UT G-201A, 5UT A-90G, 5UT A-88G, I1 G89C, 3UT G2932A	LT618812	HWS10029559
B*40:16	B*40:01:01	I2 A561C, I3 C1465T, I5 G2483T, I6 G2536A, I6 G2620T	LT618827	HWS10029565
B*40:20	B*40:02:01:01	I2 G688-, I2 G707C, I2 G709T	LT632314	HWS10029563
C*04:11	C*04:01:01:01	I2 C514T, G544A, C553G	LT618833	HWS10029575
C*14:05	C*14:02:01:01	I2 A672G	LT934437	HWS10029891
C*15:11	C*15:02:01:01	5UT A-301T, I1 G196T	LT985876	HWS10051885

Table 4: New extended full length sequenced HLA alleles with differences in the intron sequences to all alleles of the same allele group.
a I = intron, 5UT = 5′ untranslated region, 3UT = 3′ untranslated region, number indicates the position of the genomic sequence according to IPD-IMGT/HLA database

The presence of Questionable and Null alleles in this table indicates that the difference causing the aberrant expression is located in the exons that were already previously assigned. There is one allele included with the suffix Q, A*32:11Q, which has three nucleotide differences compared to A*32:01:01:01. The variation at position 563 of the cDNA results in a change of Cys164 into Phe164. This change impairs the normal disulfide bond between Cys164 and Cys101 [20], most likely affecting the expression of the molecule as previously described [21]. Three null alleles are included in table 2. Two of their respective sequences (A*02:113:01N, C*07:55N) contain a nucleotide mutation which immediately introduces a stop codon [22, 23], and in the third sequence (B*07:161N) an insertion of a nucleotide caused a frameshift, resulting in a downstream premature stop codon [7].

3.2.2 Alleles with no differences in the intron sequences compared to other alleles of the same allele group

In 39 cases differences were identified between the intron sequences of the studied allele compared to the first allele of the same allele group. In these cases the sequences were compared to other alleles of the same allele group in the order of the allele number. In table 3 the alleles are listed that showed no intron differences with another allele of the same allele group; including the differences with the first allele of the same allele group and the first ranked allele with identical intron sequences. In total there were 5 HLA-A, 16 HLA-B and 8 HLA-C alleles that belonged to this group of alleles.
The null allele in this table (B*39:25N) has a deletion of two nucleotides in exon 3, resulting in a frameshift and premature stop codon.

3.2.3 Alleles with differences in the intron sequences compared to all other alleles of the same allele group

Table 4 comprises the remaining alleles that did not fit in the previous two groups. We studied these alleles in more detail by comparing their sequences with all genomic sequences of the corresponding HLA locus currently known (IPD-IMGT/HLA database vs 3.30).

The allele A*01:37 showed, apart from the difference in exon 4, also a difference in intron 3 compared to A*01:01:01:01, namely C1032T. The position 1032 was previously thought to be a conserved position for the HLA-A locus, as all other HLA-A alleles known have a C at this position. Furthermore, the difference in exon 4 between A*01:37 and A*01:01:01:01, C755T is a rather striking change, there are only 3 other HLA-A alleles with a T at this position, A*02:131, A*11:172 and A*29:75. The sequence CAGGACAC at positions 748-755 is conserved among the classical class I genes HLA-A, B and C. No known HLA-B or –C alleles has a T at this position 755.

Comparing the intron sequences, the allele A*36:03 showed only one difference with A*36:01 in intron 2 G670C. A*36:01 is the only A*36 allele of which the full-length genomic sequence is known. Comparing this position 670 with other allele groups revealed the presence of a C in A*11, 23, 24, 29, 31, 32, 33 and 74. Since we also observed a difference at the start of exon 3 between A*36:03 and A*36:01, we compared the polymorphic positions at the 3′ end of intron 2 and the 5′ end of exon 3, for all these allele groups as illustrated in table 5A. From this comparison it is clear that the sequence of A*36:03 is identical to A*36:01 from the 5′ end of the allele up to intron 2 position 617 and from position 739 in exon 3 up to the 3′ end. The sequence from position 670 up to and including position 734 is identical to the sequence of the HLA-A*23 and *24 allele group, implying that the allele A*36:03 might originate from an inter-allelic gene conversion event with an A*36 and an A*23 or A*24 allele.

Up to now the genomic full-length sequences of 7 A*74 alleles are known, 4 of them having identical intron sequences (A*74:01:01, 74:01:02, 74:01:03 and 74:01:06), the others having differences in intron 2, 3, 5 and 7 as illustrated in table 5B. The intron sequence of A*74:03 is identical to A*74:01:01 except for position 125 in intron 1. This C was a conserved nucleotide in all HLA-A alleles, but is a T in A*74:03. The A*74:06 has one nucleotide difference with all other alleles, in exon 2, position G422C. When comparing the intron sequences, the allele A*74:06 shows 3 differences in intron 2 with the first allele of the allele group, A*74:01:01, and only one difference in intron 3 with A*74:02:01:02. This suggests two possible evolutionary origins for this allele, it either arose from A*74:02:01:02 with two point mutations (422, 1427), or it arose from a recombination between A*74:01 with A*74:02 together with a single point mutation (422). Whether intermediate alleles

exist is unclear, there remain 27 A*74 alleles of which the genomic sequence is not yet elucidated.

The B*35:42 was identified as a new allele when we started sequencing exons 1-5 for HLA class I alleles, since there were two nucleotide differences between this allele and the B*35:01:01:01 in exon 1 [24]. Sequencing full-length revealed additional differences in the 5′ UTR and in intron 1 as shown in table 5C. Since these differences were all located next to each other without any other polymorphic positions in between, we speculated that the 5′ part of the B*35:42 was derived from another allele group. When comparing this 5′ part with other alleles in the IPD-IMGT/HLA database there were 4 allele groups identified with identical sequences, B*18, 27, 37 and the alleles of the B*40:02 lineage (table 5C). Therefore, it seems plausible that the B*35:42 allele is the result of a gene conversion between a B*35 allele and an allele of either B*18, 27, 37 or 40:02 lineage, with the cross over point located in intron 1 between position 89 and 146.

The B*40:16 allele belongs, according to the exon 1 sequence, to the B*40:01 lineage. Comparing the intron sequences with those of B*40:01:01 revealed 5 differences, as indicated in table 4. Two of these differences are due to a point mutation in B*40:01:01 compared to the other B*40:01 alleles: at positions I5 2483 and I6 2620 the B*40:01:01 is the only B*40 allele with a G, all other B*40 alleles have a T nucleotide at these positions. Concerning the C at position 561 in intron 2, this nucleotide is associated with a deletion of two nucleotides at positions 566 and 567 in all other B*40 alleles known today. The B*40:16 is the only allele that has a C561 together with CT at positions 566 and 567. Interestingly, the T at position 1465 in intron 3 is unique among the HLA-B*40 alleles, all other B*40 alleles have a C at this position. Furthermore, this T1465 is only present in 4 other B alleles, all belonging to the B*15 group. The final difference I6 G2536A concerns a mutation of a previous conserved position, all other B alleles have a G at this position.

The B*40:20 allele belongs to the B*40:02 lineage and has 3 nucleotide differences with B*40:02:01:01, all located adjacent to each other in intron 2 (table 5D). We have therefore compared this part of the sequence with all other B alleles, and found this combination (.688/689, G691, C707, T709) to be unique for the B*15 and B*46 allele groups. Extending the SNP analysis to exon 3 revealed identity between the sequence of B*40:20 and the B*15/46 allele group up to and including position 792, suggesting that B*40:20 is the result of a recombination event between the B*40:02 and the B*15/46 allele group.

Also the C*04:11 showed 3 nucleotide differences with C*04:01:01:01 adjacent to each other in intron 2 (table 5E). This TAG sequence was not present in any of the other C*04 alleles, but was present in the C*01 and C*03 alleles. Comparison of the sequence around this part showed that the 3′ end of exon 2 was also different from C*04:01:01:01, but was

identical to the C*01 allele group. The C*03 allele group showed an additional difference with C*04:11 at position 473. Therefore, we can conclude that the C*04:11 is the result of a recombination event between C*04:01:01:01 and the C*01 allele group.

The C*14:05 allele differs from all other C*14 alleles in intron 2 at position 672, having a G instead of an A. Many of the C allele groups have a G at this position. Comparison of the adjacent SNPs shows that position 744 in exon 3, immediately adjacent to the SNP at 672, also shows a difference with the C*14 alleles (table 5F). An A at this position 744, without any other SNPs between 672 and 744, is also present in C*06 and C*12:03 lineage (which differs from C*12:02 lineage in exon 3). The C*14:05 might be the result of a recombination between C*14 and C*06/12:03, but since this recombination includes only 2 SNP positions it might as well be the result of a double mutation.

Comparing the intron sequences of C*15:11 with those of C*15:02:01:01 revealed two differences, one in the 5′ UTR at position -301 A to T, and one in intron 1 at position 196 G to T. Since there were also differences in exon 2 between C*15:11 and C*15:02:01:01 we have compared the full-length sequence with other C alleles. The T's at position -301 and 196 are also present in C*02, 04 and 14 alleles, but the C*04 and 14 alleles differ in exon 2 from C*15:11. However, the sequence of the C*02 alleles in exon 2 is identical to the C*15:11 allele and differs from the C*15 alleles (table 5G). Therefore, it seems plausible that the C*15:11 allele arose by a recombination event between a C*02 and a C*15 allele.

Table 5A:

	I2 *a*								**E3**				
HLA-A*	580	593	617	670	673	675	681	700	726	734	739	747	756
11	C	C	T	C	G	A	T	C	A	A	A	G	C
29/31/32/33/74	C	C	T	C	A	G	C	T	A	G	A	G/C	T/C
23/24	C	C	A	C	G	G	C	C	C	G	T	G	T
36:03	T	T	T	C	G	G	C	C	C	G	A	G	C
36:01	T	T	T	G	G	G	C	C	A	A	A	G	C

Table 5B:

	E1	**E2**	**I2**			**I3**	**I5**	**I7**				
HLA-A*	67	422	490	504	509	1427	2394	2842	2850	2862	2863	2868
74:01:01-74:01:03, A 74:01:06	G	G	G	C	C	C	A	T	C	A	C	
74:01:04	A	G	G	G	C	C	A	A	T	C	A	C
74:02:01:01	T	G	C	C	T	C	C	G	C	T	G	T
74:02:01:02	T	G	C	C	T	T	C	A	T	C	A	C
74:06	T	C	C	C	T	C	C	A	T	C	A	C

Table 5C:

	5′UT			E1		I1			E2	
HLA-B*	-201	-90	-88	25	72	89	146	148	225	231
35	G	A	A	G	C	G	T	A	T	G
35:42:01	A	G	G	C	T	C	T	A	T	G
18/27/37/40:02lineage	A	G	G	C	T	C	C	G	C	T

Table 5D:

	I2					E3							
HLA-B*	561	566	567	588	688/689	691	707	709	736	785	792	836	850
40:01 lineage	A	C	T	G	.	T	C	G	G	A	A	C	C
40:02 lineage	C	.	.	G	G	G	G	G	C	A	A	C	C
40:20	C	.	.	G	.	G	C	T	G	G	C	C	C
15/46	C	C	T	A	.	G	C	T	G	G	C	A	G

Table 5E:

	E2						I2					E3			
HLA-C*	256	272	331	348	419	432	442	471	514	544	553	694	704	737	744
01	A	T	G	C	A	G	C	C	T	A	G	G	A	T	G
04:11	G	G	A	A	A	G	C	C	T	A	G	C	A	A	T
04:01	G	G	A	A	G	A	A	A	C	G	C	C	A	A	T

Table 5F:

	I2	E3		
HLA-C*	553	672	744	835
06/12:03 lineage	G	G	A	C
14:05	C	G	A	T
14	C	A	T	T

Table 5G:

	5′UT	I1	E2		I2	
HLA-C*	-301	196	248	400	552	694
02	T	T	A	G	G	.
15:11	T	T	A	G	C	G
15	A	G	G	C	C	G

Table 5: Comparison of part of the sequence between alleles/allele groups to show the gene conversion event (reflected by the different grey shades) that resulted in the allele (A) HLA-A*36:03, (B) HLA-A*74:06, (C) HLA-B*35:42:01, (D) HLA-B*40:20, (E) HLA-C*04:11, (F) HLA-C*14:05 and (G) HLA-C*15:11.

a I = intron, E = exon, 5UT = 5′ untranslated region, number indicates the position of the genomic sequence according to IPD-IMGT/HLA database

3.3 Comparison of full-length sequences obtained by SSBT and by NGS

For 14 alleles full length sequences were obtained with both SSBT and IonTorrent NGS. Comparing the results of the full length sequences revealed no differences between the different sequencing approaches. Eleven of the alleles are present in table 2 (A*02:35:01, A*11:126, A*24:32, A*26:47, A*31:12, B*07:104, B*27:09, B*44:21, C*01:44, C*02:29 and C*03:44), two of them in table 3 (B*18:18, B*18:33) and one in table 4 (C*14:05). For the latter one, both SSBT and NGS identified a G at position 672, that is not present in any of the other C*14 alleles.

4. Discussion

In this workshop component "Extension of HLA allele sequences by full-length HLA allele-specific hemizygous Sanger sequencing (SSBT)", the full-length sequences of 145 different class I alleles were obtained and submitted to the EMBL and IPD-IMGT/HLA database. The strategy to select those alleles that were not yet fully known in the database (IPD-IMGT/HLA vs3.29 July 2017), enabled us to focus resources, efforts and skills to identify the unknown parts of the sequence in the IPD-IMGT/HLA database. During the process, the full-length sequences of 21 of these alleles have been already submitted by other laboratories, and therefore our submissions were considered as confirmatory sequences. For the other 124 alleles, we provided the newly identified full-length sequences, including 48 HLA-A alleles, 45 HLA-B alleles and 31 HLA-C alleles.

Comparing tables 2 (intron sequences identical to the first allele in the group) and 3 (intron sequences not identical to the first allele in the group) we observed that for HLA-B, some allele groups are only present in table 2 and not in table 3 and vice versa. HLA-B*07 is only present in table 2, and is completely absent in table 3, whereas B*35 is only present in table 3 and is lacking in table 2. This may imply that for the allele groups that are only present in table 2, there is little variation in the intron sequences among that group. This was indeed the case for the allele groups B*07, 27, 44, 53 and 55. The B*15 allele group showed two different general types, based on differences in the sequences of 5′ UTR, I1 and I2; the other introns are comparably similar to each other. The B*40 allele group was previously reported to consist of two separate lineages, that show differences throughout the whole genomic sequence, and are correlated with the serological groups B60 and B61 [25]. The B*41 group shows a difference at the end of intron 2, with 7 alleles having a T709 and 8 alleles G709. This position is correlating with 6 polymorphic nucleotides at the 5′ end of exon 3, indeed pointing to two different lineages within this allele group. The phylogenetic trees of the genomic sequences of HLA-B*40 and B*41, illustrating the presence of the two different lineages, are depicted in figure 1 A and B, respectively.

Interestingly, the alleles outside of the two clusters in the HLA*B*40 tree were previously assigned serologically as B48 like (B*40:10 and B*40:12) or as B21 (B*40:26) [26].

For B*55, we observed a separation into two groups at the 3′ UTR end of the sequence, based upon the sequence between position 2708 and 2718, where there are either 8 or 9 G nucleotides present. The same was observed in the B*54 and B*56 allele group, although the number of alleles sequenced is much lower than for B*55. There are 19 alleles with 9 G nucleotides, 2 B*54, 12 B*55 and 5 B*56, whereas there are only 7 alleles with 8 G nucleotides, 1 B*54, 4 B*55 and 2 B*56. The latter ones were all sequenced in 2015 or later, but unfortunately the method of sequencing, NGS or Sanger, is not mentioned. Since we had the experience with B*27:12 (described in the results), we sequenced in addition to B*56:04 and B*56:07 (table 3) 4 other individuals with 3 different B*56 alleles that we had available in our laboratory with the SSBT Sanger method and all of them had 9 G nucleotides. Whether the nucleotide difference is due to sequencing with NGS needs further investigation.

For the majority of the B alleles in table 3, there is only one nucleotide difference with the first allele of the allele group. Within three of the allele groups (B*35, B*39 and B*56), there are only a few alleles identified with the intron sequences identical to the first allele or even none, e.g. B*39:01:01:01 is the only B*39 allele with a T2407, all other B*39 alleles

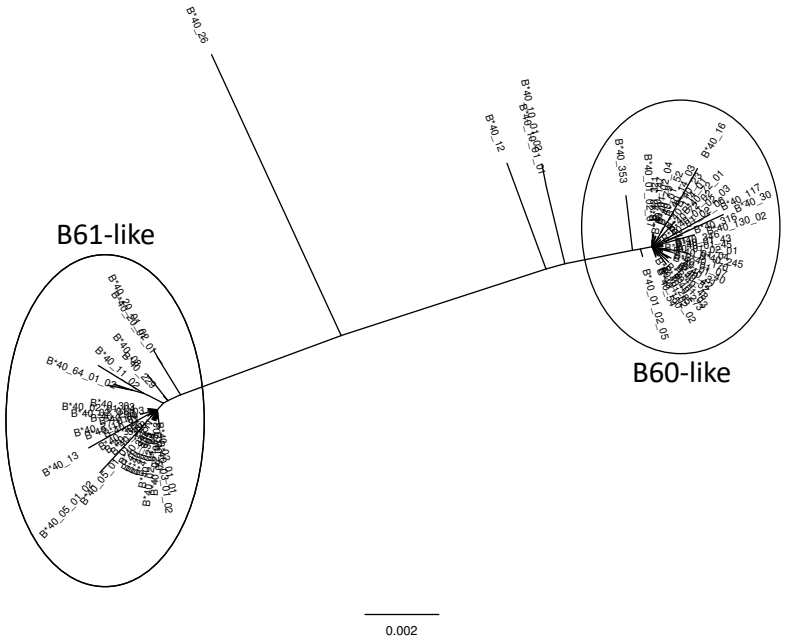


Figure 1A

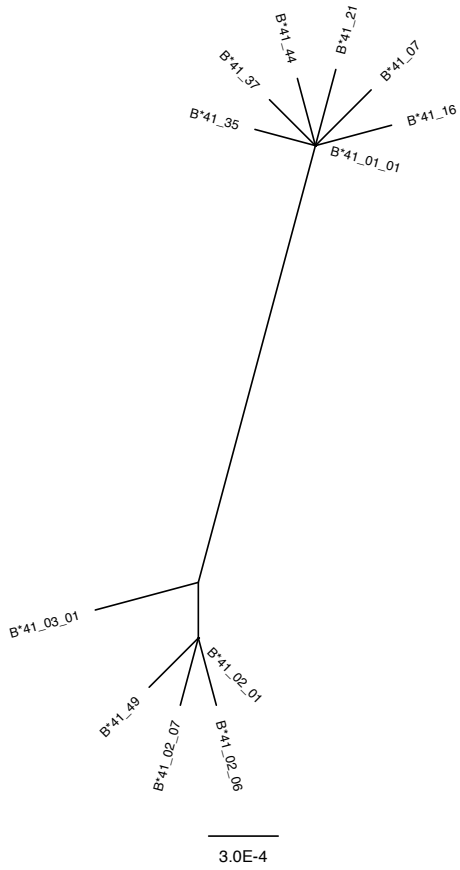


Figure 1B

Figure 1. Phylogenetic tree of HLA* B* 40 (A) and B* 41 (B).

A multiple sequence alignment was performed on full length sequences available in the IPD-IMGT/HLA-database (vs 3.32.0, [7]) for HLA-B* 40 (A) and B*41 (B) alleles. Only those alleles were included that contain continuous sequence data including at least part of the 5′ and 3′ UTR. Trees were constructed using the Clustal Omega 1.2.1 algorithm [36]. In Fig. 1A the two clouds B60-like and B61-like include the following alleles:

B60-like (ordered from top to bottom): B*40:353, B*40:01:02:07, B*40:321, B*40:147, B*40:01:02:04, B*40:79, B*40:01:52, B*40:14:03, B*40:16, B*40:23, B*40:114:01, B*40:72:01, B*40:72:02, B*40:01:02:03, B*40:01:02:06, B*40:117, B*40:30, B*40:316, B*40:130:02, B*40:346, B*40:01:43, B*40:01:45, B*40:01:02:01, B*40:01:04, B*40:245, B*40:170, B*40:01:07, B*40:01:40, B*40:31, B*40:42, B*40:52, B*40:48, B*40:01:03, B*40:273, B*40:285, B*40:01:02:02, B*40:150, B*40:54, B*40:01:02:05.

B61-like (ordered from top to bottom): B*40:20:01:02, B*40:20:01:01, B*40:08, B*40:229, B*40:11:02, B*40:06:04:02, B*40:06:04:01, B*40:06:01:02, B*40:06:01:01, B*40:64:01:02, B*40:64:01:01, B*40:04, B*40:303, B*40:02:01:04, B*40:02:01:03, B*40:296, B*40:90, B*40:27:01, B*40:11:01, B*40:37, B*40:13, B*40:302, B*40:356, B*40:309, B*40:05:01:02, B*40:05:01:01, B*40:345N, B*40:347, B*40:334, B*40:305, B*40:304, B*40:02:24, B*40:02:01:05, B*40:02:01:02, B*40:50, B*40:03:01:02, B*40:02:01:01.

have a C at this position. In the other two cases (B*14, 57) as well as in the case of B*18, there are more alleles with sequences identical to the first one of this group (see column 'compared to'), but also more alleles with sequences identical to the one in the table (see column 'no differences with'), implicating a lineage evolutionary origin. For B*18 it was found that the positions mentioned (I3 1126, I5 2170 and 3UT 3014) were associated with each other in 44/48 alleles, resulting in TAT sequence in 11 alleles and CGC in 33 alleles, although the last nucleotide at position 3014 was not always known.

There are many different reasons for elucidating the full-length sequence of HLA alleles. Although the sequence of the introns is non-coding, there might be important information located in these regions. They may uncover information about the evolutionary origin, as indicated above. Especially in those cases where limited differences are present in the exon sequences, the intron sequences may indicate possible recombination events, and can narrow the location of cross-over positions. In this respect it is worth mentioning the differences in natural selection pressure on introns and exons of the MHC system due to functionality. The non-coding regions may identify polymorphism affecting splicing that subsequently could influence expression levels and/or structure of the molecule [27]. Given that 90% of causal autoimmune disease variants are located within non-coding regions of the genome [28], intronic differences may harbor genomic elements which play a functional role in disease pathogenesis. Recently, it was identified that intron 4 of HLA-B encodes a microRNA (has-miR-6891) [29]. This microRNA has been shown to have an impact on the expression of nearly 200 transcripts in vitro, with a direct impact on metabolic pathways, including immune response networks [30]. Strikingly, we observed differences in all introns and 5′ and 3′ UTR regions, except in intron 4 for HLA-A, -B and -C (see tables 3 and 4). However, analyzing the IPD-IMGT/HLA database revealed the presence of polymorphic positions in the intron 4 sequences of all three classical HLA genes, but they are rather conserved within the allele groups.

With resolving the full-length sequence of HLA alleles, it seems possible to elucidate the evolutionary origin more easily than with exon sequencing alone. In table 4 the HLA alleles are listed that show intron sequences different from the other alleles of the same allele group, except A*74:06, which seems to be the result of either a double point mutation or a gene conversion of two different alleles from the A*74 group. Three of the alleles (A*01:37, A*74:03, B*40:16) show a change of a conserved nucleotide, most probably due to a single point mutation. In the remaining 6 cases the allele was found to be the combination of two different sequences, one derived from the same allele group, the other from another allele group, implicating that the allele was the result of an intragenic inter-allelic gene conversion event. In four of them (A*36:03, B*40:20, C*04:11, C*14:05) only a segmental gene conversion occurred, with the size of the segments ranging from 65 to 345 nucleotides. These segments might represent some of the DNA sequence cassettes

that have been reported to be exchanged between different HLA alleles to generate new variants for adaptation to the exposure of new and evolving pathogens [31]. This is supported by the fact that all segments contain at least part of an exon, encoding one or more different amino acids. In two of them (B*35:42:01, C*15:11) the 5′ end of the gene is completely exchanged, one including exon 2, the other up to and including part of intron 1. In these latter two cases it is unclear whether part of the adjacent sequence has been exchanged as well and to what extent. The B*35:42 was associated with a C*04, which is a common association in the Caucasian population, the sequence 5′ upstream of B*35:42 was not investigated. The C*15:11 was present with a B*27, which is a very uncommon association. However, C*02 with B*27 is a very common association, implying that the C*15:11 might originate from a C*02 in which a large part of the gene (from intron 2 to at least the end of the gene) is changed by a gene conversion event.

The method of group-specific Sanger sequencing for HLA class I was developed a few years ago [11, 12], selecting group-specific primers in the 5′ and 3′ UT regions according to the sequences known at that time. Recently, the sequences on both sides of the gene have been extended and for some alleles the 3′ end sequence was identified up to 1100 bp after the stop codon. The question that arises here is what part does still belong to the gene to enable correct allele assignment. For HLA-C it is known that there is a SNP variation 35 kb upstream of the gene, that might affect the expression level of the cell surface encoded molecule [32] and recently also polymorphism at the 5′ end of the gene was reported to regulate NK cell-specific HLA-C expression [33]. Also for class II there is discussion to what extend the sequence is part of the gene and should be included to define allelic resolution. It remains to be determined whether individual STR variation should be included as HLA allele sequence. In fact, the mutation rate of STR sequences, especially dinucleotide repeats, is several orders of magnitude higher than the mutation rate of unique DNA sequences in the genome [34]. Therefore, including STR variation as HLA variation is questionable and will lead to an exponential increase in the number of alleles. Furthermore, both Sanger sequencing as well as NGS on all platforms encounter problems to analyze STRs in a reliable way due to strand-slippage replication [35], with the consideration that there might already be some errors in the IPD-IMGT/HLA database in these STR rich regions. With the single molecule sequencing that enables long reads, the fast identification of many full-length allele sequences is envisaged for the near future. Therefore, it is now time for the HLA community to discuss and define which parts of the genome belong to a specific gene and must be identified to acquire allelic resolution typing.

Since this workshop component has been successful in resolving full-length sequences of HLA alleles with unknown parts, we will continue this component in the upcoming 18*th* International Histocompatibility Workshop and extend it to class II alleles. Those laboratories that have class I and/or class II alleles available of which the full-length

sequences are not known, are welcome to participate by submitting these alleles. Further information will soon be available on the IHIWS website or can already be obtained by contacting the corresponding author.

Acknowledgements

The authors like to thank Christel Meertens for technical and Diana van Bakel for secretarial assistance.

References

1. Leffler EM, Gao Z, Pfeifer S, Segurel L, Auton A, Venn O, *et al.*: Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 2013;339:1578-82.
2. Parham P, Ohta T: Population biology of antigen presentation by MHC class I molecules. Science 1996;272:67-74.
3. Erlich H: HLA DNA typing: past, present, and future. Tissue Antigens 2012;80:1-11.
4. Santamaria P, Lindstrom AL, Boyce-Jacino MT, Myster SH, Barbosa JJ, Faras AJ, *et al.*: HLA class I sequence-based typing. Hum Immunol 1993;37:39-50.
5. Lazaro A, Tu B, Yang R, Xiao Y, Kariyawasam K, Ng J, *et al.*: Human leukocyte antigen (HLA) typing by DNA sequencing. Methods Mol Biol. 2013;1034:161-95.
6. Mack SJ: A gene feature enumeration approach for describing HLA allele polymorphism. Human Immunology 2015;76:975-81.
7. Robinson J, Halliwell JA, Hayhurst JH, Flicek P, Parham P, Marsh SGE: The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 2015;43:D423-31.
8. Albrecht V, Zweiniger C, Surendranath V, Lang K, Schöfl G, Dahl A, *et al.*: Dual redundant sequencing strategy: Full-length gene characterisation of 1056 novel and confirmatory HLA alleles. HLA 2017;90:79-87.
9. Hosomichi K, Shiina T, Tajima A, Inoue I: The impact of next-generation sequencing technologies on HLA research. J Hum Genet 2015;60:665-73.
10. Goodwin S, McPherson JD, McCombie WR: Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17:333-51.
11. Voorter CE, Palusci F, Tilanus MG: Sequence-based typing of HLA: an improved group-specific full-length gene sequencing approach. In Beksac M,(ed) Methods Mol Biol, Humana Press, 2014, vol 1109. p 101-14.
12. Voorter CE, Groeneweg M, Groeneveld L, Tilanus MG: Uncommon HLA alleles identified by hemizygous ultra-high Sanger sequencing: haplotype associations and reconsideration of their assignment in the Common and Well-Documented catalogue. Hum Immunol 2016;77:184-90.
13. Kebschull M, Zador AM: Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res 2015;43:el43.
14. Pääbo S, Irwin DM, Wilson AC: DNA damage promotes jumping between templates during enzymatic amplification. J Biol Chem 1990;265:4718-21.
15. Holcomb CL, Rastrou M, Williams TC, Goodridge D, Lazaro AM, Tilanus M, *et al.*: Next-generation sequencing can reveal in vitro-generated PCR crossover products: some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. Tissue Antigens 2014;83:32-40.
16. Miller SA, Dykes DD, Polesky HF: A simple salting out procedure for extracting DNA from human nucleated cells. Nucl Acids Res 1988;16:1215.
17. Tilanus MGJ: The power of Oxford Nanopore MinION in human leukocyte antigen immunogenetics. Annals of Blood 2017;

18. Matern BM, Groeneweg M, Voorter CEM, Tilanus MGJ: Saddlebags: A software interface for submitting full-length HLA allele sequences to the EMBL-ENA nucleotide database. Hla 2018;91:29-35.
19. Shin S, Park J: Characterization of sequence-specific errors in various next-generation sequencing systems. Mol Biosyst 2016;12:914-22.
20. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC: Structure of the human class I histocompatibility antigen, HLA-A2. Nature 1987;329:506-12.
21. Tang TF, Hou L, Tu B, Hwang WY, Yeoh AE, Ng J, *et al.*: Identification of nine new HLA class I alleles in volunteers from the Singapore stem cell donor registries. Tissue Antigens 2006;68:518-20.
22. Hammond L, Limmer M, Dyer CR, Rantes CH, Dunn PP: A new HLA-A*02 null allele, HLA-A*9213N. Tissue Antigens 2008;72:176-7.
23. Badrinath S, Blasczyk R, Bade-Doeding C: Non-expression of HLA-C*07:55N is caused by a premature stop codon in exon 3. Tissue Antigens 2011;79:139.
24. Swelsen WTN, Voorter CEM, Berg van den-Loonen EM: Sequence analysis of exons 1, 2, 3, 4 and 5 of the HLA-B5/35 cross-reacting group. Tissue Antigens 2002;60:224-34.
25. Dawkins RL, Houliston JW: Joint report: BW60. In Terasaki PI,(ed) Histocompatibility Testing,Los Angeles, University of California, 1980. p 454-7.
26. Holdsworth R, Hurley CK, Marsh SGE, Lau M, Noreen HJ, Kempenich JH, *et al.*: The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. Tissue Antigens 2009;73:95-170.
27. Voorter CE, Gerritsen KE, Groeneweg M, Wieten L, Tilanus MG: The role of gene polymorphism in HLA class I splicing. Int J Immunogenet 2016;43:65-78.
28. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, *et al.*: Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2015;518:337-43.
29. Ladewig E, Okamura K, Flynt AS, Westholm JO, Lai EC: Discovery of hundreds of mirtrons in mouse and human small RNA data. Genome Res 2018;22:1634-45.
30. Chitnis N, Clark PM, Kamoun M, Stolle C, Brad Johnson F, Monos DS: An Expanded Role for HLA Genes: HLA-B Encodes a microRNA that Regulates IgA and Other Immune Response Transcripts. Front Immunol 2017;8:583.
31. Klitz W, Hedrick P, Louis EJ: New reservoirs of HLA alleles: pools of rare variants enhance immune defense. Trends Genet 2012;28:480-6.
32. Thomas R, Apps R, Qi Y, Gao X, Male V, O'HUigin C, *et al.*: HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. Nat Genet 2009;41:1290-4.
33. Li H, Ivarsson MA, Walker-Sperling VE, Subleski J, Johnson JK, Wright PW, *et al.*: Identification of an elaborate NK-specific system regulating HLA-C expression. PLoS Genet 2018;14:e1007163.
34. Fan H, Chu J-Y: A Brief Review of Short Tandem Repeat Mutation. Geno Prot Bioinfo 2007;5:7-14.
35. Hosseinzadeh-Colagar A, Haghighatnia MJ, Amiri Z, Mohadjerani M, Tafrihi M: Microsatellite (SSR) amplification by PCR usually led to polymorphic bands: Evidence which shows replication slippage occurs in extend or nascent DNA strands. Mol Biol Res Commun 2016;5:167-74.

36. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Weizhong L, *et al.*: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:1-6.

CHAPTER 4

Long-read nanopore sequencing validated for HLA typing in routine diagnostics

B.M. Matern[1], T.I. Olieslagers[1], M. Groeneweg, Duygu B, L. Wieten, M.G.J. Tilanus, C.E.M. Voorter

Transplantation Immunology, Tissue Typing Laboratory, Maastricht University Medical Center, Maastricht, The Netherlands

[1] Both authors contributed equally to this paper

Manuscript Submitted (Journal of Molecular Diagnostics)

Abstract

Matching of the HLA gene polymorphisms by high-resolution DNA sequence analysis is the gold standard for determining compatibility between patient and donor for hematopoietic stem cell transplantation. Single molecule sequencing (PacBio or MinION) is a newest (third) generation sequencing approach. MinION is a nanopore sequencing platform, which provides long targeted DNA sequences. The long reads provide unambiguous phasing, but the initial high error profile prevented its use in high-impact applications such as Human Leukocyte Antigen (HLA) typing for HLA matching of donor and recipient in the transplantation setting. Ongoing developments on the instrumentation and basecalling software have improved the per-base accuracy of $1D^2$ nanopore reads tremendously. In the current study, we used 2 validation panels of samples covering 70 of the 71 known HLA class I allele groups to compare 3*rd* field sequences obtained by MinION with Sanger sequence-based typing (SSBT) showing a 100% concordance between both data sets. In addition, the first validation panel was used to set the acceptance criteria for the use of MinION in a routine setting. The acceptance criteria were subsequently confirmed with the 2*nd* validation panel.

In summary, the present study describes the validation and implementation of the nanopore sequencing HLA class I typing method and illustrates that nanopore sequencing technology has advanced to a point where it can be used in routine diagnostics with high accuracy.

Acronyms / Definitions

HLA: Human Leukocyte Antigen
MHC: Major Histocompatibility Complex
SCT: Stem Cell Transplantation
SOT: Solid Organ Transplantation
SBT: Sequence-based Typing
SSBT: Sanger Sequence-based Typing
NGS: Next-Generation Sequencing
TGS: Third-Generation Sequencing
STR: Short Tandem Repeat
UTR: Untranslated Region
SNP: Single Nucleotide Polymorphism

Introduction

Human Leukocyte Antigen (HLA) is the human Major Histocompatibility Complex (MHC), a group of genes comprising the most polymorphic loci in the human genome.[1] The HLA genes are encoded within the short arm of the human chromosome 6, and they are grouped by both function and morphology into two general classes, HLA classes I and II. The hyperpolymorphism of HLA is demonstrated in the IPD-IMGT/HLA database,[2] which currently lists 18,691 class I and 7,065 class II HLA alleles (release 3.38.0). The nucleotide polymorphism is reflected in the protein polymorphism, which allows the HLA class I or class II molecules to present a wide variety of intra- and extracellular antigens, respectively.

HLA polymorphism enables the immune system to respond to a large variety of pathogens and diseases, but creates challenges in performing solid-organ (SOT) and stem-cell transplantation (SCT) requiring HLA typing. Both SOT and SCT involve the introduction of non-self tissue to the body, and have the inherent risk of adverse immune response. In SOT, mismatched donor HLA can induce the production of donor-specific anti-HLA antibodies, which can bind to the HLA on the transplanted tissue,[3] triggering antibody-mediated organ rejection. Although matching of patient and donor HLA alleles may not be possible due to low organ availability, high-resolution typing of HLA alleles identifies the amino acid sequence and consequent epitope structure of the HLA molecule, providing insight in the targets of the anti-HLA antibodies.[4, 5] In SCT, T-cells in the donor allograft may recognize the HLA-peptide complexes expressed on recipient tissue as non-self, triggering global immune activation leading to graft-versus-host disease. High-resolution HLA typing for SCT is especially critical, since allele mismatches are linked with sometimes fatal adverse side-effects.[6, 7]

High-resolution HLA typing has further applications in the area of drug hypersensitivity,[8] like abacavir and carbamazepine. HLA-B*57:01 has been correlated with hypersensitivity to abacavir,[9] a drug commonly used to treat HIV, whereas HLA-B*15:02 has been correlated with hypersensitivity to carbamazepine,[10] a drug used in epilepsy treatment. Typing for the HLA-B alleles informs the caregiver's choice of treatment methods and prevents adverse drug interactions. Furthermore, the HLA region is correlated with the highest numbers of human diseases in the human genome,[11] and high-resolution sequencing allows refinement of our understanding of the function of HLA and its relationship to the causality of diseases.

In recent decades, increased understanding of the HLA genetics, advances in DNA sequencing technology, and a general lack of allele-specific antisera have led to a shift in HLA typing methodology. Serological methods have often been replaced or supplemented by DNA sequence-based methods, allowing analysis of HLA at the nucleotide level.[12] Sanger Sequencing was developed in 1977,[13] and Sanger Sequence Based Typing (SSBT)

has been the gold standard for HLA allele assignment for many years. However, with the increase in available HLA allele sequences the heterozygous Sanger sequencing approach became more and more cumbersome, due to the increase in ambiguous typing results, which needed additional sequencing to resolve. This problem could be circumvented by group-specific full length amplification and separate sequencing of the alleles,[14] albeit this requires a preceding low resolution typing to determine the allele groups. Technological advances led to Next Generation Sequencing (NGS), intended to be faster and cheaper than Sanger-based sequencing and with the huge advantage of separate allele sequencing. These technologies, including reversible terminator (Illumina),[15] and semiconductor (Ion Torrent)[16] methods are commonly used for HLA typing,[17] but have the disadvantage of short sequences, impairing with correct alignment and phasing of the alleles. Third Generation Sequencing (TGS) is the term used to describe a new era of sequencing technologies that are focused on the analysis of single molecules, i.e. long stretches of DNA without the need to fragment the DNA into smaller pieces as is the case for NGS based techniques. TGS includes single polymerase molecule (Pacific Biosystems)[18] and nanopore-based sequencing (Oxford Nanopore)[19] technologies.

The MinION[20] is a portable nanopore sequencing platform that generates ultra-long reads and requires little initial equipment investment. MinION uses a flowcell which contains a membrane with a grid of embedded nanopores, each of which is capable of binding to a DNA molecule. An electrical potential difference between both sides of the membrane is applied, generating a current across the membrane. Single-stranded DNA is passing through the pore with help from an accompanying motor protein.[21] As nucleotides pass through the pore, disruptions in the current and resulting electrical signal are measured by the MinION integrated circuits. This signal is characteristic of the bases that are present within the pore, due to the varying size and morphology of the nucleotides. Software tools can translate the electrical signal back to the original nucleotide sequence, in a process known as "basecalling".[22, 23] The MinION is capable of natively sequencing a piece of single-stranded DNA, in a process known as 1D sequencing. A quality improvement is provided by MinION $1D^2$ protocols, where the sample preparation results in double-stranded DNA with adapter proteins attached to both ends, allowing both strands of the DNA to be individually sequenced. The MinION basecallers pair two complementary single-stranded reads in-silico, resulting in a single, higher-accuracy read. The quality of the basecalling of a MinION read is represented by the per-base Phred quality scores,[24] which were averaged over the length of each $1D^2$ read to calculate the mean. The mean reported Phred score over all the reads was found to be 18.5, which corresponds to 98.6% read accuracy. For the hyperpolymorphic HLA genes, in which each nucleotide difference can actually account for another allele, a high level of accuracy is essential to obtain reliable results. With reaching this high level of accuracy, typing of the hyperpolymorphic HLA genes by this MinION approach came into the picture.

We and others earlier described the potential of using the MinION as sequencing platform for the analysis of HLA,[25-32] but up till now it has not been used in routine diagnostics. In the current paper we describe the complete validation process of full length HLA class I single molecule sequencing and typing method using the Oxford Nanopore MinION and the implementation in the routine diagnostic setting.

Material and Methods

Samples and Validation process

In order to validate the MinION approach for reliable HLA typing, quality standards and validation metrics needed to be created. To this end, a panel of samples with known HLA high resolution typing was sequenced and typed with the standard 1D^2 MinION protocol. The initial panel consisted of 33 samples, which cover 70 of the 71 known class I allele groups, excluding HLA-B*83, which was not available in our laboratory (see table 1). This panel was subjected to the MinION 1D^2 approach, with a focus on comparing results with our Sanger sequencing results (see supplementary table 1), and determining read quality and coverage statistics. For each sample in the initial validation, read coverage was measured on each MinION run, for each demultiplexed MinION barcode, for each HLA locus within a barcode, and for both HLA alleles at a locus.

Acceptance criteria were defined in the initial validation panel, and the defined criteria were verified by sequencing and HLA-typing a second panel of 67 samples (402 alleles) from our laboratory in a secondary validation phase parallel to Sanger sequencing (supplementary table 2). The samples were sequenced and analyzed using the combined analysis approach, with optimizations based on the initial validation. The samples were also typed using full-length allele-specific Sanger SBT (SSBT) approach.[14] Typing results from MinION were compared with the SSBT results. Accuracy of HLA allele assignment of these samples was considered based on criteria defined in the initial validation.

MinION Sequencing

The MinION sequencing and analysis method is outlined in Figure 1. The procedure starts with the PCR amplification of 300 ng DNA, purified and isolated from peripheral blood samples and according to the descriptions outlined in Voorter et al.[14] The full-length gene-specific amplification primers are located in the 5′ and 3′ UTRs of the HLA-A, -B and –C genes (Table 2). In order to be able to sequence several samples simultaneously, specific Oxford Nanopore tag and barcode sequences were added to the primers. The tag sequence coordinates the base calling and demultiplexing software, and the barcode sequence is used to sort multiplexed samples. After amplification, presence of PCR products was confirmed by agarose gel electrophoresis, and PCR products were

purified using CleanPCR magnetic beads (GC Biotech, Waddinxveen, the Netherlands). In addition, by using a bead versus DNA ratio of 1:1 during the purification, primer dimers were removed simultaneously. Subsequently, up to nine samples (27 PCR products) were equimolar pooled, with equal distribution between loci, to a total quantity of 1300 ng DNA.

Gene	Direction	Sequence, 5′-3′	IMGT/HLA gDNA position
HLA-A	Forward	GGATACTCACGACGCGGAC	-137 - -119
HLA-A	Reverse	GGGAGCACAGGTCAGCGTGGGAAG	3075 - 3098
HLA-B	Forward	GGCAGACAGTGTGACAAAGAGGC	-420 - -398
HLA-B	Reverse	CTGGGGAGGAAACACAGGTCAGCATGGGAAC	3040 - 3070
HLA-C	Forward	TCAGGCACACAGTGTGACAAAGAT	-327 - -304
HLA-C	Reverse	TCGGGGAGGGAACACAGGTCAGTGTGGGGAC	3067 -3098

Table 2. Amplification Primers. This table contains the gene-specific amplification primers. Only the sequence that complements the HLA UTR sequence is shown, the amplification primers also include a section of sequence containing MinION Tag sequences, as well as DNA barcodes, as provided by Oxford Nanopore.

Pooled samples were further prepared for MinION sequencing using the 1D^2 sequencing kit (SQK-LSK308, Oxford Nanopore) and following manufacturer's protocol with a few minor adjustments. In short, the amplicon strands are end-repaired and dA-tailed using the NEBNext End Repair/dA tailing module (NEBnext). These end-repaired amplicons were purified using Clean PCR beads and ligated with 1D^2 adapters, which allows the nanopore to capture the complement strand immediately after the template. After another purification step, sequencing adapters were ligated onto the amplicons, which ensures that the DNA strands can enter the nanopore. The MinION, with an attached flow cell, was connected to a computer and quality control was performed, which checks for available and active nanopores, the R9.5 flow cell was primed and 75 µl of the prepared 1D^2 library was loaded into the flow cell for sequencing.

The sequencing run was performed for 16 hours, and was controlled by MinKNOW software, which collects the 1D read data. Basecalling was performed initially using Albacore software (v2.3.1, Oxford Nanopore, Oxford, UK), later updated to Guppy software (v3.2.4). The basecaller first converts electrical signal from the 1D reads into a nucleotide sequence, and the sequence within 1D reads from complementary strands are subsequently paired and combined into higher-accuracy 1D^2 reads. 1D^2 reads containing the MinION tag and barcode sequences were demultiplexed inititally by Porechop (v0.2.3) [33], later updated to the same Guppy software. Porechop/Guppy removes the non-HLA tag and barcode sequences from the reads, and sorts the reads into fastq files corresponding to each barcode sequence.

Data analysis and interpretation were performed by a combined approach, using two separate software packages: SeqPilot SeqNext (NGS) HLA module (v4.4.0, JSI, Ettenheim, Germany) and GenDx NGSengine (v2.11.0.11444, GenDx, Utrecht, The Netherlands). Both are configured to ignore the regions containing amplification primers. Analysis of the sorted HLA 1D^2 read data was performed in both software packages independently, and any discrepancies between the two programs were analyzed in detail.

Sanger SBT

HLA sequences obtained by MinION sequencing were compared with typing results from full length group-specific Sanger SBT,[14] which is considered the gold standard for HLA typing in our laboratory. In brief, DNA was isolated and amplified using group- and allele-specific primers for the HLA class I genes. The DNA product was sequenced on a Sanger 3730 analyser (Applied Biosystems), and sequence analysis and allele calling were performed using the JSI SeqPilot SeqHLA module.

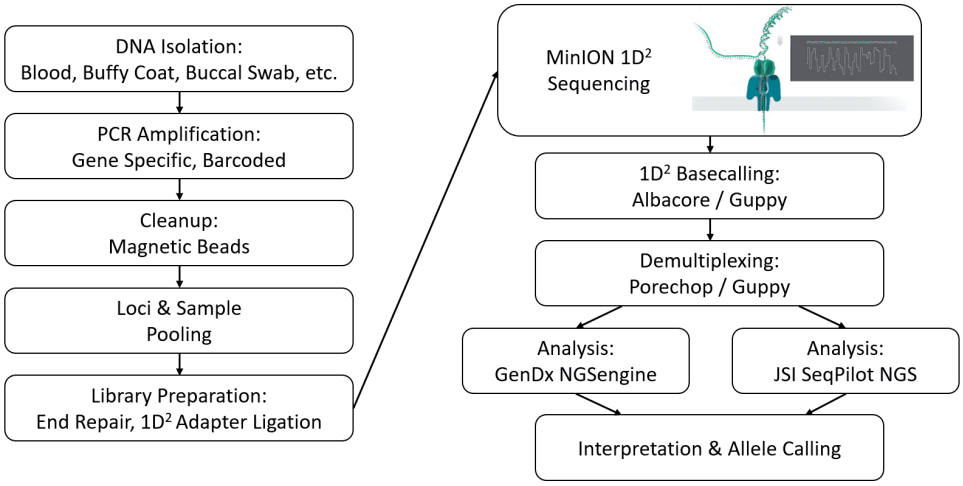


Figure 1. MinION Sequencing and Analysis

The first column depicts the steps of sample preparation, while data generation and analysis is in the second column. Data analysis is performed in a combined analysis approach, where the results from two software packages are compared for accurate allele calls.

Results

Initial Validation

For the initial validation panel, consisting of 33 samples covering 70 of the 71 known HLA class I allele groups (see table 1), HLA-A, B and C were successfully amplified and sequenced using the MinION sequencing method. Data were analyzed and interpreted using two separate software packages: SeqPilot SeqNext (NGS) HLA module (v4.4.0, JSI, Ettenheim, Germany) and GenDx NGSengine (v2.11.0.11444). Since HLA analysis software programs specifically designed for MinION data were not yet available, we have chosen to use a combination of two different HLA analysis programs, both able to deal with data from all common NGS platforms and kits. Comparisons of the typing results with the results of Sanger full-length class I HLA sequence based typing (SSBT) reveals that the MinION $1D^2$ sequencing protocol and redundant analysis approach was 100% concordant with the SSBT typings to third-field resolution (supplementary table 1).

While the majority of the samples was correctly typed to three fields in both software tools immediately, some manual interpretation was necessary in a small percentage of the typing results. Due to the presence of homopolymer stretches, in 14.1% of the cases one of the nucleotides within the homopolymer sequence was ignored in the NGSEngine software, resulting in an ambiguous typing result (see *A* in supplementary table 1). This nucleotide was however correctly identified with the other analysis program, resulting in a correct allele assignment and therefore a correct final HLA typing. For 3 alleles (1.5%) a discrepancy was observed between the allele call obtained with NGSengine and SeqPilot NGS (see *B* in supplementary table 1). In these cases, SeqPilot correctly assigned the allele to two fields, but a region in the introns containing an STR (short tandem repeat) sequence was misaligned, resulting in misalignment of the exon and therefore incorrect third-field allele call. These alleles were, however, correctly typed to the third field by NGSengine, and manual inspection of the analysis details easily resolved the discrepancy. In a single sample NGSengine was unable to assign an allele for HLA-A (see *C* in supplementary table 1), while the same MinION sequence data in SeqPilot NGS gave the correct typing without any problems. Repeating the sample did not solve the problem, whereas other samples with the same typing did not demonstrate this problem. Another problem observed during this initial validation was co-amplification of HLA-Y with HLA-A in samples that were positive for HLA-A*30, *31, *33 and *34. Therefore it was decided to include all pseudogenes in the program analysis to remove any off-target reads.

During initial validation difficult positions and regions were monitored. In most cases it concerned homopolymer or repeat sequence regions in the introns of the genes, not affecting the HLA typing result at the 3rd field level. In these cases the difficult region was automatically ignored in the program. In few cases a difficult region was present in an

Sample ID	HLA-A*		HLA-B*		HLA-C*	
1	01:01:01:01	30:01:01	15:10:01	42:01:01	03:04:02	17:01:01
2	02:06:01:01	02:06:01:04	51:01:01	59:01:01	01:02:01	14:02:01
3	02:01:01:01	-	52:01:01	73:01	07:01:01	15:05:01
4	01:02	66:01:01:01	58:01:01	58:02:01	03:02:02:01	06:02:01
5	02:01:01	36:01	15:03:01:02	51:01:01	01:02:01	12:03:01
6	02:01:01	31:01:02:01	15:01:01:01	67:01:01	07:02:01	-
7	26:01:01:01	30:02:01:01	18:01:01	40:01:02	03:04:01:01	05:01:01:01
8	24:02:01:01	32:01:01:01	14:01:01:01	18:01:01	07:01:01	08:02:01:02
9	01:01:01:01	31:01:02:01	08:01:01	40:01:02	03:04:01:01	07:01:01
10	01:01:01:01	03:01:01:01	08:01:01	45:01:01	06:02:01:03	07:01:01
11	02:06:01:01	30:02:01:01	18:01:01	39:08	05:01:01	07:02:01
12	29:02:01:01	69:01:01:01	39:06:02:01	55:01:01	01:02:01	07:02:01
13	03:01:01:05	66:01:01:01	15:03:01:02	52:01:01:01	02:10:01:01	12:02:02:01
14	43:01	74:01:01	15:03:01:02	44:03	02:10:01:01	08:04:01
15	02:01:01	34:01:01	40:02:01	56:02:01	01:02:01	15:02:01:01
16	30:01:01	33:01:01:01	53:01:01	81:01	04:01:01	08:04:01
17	02:01:01	24:02:01:01	44:02:01:01	49:03	05:01:01:02	07:01:01
18	03:01:01:05	25:01:01:01	37:01:01:01	47:01:01:03	06:02:01:01	-
19	02:03:01	02:07:01	38:02:01	46:01:01	01:02:01	07:02:01
20	01:01:01:11	02:01:01	35:04:01	82:01:01:01	03:02:02:01	04:01:01
21	02:01:01	-	44:09	50:01:01:01	05:01:01:02	-
22	24:02:01:01	26:02:01	40:06:01:01	54:01:01	01:02:01	08:01:01:01
23	24:02:01	33:03:01	15:07:01:02	15:16:01:02	03:03:01	14:02:01:02
24	26:01:01:01	74:01:01	81:01	78:01:01:02	16:01:01	18:01
25	23:01:01:01	32:01:01:01	41:02:01	44:03:01	04:01:01	17:03:01
26	11:01:01:01	24:02:01	27:06	48:01:01	01:02:01	08:01:01:01
27	01:01:38L	02:01:01	15:17:01:01	57:01:01:01	06:02:01	07:01:02
28	02:01:01	25:01:01:01	15:78:01	38:01:01:01	03:04:01:01	12:03:01:01
29	30:09	80:01:01:02	07:02:01	81:01:01	07:02:01	18:02
30	11:01:01:01	68:01:02:01	40:01:02	55:01:01	01:02:01	07:02:01
31	24:02:01:01	29:01:01:01	07:05:01	27:02:01:04	02:02:02	15:05:02
32	24:17	33:03:01	07:02:01	15:02:01:01	07:02:01:03	08:01:01:01
33	02:01:01:01	24:02:01	07:02:01	13:02:01:01	06:02:01:01	07:02:01

Table 1. Samples included in the initial validation.
Samples were selected to cover 70 of the 71 known HLA-A,B,C allele groups. HLA-B*83 was not included, as it was not available in our laboratory.

exon, resulting in an ambiguous typing result if this region would be ignored. However, in most cases the second analysis program was able to analyse this region reliably and therefore the ambiguity was resolved by this second program. In some cases the difficult region concerned a homopolymer in an exon, *e.g.* the homopolymer C region in exon 4 of HLA-A which bears either 7, 8 or 9 C nucleotides and is difficult to distinguish. Since this region is defining several null alleles, we recommend to resolve these cases by either Sanger sequencing of exon 4 or another method which enables detection of null alleles.

Read Distribution

During the initial validation, read distribution between barcodes, loci, and alleles were compared and used to identify imbalances in amplification or sequencing, as well as to define minimum coverage values. During this process it was noticed that the same barcodes gave consistent imbalanced sequence results between different runs and this was independent of the sample. The imbalance in read coverage between the different barcodes could not be explained by differences in amplification efficiency since equimolar pooling of the amplicons was performed before library preparation. Therefore, we assumed that it was an effect of the library preparation, most probably due to the ligation of the adapters. In total, 24 barcode sequences were tested and 2 panels of barcodes were designed to optimize balanced read coverage between different samples. Since each run consists of 9 samples the balance was optimized to be between 10 and 13% for each barcode.

With the equimolar pooling of HLA-A, -B and –C of one sample, the read coverage between the different loci from one sample were comparable and therefore no adaptation in the pooling process was needed. Concerning the allelic distribution, we used the heterozygous positions to determine the median percentage of the second allele being the second most abundant nucleotide at this position. In our initial validation values for the second allele were varying between 36% and 47%, irrespective of the locus. Since all typing results were correct, the criterion of allelic distribution was set at a minimum of 35% for the second allele. In samples that were homozygous for one or more of the HLA genes typed, the software program SeqPilot showed in 3 cases the presence of a second allele (see *D* in supplementary table 1), but in all cases the read proportion was significantly less than the threshold of 35%.

During this initial validation the read coverage per allele was assessed in the NGSEngine program as well. Since the barcode panel was not yet optimized during this validation, the read coverage per allele varied enormously, from 26 to 4070 reads. Even the allele with only 26 reads was correctly typed, indicating that the minimum coverage can be rather low. A minimum coverage per allele was set conservatively at 150 reads, ensuring that typing results are supported by sufficient coverage.

Second Validation

The criteria established with the initial validation, as described above, were verified with a second validation panel. This panel consisted of 67 diagnostic samples that were sequenced for HLA-A, -B and –C with both group-specific Sanger SBT and MinION and analyzed with SeqPilot NGS and NGSengine software packages. These samples comprised 15 different HLA A, 22 different HLA-B and 13 different HLA-C allele groups (supplementary table 2).

In total 198 HLA typing results (98.5%) obtained to the third field level by MinION sequencing were found to be identical to third field level typing by Sanger sequencing, taking the software differences as identified in the initial validation into account (see *A* and *B* supplementary table 2). In one sample that was homozygous for HLA-C the NGSEngine software showed the presence of a second allele, but the proportion was below the threshold of 35% (see *C* in supplementary table 2) In two cases (1%), no typing result was obtained by both programs (see *D* in supplementary table 2), because of lack of amplification of one of the HLA genes. In one case (0.5%) a correct typing result was obtained by one program, whereas the other program was not able to analyse the sequences due to insufficient number of reads (see *E* in supplementary table 2). The criterion of 150 reads per allele and minimal allele distribution of 35% - 65% were verified in all heterozygous samples. The read coverage in the homozygous samples always exceeded 150 reads, and typing results were evaluated in detail by two different persons to ascertain homozygosity.

Discussion

In the current study, the MinION full-length HLA sequencing approach with a combined analysis strategy using SeqPilot NGS and NGSengine software packages, was validated in two steps; an initial validation with 33 samples including all HLA class I allele groups except HLA-B*83, and a second validation running 67 diagnostic samples in parallel with Sanger sequencing. All sample typing results were compared with our previously described group-specific Sanger sequence based typing approach and all were verified to be correct. Validation was focused on three-field analysis, because one of the programs was not able to take the intron sequences into account in the HLA typing results, and because there is still a huge lack of full length reference allele sequences. Furthermore, the presence of homopolymeric regions in the introns also interfered with fourth field allele assignment. Overall, this study showed that the MinION nanopore approach for high resolution HLA typing was valid for diagnostic purposes and as such is now implemented in the routine laboratory setting for high resolution typing of HLA class I.

The current NGS approaches for HLA typing are based on Illumina, Ion Torrent, and PacBio technologies, which each have advantages and disadvantages. One of the major advantages of MinION single molecule sequencing is the generation of long reads that span the entire HLA gene, which allows unambiguous phasing of polymorphism across the gene. Since the HLA genes are very homologous, correct separation of genes and alleles can be a challenge with small DNA pieces. However, the long reads provided by MinION can be easily separated by locus and allele, allowing accurate analysis of multiple HLA genes and alleles. Furthermore, the MinION technology is based on the sequencing of each different PCR strand by directing it through a nanopore, ensuring the sequencing of each PCR strand only once, instead of multiple times like in other approaches. This procedure ensures detection of all different variants in a sample.

Another major advantage is the low price of the equipment, making it feasible to introduce it in even small HLA laboratories with a limited number of samples. Furthermore, the small size of the MinION allows its use in virtually any laboratory without the need for enormous dedicated desk space for equipment. Concerning turnover time, sequencing and allele calling using the MinION approach requires a similar timeframe to other NGS methods,[34] with comparable or even less hands-on time. For high throughput laboratories Oxford Nanopore provides the GridION and PromethION equipment, which can run 5 and 24/48 flow cells at once, respectively.

Analysis of read quality can be challenging, and it is important to choose read quality metrics that are meaningful and unambiguous. The Phred quality scores as reported within the MinION reads are estimates based on the confidence the neural network within the basecaller has in its interpretation of an electric signal. Since the neural network is trained based on reads from known non-HLA DNA sequences, any variation between the nature of the training samples and the targeted DNA may have an influence on read quality. Additionally, as with many sequencing platforms, MinION faces challenges in analyzing regions containing STRs and homopolymers, and STR/homopolymer length is regularly underestimated in the reads. The reason for this is again that the training of the neural network has been established with mostly micro-organism sequences, that don't have any STR or homopolymer regions. This is different to the reason for problems with STR/homopolymers in other NGS methods, which are based on incorporation of nucleotides in the sequence. For sequencing with the MinION, adequate training of the neural network with human sequences or even HLA sequences might solve this problem and open the door to ultrahigh resolution typing of both class I and class II to the allelic resolution level.

The use of $1D^2$ MinION reads, which are created by pairing two single-stranded reads into a single sequence, provides both advantages and challenges. Read pairing and discarding of unpaired reads reduces the amount of data by over half, which is reflected in the final

read coverage. On the other hand, this $1D^2$ method enables a higher reliability in SNP calling. Since random base calling errors can be resolved with additional read coverage, the use of 1D sequencing seems to be feasible in the future for the described MinION approach as well.

MinION is a long-read sequencing platform, and the length of a sequenced region is in fact limited by the sample preparation, and not by the sequencing platform. Read lengths of 2MB have been reported for the MinION platform,[35] which is sufficient for sequencing the whole HLA class I or class II region. Methods like probe capturing[36] or genomic fishing might enable isolation of the MHC region, providing the possibility to abolish PCR amplification and even determine the arrangement of HLA haplotypes by MinION sequencing in the future.

MinION and nanopore sequencing have evolved rapidly from a low quality sequencing platform for research purposes to the now available high standard third generation sequencing method, that can be used to define the hyperpolymorphic HLA genes. Further development and improvement of the 1D procedure will provide higher coverage, lower hands-on time and faster base-calling technology and therefore improve HLA typing in the near future. Overall, the present study describes the validation and implementation of the nanopore sequencing HLA class I typing method and illustrates that nanopore sequencing technology has advanced to a point where it can be used in routine diagnostics with high accuracy.

Acknowledgements:

The authors thank Christel Meertens, Fausto Palucci, Sophie Onclin and Stefan Molenbroeck for their assistance in the validation process. Thanks to Diana van Bakel for her contributions to manuscript submission.

References

1. Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, Donnelly P, McVean G, Przeworski M: Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. Science 2013, 339:1578.
2. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh Steven G E: The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Research 2015, 43:D423-D431.
3. Patel R, Terasaki PI: Significance of the Positive Crossmatch Test in Kidney Transplantation. New England Journal of Medicine 1969, 280:735-739.
4. Duquesnoy RJ: Human leukocyte antigen epitope antigenicity and immunogenicity. Curr Opin Organ Transplant 2014, 19:428-435.
5. El-Awar N, Jucaud V, Nguyen A: HLA Epitopes: The Targets of Monoclonal and Alloantibodies Defined. Journal of Immunology Research 2017, 2017:3406230.
6. Fürst D, Müller C, Vucinic V, Bunjes D, Herr W, Gramatzki M, Schwerdtfeger R, Arnold R, Einsele H, Wulf G, Pfreundschuh M, Glass B, Schrezenmeier H, Schwarz K, Mytilineos J: High-resolution HLA matching in hematopoietic stem cell transplantation: a retrospective collaborative analysis. Blood 2013, 122:3220-3229.
7. Mayor NP, Hayhurst JD, Turner TR, Szydlo RM, Shaw BE, Bultitude WP, Sayno J-R, Tavarozzi F, Latham K, Anthias C, Robinson J, Braund H, Danby R, Perry J, Wilson MC, Bloor AJ, McQuaker IG, MacKinnon S, Marks DI, Pagliuca A, Potter MN, Potter VT, Russell NH, Thomson KJ, Madrigal JA, Marsh SGE: Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. Biology of Blood and Marrow Transplantation 2019, 25:443-450.
8. Fan WL, Shiao MS, Hui RC, Su SC, Wang CW, Chang YC, Chung WH: HLA Association with Drug-Induced Adverse Reactions. Journal of Immunology Research 2017, 2017:3186328.
9. Illing PT, Purcell AW, McCluskey J: The role of HLA genes in pharmacogenomics: unravelling HLA associated adverse drug reactions. Immunogenetics 2017, 69:617-630.
10. Leckband SG, Kelsoe JR, Dunnenberger HM, George AL, Jr., Tran E, Berger R, Muller DJ, Whirl-Carrillo M, Caudle KE, Pirmohamed M: Clinical Pharmacogenetics Implementation Consortium guidelines for HLA-B genotype and carbamazepine dosing. Clin Pharmacol Ther 2013, 94:324-328.
11. Dendrou CA, Petersen J, Rossjohn J, Fugger L: HLA variation and disease. Nature Reviews Immunology 2018, 18:325-339.
12. Erlich H: HLA DNA typing: past, present, and future. Tissue Antigens 2012, 80:1-11.
13. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 1977, 74:5463-5467.
14. Voorter CEM, Palusci F, Tilanus MGJ: Sequence-Based Typing of HLA: An Improved Group-Specific Full-Length Gene Sequencing Approach. Bone Marrow and Stem Cell Transplantation. Edited by Beksaç M. New York, NY: Springer New York, 2014. pp. 101-114.
15. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush

MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ: Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008, 456:53-59.

16. Merriman B, Rothberg JM: Progress in ion torrent semiconductor chip based sequencing. Electrophoresis 2012, 33:3397-3417.
17. Monos D, Maiers MJ: Progressing towards the complete and thorough characterization of the HLA genes by NGS (or single-molecule DNA sequencing): Consequences, opportunities and challenges. Human immunology 2015, 76:883-886.
18. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S: Real-Time DNA Sequencing from Single Polymerase Molecules. Science 2009, 323:133-138.
19. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA: The potential and challenges of nanopore sequencing. Nature Biotechnology 2008, 26:1146-1153.

20. Jain M, Olsen HE, Paten B, Akeson M: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology 2016, 17:239.
21. Lieberman KR, Cherf GM, Doody MJ, Olasagasti F, Kolodji Y, Akeson M: Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. Journal of the American Chemical Society 2010, 132:17961-17972.
22. Wick RR, Judd LM, Holt KE: Comparison of Oxford Nanopore basecalling tools. 2018.
23. Wick RR, Judd LM, Holt KE: Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology 2019, 20:129.
24. Ewing B, Green P: Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998, 8:186-194.
25. Montgomery MC, Liu C, Petraroia R, Weimer ET: Using Nanopore Whole-Transcriptome Sequencing for Human Leukocyte Antigen Genotyping and Correlating Donor Human Leukocyte Antigen Expression with Flow Cytometric Crossmatch Results. The Journal of Molecular Diagnostics 2020, 22:101-110.
26. Matern BM, Olieslagers TI, Voorter CEM, Groeneweg M, Tilanus MGJ: Insights into the polymorphism in HLA-DRA and its evolutionary relationship with HLA haplotypes. HLA 2019.
27. Duke JL, Mosbruger TL, Ferriola D, Chitnis N, Hu T, Tairis N, Margolis DJ, Monos DS: Resolving MiSeq-Generated Ambiguities in HLA-DPB1 Typing by Using the Oxford Nanopore Technology. The Journal of Molecular Diagnostics 2019, 21:852-861.
28. Lang K, Surendranath V, Quenzel P, Schofl G, Schmidt AH, Lange V: Full-Length HLA Class I Genotyping with the MinION Nanopore Sequencer. Methods Mol Biol 2018, 1802:155-162.
29. Liu C, Xiao F, Hoisington-Lopez J, Lang K, Quenzel P, Duffy B, Mitra RD: Accurate Typing of Human Leukocyte Antigen Class I Genes by Oxford Nanopore Sequencing. The Journal of molecular diagnostics : JMD 2018, 20:428-435.
30. Chua EW, Ng PY: MinION: A Novel Tool for Predicting Drug Hypersensitivity? Frontiers in pharmacology 2016, 7:156.
31. Ammar R, Paton TA, Torti D, Shlien A, Bader GD: Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Research 2015, 4:17.
32. Tilanus MGJ: The power of Oxford Nanopore MinION in human leukocyte antigen immunogenetics. Annals of Blood 2017, 2.
33. Wick R: Porechop.
34. Mayor NP, Robinson J, McWhinnie AJM, Ranade S, Eng K, Midwinter W, Bultitude WP, Chin C-S, Bowman B, Marks P, Braund H, Madrigal JA, Latham K, Marsh SGE: HLA Typing for the Next Generation. PloS one 2015, 10:e0127153-e0127153.
35. Payne A, Holmes N, Rakyan V, Loose M: Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. bioRxiv 2018:312256.
36. Dapprich J, Ferriola D, Mackiewicz K, Clark PM, Rappaport E, D'Arcy M, Sasson A, Gai X, Schug J, Kaestner KH, Monos D: The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. BMC genomics 2016, 17:486-486.

Supplementary Table 1. Typing results of the initial validation.

This table shows the *HLA-A*, *-B* and *-C* typing results of the 33 samples included in the initial validation panel. Typing results obtained using MinION sequencing and analysis by NGSengine or SeqPilot NGS are compared with the typing result obtained using Sanger sequencing and analysis by SeqPilot SBT.

* Ambiguous typing result in NGSengine caused by an ignored nucleotide in an homopolymeric region.

† Discrepancy between the typing result of NGSengine and SeqPilot NGS, caused by a misalignment of an STR region in SeqPilot NGS.

‡ No typing result by analysis using NGSengine.

§ Incorrect second allele assignment of a homozygous sample by SeqPilot NGS.

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot FL Sanger
1	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*30:01:01	*30:01:01	*30:01:01
2	HLA-A Allele 1*	*02:06:01 *02:602	*02:06:01	*02:06:01
3	HLA-A Allele 1*	*02:01:01 *02:59 *02:388	*02:01:01	*02:01:01
4	HLA-A Allele 1	*01:02	*01:02	*01:02
	HLA-A Allele 2	*66:01:01	*66:01:01	*66:01:01
5	HLA-A Allele 1	*02:01:01	*02:01:01	*02:01:01
	HLA-A Allele 2	*36:01	*36:01	*36:01
6	HLA-A Allele 1	*02:01:01	*02:01:01	*02:01:01
	HLA-A Allele 2	*31:01:02	*31:01:02	*31:01:02
7	HLA-A Allele 1	*26:01:01	*26:01:01	*26:01:01
	HLA-A Allele 2	*30:02:01	*30:02:01	*30:02:01
8	HLA-A Allele 1‡	no typing	*24:02:01	*24:02:01
	HLA-A Allele 2‡	no typing	*32:01:01	*32:01:01
9	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*31:01:02	*31:01:02	*31:01:02
10	HLA-A Allele 1*	*01:01:01 *01:04N	*01:01:01	*01:01:01
	HLA-A Allele 2*	*03:01:01 *03:21N *03:279N	*03:01:01	*03:01:01
11	HLA-A Allele 1	*02:06:01	*02:06:01	*02:06:01
	HLA-A Allele 2	*30:02:01	*30:02:01	*30:02:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot FL Sanger
12	HLA-A Allele 1	*29:02:01	*29:02:01	*29:02:01
	HLA-A Allele 2	*69:01:01	*69:01:01	*69:01:01
13	HLA-A Allele 1	*03:01:01	*03:01:01	*03:01:01
	HLA-A Allele 2	*66:01:01	*66:01:01	*66:01:01
14	HLA-A Allele 1	*43:01	*43:01	*43:01
	HLA-A Allele 2	*74:01:01	*74:01:01	*74:01:01
15	HLA-A Allele 1	*02:01:01	*02:01:01	*02:01:01
	HLA-A Allele 2	*34:01:01	*34:01:01	*34:01:01
16	HLA-A Allele 1	*30:01:01	*30:01:01	*30:01:01
	HLA-A Allele 2	*33:01:01	*33:01:01	*33:01:01
17	HLA-A Allele 1	*02:01:01	*02:01:01	*02:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
18	HLA-A Allele 1	*03:01:01	*03:01:01	*03:01:01
	HLA-A Allele 2	*25:01:01	*25:01:01	*25:01:01
19	HLA-A Allele 1	*02:03:01	*02:03:01	*02:03:01
	HLA-A Allele 2	*02:07:01	*02:07:01	*02:07:01
20	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*02:01:01	*02:01:01	*02:01:01
21	HLA-A Allele 1*	*02:01:01 *02:59 *02:388	*02:01:01	*02:01:01
22	HLA-A Allele 1	*24:02:01	*24:02:01	*24:02:01
	HLA-A Allele 2	*26:02:01	*26:02:01	*26:02:01
23	HLA-A Allele 1*	*24:02:01 *24:11N	*24:02:01	*24:02:01
	HLA-A Allele 2*	*33:03:01 *33:73N	*33:03:01	*33:03:01
24	HLA-A Allele 1	*26:01:01	*26:01:01	*26:01:01
	HLA-A Allele 2	*74:01:01	*74:01:01	*74:01:01
25	HLA-A Allele 1	*23:01:01	*23:01:01	*23:01:01
	HLA-A Allele 2*	*32:01:01 *32:01:23 *32:102	*32:01:01	*32:01:01
26	HLA-A Allele 1*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
27	HLA-A Allele 1	*01:01:38L	*01:01:38L	*01:01:38L
	HLA-A Allele 2*	*02:01:01 *02:59 *02:388	*02:01:01	*02:01:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot FL Sanger
28	HLA-A Allele 1*	*02:01:01 *02:59 *02:388	*02:01:01	*02:01:01
	HLA-A Allele 2*	*25:01:01 *25:32	*25:01:01	*25:01:01
29	HLA-A Allele 1	*30:09	*30:09	*30:09
	HLA-A Allele 2	*80:01:01	*80:01:01	*80:01:01
30	HLA-A Allele 1*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
	HLA-A Allele 2	*68:01:02	*68:01:02	*68:01:02
31	HLA-A Allele 1	*24:02:01	*24:02:01	*24:02:01
	HLA-A Allele 2	*29:01:01	*29:01:01	*29:01:01
32	HLA-A Allele 1	*24:17	*24:17	*24:17
	HLA-A Allele 2	*33:03:01	*33:03:01	*33:03:01
33	HLA-A Allele 1	*02:01:01	*02:01:01	*02:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
1	HLA-B Allele 1	*15:10:01	*15:10:01	*15:10:01
	HLA-B Allele 2	*42:01:01	*42:01:01	*42:01:01
2	HLA-B Allele 1	*51:01:01	*51:01:01	*51:01:01
	HLA-B Allele 2	*59:01:01	*59:01:01	*59:01:01
3	HLA-B Allele 1	*52:01:01	*52:01:01	*52:01:01
	HLA-B Allele 2	*73:01	*73:01	*73:01
4	HLA-B Allele 1	*58:01:01	*58:01:01	*58:01:01
	HLA-B Allele 2	*58:02:01	*58:02:01	*58:02:01
5	HLA-B Allele 1	*15:03:01	*15:03:01	*15:03:01
	HLA-B Allele 2	*51:01:01	*51:01:01	*51:01:01
6	HLA-B Allele 1	*15:01:01	*15:01:01	*15:01:01
	HLA-B Allele 2	*67:01:01	*67:01:01	*67:01:01
7	HLA-B Allele 1†	*18:01:01	*18:01:25	*18:01:01
	HLA-B Allele 2	*40:01:02	*40:01:02	*40:01:02
8	HLA-B Allele 1	*14:01:01	*14:01:01	*14:01:01
	HLA-B Allele 2	*18:01:01	*18:01:01	*18:01:01
9	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*40:01:02	*40:01:02	*40:01:02
10	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*45:01:01	*45:01:01	*45:01:01
11	HLA-B Allele 1†	*18:01:01	*18:01:25	*18:01:01
	HLA-B Allele 2	*39:08	*39:08	*39:08

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot FL Sanger
12	HLA-B Allele 1	*39:06:02	*39:06:02	*39:06:02
	HLA-B Allele 2	*55:01:01	*55:01:01	*55:01:01
13	HLA-B Allele 1	*15:03:01	*15:03:01	*15:03:01
	HLA-B Allele 2	*52:01:01	*52:01:01	*52:01:01
14	HLA-B Allele 1	*15:03:01	*15:03:01	*15:03:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01
15	HLA-B Allele 1	*40:02:01	*40:02:01	*40:02:01
	HLA-B Allele 2	*56:02:01	*56:02:01	*56:02:01
16	HLA-B Allele 1	*53:01:01	*53:01:01	*53:01:01
	HLA-B Allele 2	*81:01	*81:01	*81:01
17	HLA-B Allele 1	*44:02:01	*44:02:01	*44:02:01
	HLA-B Allele 2	*49:03	*49:03	*49:03
18	HLA-B Allele 1	*37:01:01	*37:01:01	*37:01:01
	HLA-B Allele 2	*47:01:01	*47:01:01	*47:01:01
19	HLA-B Allele 1	*38:02:01	*38:02:01	*38:02:01
	HLA-B Allele 2	*46:01:01	*46:01:01	*46:01:01
20	HLA-B Allele 1	*35:04:01	*35:04:01	*35:04:01
	HLA-B Allele 2	*82:01	*82:01	*82:01
21	HLA-B Allele 1	*44:09	*44:09	*44:09
	HLA-B Allele 2	*50:01:01	*50:01:01	*50:01:01
22	HLA-B Allele 1	*40:06:01	*40:06:01	*40:06:01
	HLA-B Allele 2	*54:01:01	*54:01:01	*54:01:01
23	HLA-B Allele 1	*15:07:01	*15:07:01	*15:07:01
	HLA-B Allele 2	*15:16:01	*15:16:01	*15:16:01
24	HLA-B Allele 1	*78:01:01	*78:01:01	*78:01:01
	HLA-B Allele 2	*81:01	*81:01	*81:01
25	HLA-B Allele 1	*41:02:01	*41:02:01	*41:02:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01
26	HLA-B Allele 1	*27:06	*27:06	*27:06
	HLA-B Allele 2	*48:01:01	*48:01:01	*48:01:01
27	HLA-B Allele 1	*15:17:01	*15:17:01	*15:17:01
	HLA-B Allele 2	*57:01:01	*57:01:01	*57:01:01
28	HLA-B Allele 1	*15:78:01	*15:78:01	*15:78:01
	HLA-B Allele 2	*38:01:01	*38:01:01	*38:01:01
29	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*81:01	*81:01	*81:01
30	HLA-B Allele 1	*40:01:02	*40:01:02	*40:01:02
	HLA-B Allele 2	*55:01:01	*55:01:01	*55:01:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot FL Sanger
31	HLA-B Allele 1	*07:05:01	*07:05:01	*07:05:01
	HLA-B Allele 2	*27:02:01	*27:02:01	*27:02:01
32	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*15:02:01	*15:02:01	*15:02:01
33	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*13:02:01	*13:02:01	*13:02:01
1	HLA-C Allele 1	*03:04:02	*03:04:02	*03:04:02
	HLA-C Allele 2†	*17:01:01	*17:01:02	*17:01:01
2	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*14:02:01	*14:02:01	*14:02:01
3	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2	*15:05:01	*15:05:01	*15:05:01
4	HLA-C Allele 1	*03:02:02	*03:02:02	*03:02:02
	HLA-C Allele 2	*06:02:01	*06:02:01	*06:02:01
5	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*12:03:01	*12:03:01	*12:03:01
6	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2§	-	*07:02:11	-
7	HLA-C Allele 1	*03:04:01	*03:04:01	*03:04:01
	HLA-C Allele 2	*05:01:01	*05:01:01	*05:01:01
8	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2	*08:02:01	*08:02:01	*08:02:01
9	HLA-C Allele 1	*03:04:01	*03:04:01	*03:04:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01
10	HLA-C Allele 1	*06:02:01	*06:02:01	*06:02:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01
11	HLA-C Allele 1	*05:01:01	*05:01:01	*05:01:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
12	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
13	HLA-C Allele 1	*02:10:01	*02:10:01	*02:10:01
	HLA-C Allele 2	*12:02:02	*12:02:02	*12:02:02
14	HLA-C Allele 1	*02:10:01	*02:10:01	*02:10:01
	HLA-C Allele 2	*08:04:01	*08:04:01	*08:04:01
15	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*15:02:01	*15:02:01	*15:02:01
16	HLA-C Allele 1	*04:01:01	*04:01:01	*04:01:01
	HLA-C Allele 2	*08:04:01	*08:04:01	*08:04:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot FL Sanger
17	HLA-C Allele 1	*05:01:01	*05:01:01	*05:01:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01
18	HLA-C Allele 1	*06:02:01	*06:02:01	*06:02:01
	HLA-C Allele 2§	-	*06:123	-
19	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
20	HLA-C Allele 1	*03:02:02	*03:02:02	*03:02:02
	HLA-C Allele 2	*04:01:01 *04:134 *04:217N	*04:01:01	*04:01:01
21	HLA-C Allele 1	*05:01:01	*05:01:01	*05:01:01
	HLA-C Allele 2§	-	*05:142	-
22	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2*	*08:01:01 *08:130N	*08:01:01	*08:01:01
23	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*14:02:01	*14:02:01	*14:02:01
24	HLA-C Allele 1	*16:01:01	*16:01:01	*16:01:01
	HLA-C Allele 2	*18:01	*18:01	*18:01
25	HLA-C Allele 1*	*04:01:01 *04:134 *04:217N	*04:01:01	*04:01:01
	HLA-C Allele 2	*17:03:01	*17:03:01	*17:03:01
26	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2*	*08:01:01 *08:130N	*08:01:01	*08:01:01
27	HLA-C Allele 1	*06:02:01	*06:02:01	*06:02:01
	HLA-C Allele 2	*07:01:02	*07:01:02	*07:01:02
28	HLA-C Allele 1	*03:04:01	*03:04:01	*03:04:01
	HLA-C Allele 2	*12:03:01	*12:03:01	*12:03:01
29	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2	*18:02	*18:02	*18:02
30	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
31	HLA-C Allele 1	*02:02:02	*02:02:02	*02:02:02
	HLA-C Allele 2	*15:05:02	*15:05:02	*15:05:02
32	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2*	*08:01:01 *08:130N	*08:01:01	*08:01:01
33	HLA-C Allele 1	*06:02:01	*06:02:01	*06:02:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01

Supplementary Table 2. Typing results of the second validation.

This table shows the *HLA-A*, *-B* and *-C* typing results of the 67 samples included in the second validation panel. Typing results obtained using MinION sequencing and analysis by NGSengine or SeqPilot NGS are compared with the typing result obtained using Sanger sequencing and analysis by SeqPilot SBT.

* Ambiguous typing result in NGSengine caused by an ignored nucleotide in an homopolymeric region.

† Discrepancy between the typing result of NGSengine and SeqPilot NGS, caused by a misalignment of an STR region in SeqPilot NGS.

‡ Incorrect second allele assignment of a homozygous sample by NGSEngine.

§ No typing result by NGSengine and SeqPilot NGS, due to no amplification of one of the HLA genes.

‖ No typing result by SeqPilot NGS, due to insufficient number of reads.

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
1	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
2	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*25:01:01	*25:01:01	*25:01:01
3	HLA-A Allele 1*	*32:01:01 *32:01:23 *32:14	*32:01:01	*32:01:01
4	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*31:01:02	*31:01:02	*31:01:02
5	HLA-A Allele 1*	*03:01:01 *03:279N	*03:01:01	*03:01:01
	HLA-A Allele 2	*26:01:01	*26:01:01	*26:01:01
6	HLA-A Allele 1*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
	HLA-A Allele 2	*26:01:01	*26:01:01	*26:01:01
7	HLA-A Allele 1	*25:01:01	*25:01:01	*25:01:01
	HLA-A Allele 2	*68:01:02	*68:01:02	*68:01:02
8	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*68:01:01	*68:01:01	*68:01:01
9	HLA-A Allele 1*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
	HLA-A Allele 2	*68:01:02	*68:01:02	*68:01:02
10	HLA-A Allele 1	*23:17:01	*23:17	*23:17
	HLA-A Allele 2	*74:01:01	*74:01:01	*74:01:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
11	HLA-A Allele 1	*26:01:01	*26:01:01	*26:01:01
	HLA-A Allele 2	*68:01:02	*68:01:02	*68:01:02
12	HLA-A Allele 1*	*03:01:01 *03:279N	*03:01:01	*03:01:01
	HLA-A Allele 2*	*32:01:01 *32:01:23	*32:01:01	*32:01:01
13	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
14	HLA-A Allele 1*	*02:01:01 *02:01:132 *02:740	*02:01:01	*02:01:01
15	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
16	HLA-A Allele 1*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
	HLA-A Allele 2	*11:01:01	*11:01:01	*11:01:01
17	HLA-A Allele 1*	*03:01:01 *03:279N	*03:01:01	*03:01:01
	HLA-A Allele 2	*68:01:02	*68:01:02	*68:01:02
18	HLA-A Allele 1	*26:01:01	*26:01:01	*26:01:01
	HLA-A Allele 2	*30:02:01	*30:02:01	*30:02:01
19	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
20	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2*	*03:01:01 *03:279N	*03:01:01	*03:01:01
21	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
22	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
23	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
24	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2*	*32:01:01 *32:01:23	*32:01:01	*32:01:01
25	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*68:01:02	*68:01:02	*68:01:02

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
26	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
27	HLA-A Allele 1	*24:02:01	*24:02:01	*24:02:01
	HLA-A Allele 2	*26:01:01	*26:01:01	*26:01:01
28	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
29	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*29:02:01	*29:02:01	*29:02:01
30	HLA-A Allele 1*	*03:01:01 *03:279N	*03:01:01	*03:01:01
31	HLA-A Allele 1*	*03:01:01 *03:279N	*03:01:01	*03:01:01
	HLA-A Allele 2	*31:01:02	*31:01:02	*31:01:02
32	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2*	*02:01:01 *02:740	*02:01:01	*02:01:01
33	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2*	*03:01:01 *03:279N	*03:01:01	*03:01:01
34	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2*	*03:01:01 *03:279N	*03:01:01	*03:01:01
35	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
36	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*29:02:01	*29:02:01	*29:02:01
37	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2*	*02:01:01 *02:740	*02:01:01	*02:01:01
38	HLA-A Allele 1	*02:05:01	*02:05:01	*02:05:01
	HLA-A Allele 2	*31:01:02	*02:06:01	*31:01:02
39	HLA-A Allele 1*	*02:01:01 *02:740	*02:05:01	*02:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
40	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
41	HLA-A Allele 1	*29:02:01	*29:02:01	*29:02:01
	HLA-A Allele 2*	*32:01:01 *32:01:23	*32:01:01	*32:01:01
42	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*25:01:01	*25:01:01	*25:01:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
43	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*02:06:01	*02:06:01	*02:06:01
44	HLA-A Allele 1	*03:01:01	*03:01:01	*03:01:01
	HLA-A Allele 2	*30:01:01	*30:01:01	*30:01:01
45	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2*	*02:01:01 *02:740	*02:01:01	*02:01:01
46	HLA-A Allele 1*	*03:01:01 *03:206 *03:279N	*03:01:01 *03:279N	*03:01:01
47	HLA-A Allele 1*	*03:01:01 *03:279N	*03:01:01	*03:01:01
	HLA-A Allele 2*	*32:01:01 *32:01:23	*32:01:01	*32:01:01
48	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2*	*02:01:01 *02:740	*02:01:01	*02:01:01
49	HLA-A Allele 1§	No PCR Product.	No PCR Product.	*01:01:01
	HLA-A Allele 2§			*11:01:01
50	HLA-A Allele 1	*24:02:01	*24:02:01	*24:02:01
	HLA-A Allele 2*	*32:01:01 *32:01:23	*32:01:01	*32:01:01
51	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
52	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*29:02:01	*29:02:01	*29:02:01
53	HLA-A Allele 1*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
54	HLA-A Allele 1	*24:02:01	*24:02:01	*24:02:01
	HLA-A Allele 2*	*32:01:01 *32:01:23	*32:01:01	*32:01:01
55	HLA-A Allele 1	*24:02:01	*24:02:01	*24:02:01
	HLA-A Allele 2*	*32:01:01 *32:01:23	*32:01:01	*32:01:01
56	HLA-A Allele 1*	*03:01:01 *03:279N	*03:01:01	*03:01:01
	HLA-A Allele 2	*26:08	*26:08	*26:08
57	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2	*02:01:01	*02:01:01	*02:01:01
58	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2*	*02:01:01 *02:740	*02:01:01	*02:01:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
59	HLA-A Allele 1*	*11:01:01 *11:01:53 *11:01:64	*11:01:01	*11:01:01
	HLA-A Allele 2	*24:02:01	*24:02:01	*24:02:01
60	HLA-A Allele 1	*01:01:01	*01:01:01	*01:01:01
	HLA-A Allele 2*	*02:01:01 *02:740	*02:01:01	*02:01:01
61	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*29:02:01	*29:02:01	*29:02:01
62	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2*	*03:01:01 *03:279N	*03:01:01	*03:01:01
63	HLA-A Allele 1	*26:01:01	*26:01:01	*26:01:01
	HLA-A Allele 2	*68:03:01	*68:03:01	*68:03:01
64	HLA-A Allele 1	*33:03:01	*33:03:01	*33:03:01
	HLA-A Allele 2	*74:01:01	*74:01:01	*74:01:01
65	HLA-A Allele 1*	*02:01:01 *02:740	*02:01:01	*02:01:01
	HLA-A Allele 2	*02:05:01	*02:05:01	*02:05:01
66	HLA-A Allele 1	*31:01:02	*31:01:02	*31:01:02
	HLA-A Allele 2	*68:01:01	*68:01:01	*68:01:01
67	HLA-A Allele 1*	*01:01:01 *01:01:63 *01:01:68	*01:01:01	*01:01:01
1	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*44:02:01	*44:02:01	*44:02:01
2	HLA-B Allele 1	*15:01:01	*15:01:01	*15:01:01
	HLA-B Allele 2	*18:01:01	*18:01:01	*18:01:01
3	HLA-B Allele 1	*27:05:02	*27:05:02	*27:05:02
	HLA-B Allele 2	*35:03:01	*35:03:01	*35:03:01
4	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*15:01:01	*15:01:01	*15:01:01
5	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*38:01:01	*38:01:01	*38:01:01
6	HLA-B Allele 1	*51:01:01	*51:01:01	*51:01:01
7	HLA-B Allele 1†	*18:01:01	*18:01:25	*18:01:01
	HLA-B Allele 2	*40:01:02	*40:01:02	*40:01:02
8	HLA-B Allele 1	*27:05:02	*27:05:02	*27:05:02
	HLA-B Allele 2	*44:27:01	*44:27:01	*44:27:01
9	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*27:05:02	*27:05:02	*27:05:02

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
10	HLA-B Allele 1	*14:01:01	*14:01:01	*14:01:01
	HLA-B Allele 2	*15:03:01	*15:03:01	*15:03:01
11	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*45:01:01	*45:01:01	*45:01:01
12	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*51:01:01	*51:01:01	*51:01:01
13	HLA-B Allele 1	*39:01:01	*39:01:01	*39:01:01
	HLA-B Allele 2	*40:06:01	*40:06:01	*40:06:01
14	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
15	HLA-B Allele 1	*15:01:01	*15:01:01	*15:01:01
	HLA-B Allele 2	*51:01:01	*51:01:01	*51:01:01
16	HLA-B Allele 1	*35:01:01	*35:01:01	*35:01:01
	HLA-B Allele 2	*40:01:02	*40:01:02	*40:01:02
17	HLA-B Allele 1	*35:01:01	*35:01:01	*35:01:01
	HLA-B Allele 2	*40:01:02	*40:01:02	*40:01:02
18	HLA-B Allele 1	*38:01:01	*38:01:01	*38:01:01
	HLA-B Allele 2	*56:01:01	*56:01:01	*56:01:01
19	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*44:02:01	*44:02:01	*44:02:01
20	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*49:01:01	*49:01:01	*49:01:01
21	HLA-B Allele 1	*41:01:01	*41:01:01	*41:01:01
	HLA-B Allele 2	*44:02:01	*44:02:01	*44:02:01
22	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*55:01:01	*55:01:01	*55:01:01
23	HLA-B Allele 1	*35:01:01	*35:01:01	*35:01:01
	HLA-B Allele 2	*56:01:01	*56:01:01	*56:01:01
24	HLA-B Allele 1	*40:02:01	*40:02:01	*40:02:01
	HLA-B Allele 2	*44:02:01	*44:02:01	*44:02:01
25	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*44:02:01	*44:02:01	*44:02:01
26	HLA-B Allele 1	*15:01:01	*15:01:01	*15:01:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01
27	HLA-B Allele 1	*27:05:02	*27:05:02	*27:05:02
28	HLA-B Allele 1	*08:01:20	*08:01:20	*08:01:20
	HLA-B Allele 2	*40:01:02	*40:01:02	*40:01:02
29	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
30	HLA-B Allele 1	*15:01:01	*15:01:01	*15:01:01
	HLA-B Allele 2	*51:01:01	*51:01:01	*51:01:01
31	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*40:01:02	*40:01:02	*40:01:02
32	HLA-B Allele 1	*39:01:01	*39:01:01	*39:01:01
	HLA-B Allele 2	*44:02:01	*44:02:01	*44:02:01
33	HLA-B Allele 1	*15:01:01	*15:01:01	*15:01:01
	HLA-B Allele 2	*40:01:02	*40:01:02	*40:01:02
34	HLA-B Allele 1	*27:05:02	*27:05:02	*27:05:02
	HLA-B Allele 2	*44:02:01	*44:02:01	*44:02:01
35	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*15:01:01	*15:01:01	*15:01:01
36	HLA-B Allele 1	*35:02:01	*35:02:01	*35:02:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01
37	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*15:01:01	*15:01:01	*15:01:01
38	HLA-B Allele 1	*40:11:01	*40:11:01	*40:11:01
	HLA-B Allele 2	*50:02	*50:02	*50:02
39	HLA-B Allele 1	*15:15	*15:15	*15:15
	HLA-B Allele 2	*15:39:01	*15:39:01	*15:39:01
40	HLA-B Allele 1	*40:06:01	*40:06:01	*40:06:01
	HLA-B Allele 2	*46:01:01	*46:01:01	*46:01:01
41	HLA-B Allele 1	*44:02:01	*44:02:01	*44:02:01
42	HLA-B Allele 1	*18:01:01	*18:01:01	*18:01:01
	HLA-B Allele 2	*56:01:01	*56:01:01	*56:01:01
43	HLA-B Allele 1	*15:17:01	*15:17:01	*15:17:01
	HLA-B Allele 2	*39:05:01	*39:05:01	*39:05:01
44	HLA-B Allele 1	*07:161N	*07:161N	*07:161N
	HLA-B Allele 2	*42:02:01	*42:02:01	*42:02:01
45	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*08:01:01	*08:01:01	*08:01:01
46	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*35:01:01	*35:01:01	*35:01:01
47	HLA-B Allele 1	*35:01:01	*35:01:01	*35:01:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01
48	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*39:01:01	*39:01:01	*39:01:01
49	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
50	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*39:01:01	*39:01:01	*39:01:01
51	HLA-B Allele 1	*35:01:01	*35:01:01	*35:01:01
	HLA-B Allele 2	*57:01:01	*57:01:01	*57:01:01
52	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01
53	HLA-B Allele 1	*15:01:01	*15:01:01	*15:01:01
	HLA-B Allele 2	*35:01:01	*35:01:01	*35:01:01
54	HLA-B Allele 1‖	*40:02:01	no typing	*40:02:01
	HLA-B Allele 2‖	*50:01:01	no typing	*50:01:01
55	HLA-B Allele 1†	*18:01:01	*18:01:25	*18:01:01
	HLA-B Allele 2	*39:01:01	*39:01:01	*39:01:01
56	HLA-B Allele 1	*45:01:01	*45:01:01	*45:01:01
	HLA-B Allele 2	*49:01:01	*49:01:01	*49:01:01
57	HLA-B Allele 1	*39:01:01	*39:01:01	*39:01:01
	HLA-B Allele 2	*44:02:01	*44:02:01	*44:02:01
58	HLA-B Allele 1	*39:01:01	*39:01:01	*39:01:01
	HLA-B Allele 2	*44:27:01	*44:27:01	*44:27:01
59	HLA-B Allele 1	*35:01:01	*35:01:01	*35:01:01
	HLA-B Allele 2	*58:01:01	*58:01:01	*58:01:01
60	HLA-B Allele 1	*08:01:01	*08:01:01	*08:01:01
	HLA-B Allele 2	*15:01:01	*15:01:01	*15:01:01
61	HLA-B Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01
62	HLA-B Allele 1	*15:01:01	*15:01:01	*15:01:01
	HLA-B Allele 2	*44:03:01	*44:03:01	*44:03:01
63	HLA-B Allele 1	*14:01:01	*14:01:01	*14:01:01
	HLA-B Allele 2	*35:48	*35:48	*35:48
64	HLA-B Allele 1	*49:01:01	*49:01:01	*49:01:01
	HLA-B Allele 2	*51:01:01	*51:01:01	*51:01:01
65	HLA-B Allele 1	*35:17:01	*35:17:01	*35:17:01
	HLA-B Allele 2	*58:01:01	*58:01:01	*58:01:01
66	HLA-B Allele 1	*35:12:01	*35:12:01	*35:12:01
	HLA-B Allele 2	*45:01:01	*45:01:01	*45:01:01
67	HLA-B Allele 1	*55:01:01	*55:01:01	*55:01:01
	HLA-B Allele 2	*57:01:01	*57:01:01	*57:01:01
1	HLA-C Allele 1	*05:01:01	*05:01:01	*05:01:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
2	HLA-C Allele 1	*03:04:01	*03:04:01	*03:04:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
3	HLA-C Allele 1	*02:02:02	*02:02:02	*02:02:02
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
4	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
5	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
6	HLA-C Allele 1	*05:01:01	*05:01:01	*05:01:01
	HLA-C Allele 2	*15:02:01	*15:02:01	*15:02:01
7	HLA-C Allele 1*	*03:04:01 *03:172	*03:04:01	*03:04:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
8	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*07:04:01	*07:04:01	*07:04:01
9	HLA-C Allele 1	*02:02:02	*02:02:02	*02:02:02
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
10	HLA-C Allele 1	*02:10:01	*02:10:01	*02:10:01
	HLA-C Allele 2	*08:02:01	*08:02:01	*08:02:01
11	HLA-C Allele 1	*06:02:01	*06:02:01	*06:02:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
12	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2	*15:02:01	*15:02:01	*15:02:01
13	HLA-C Allele 1*	*12:03:01 *12:225	*12:03:01	*12:03:01
	HLA-C Allele 2	*15:02:01	*15:02:01	*15:02:01
14	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
15	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*14:02:01	*14:02:01	*14:02:01
16	HLA-C Allele 1*	*03:04:01 *03:172	*03:04:01	*03:04:01
	HLA-C Allele 2	*04:01:01	*04:01:01	*04:01:01
17	HLA-C Allele 1*	*03:04:01 *03:172	*03:04:01	*03:04:01
	HLA-C Allele 2*	*04:01:01 *04:134 *04:217N	*04:01:01	*04:01:01
18	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
19	HLA-C Allele 1	*05:01:01	*05:01:01	*05:01:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
20	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
21	HLA-C Allele 1	*05:01:01	*05:01:01	*05:01:01
	HLA-C Allele 2†	*17:01:01	*17:01:02	*17:01:01
22	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01
23	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*04:01:01	*04:01:01	*04:01:01
24	HLA-C Allele 1	*02:02:02	*02:02:02	*02:02:02
	HLA-C Allele 2	*05:01:01	*05:01:01	*05:01:01
25	HLA-C Allele 1	*02:02:02	*02:02:02	*02:02:02
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01
26	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*16:01:01	*16:01:01	*16:01:01
27	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
28	HLA-C Allele 1	*03:04:01	*03:04:01	*03:04:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01
29	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2	*16:01:01	*16:01:01	*16:01:01
30	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*03:03:01	*03:03:01	*03:03:01
31	HLA-C Allele 1	*03:04:01	*03:04:01	*03:04:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
32	HLA-C Allele 1	*07:04:01	*07:04:01	*07:04:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
33	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*03:04:01	*03:04:01	*03:04:01
34	HLA-C Allele 1	*02:02:02	*02:02:02	*02:02:02
	HLA-C Allele 2	*05:01:01	*05:01:01	*05:01:01
35	HLA-C Allele 1	*03:04:01	*03:04:01	*03:04:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
36	HLA-C Allele 1	*04:01:01	*04:01:01	*04:01:01
	HLA-C Allele 2	*16:01:01	*16:01:01	*16:01:01
37	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
38	HLA-C Allele 1*	*03:04:01 *03:172	*03:04:01	*03:04:01
	HLA-C Allele 2	*06:02:01	*06:02:01	*06:02:01
39	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*03:03:01	*03:03:01	*03:03:01
40	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*08:01:01 *08:130N	*08:01:01	*08:01:01
41	HLA-C Allele 1	*07:04:01	*07:04:01	*07:04:01
42	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
43	HLA-C Allele 1	*07:01:02	*07:01:02	*07:01:02
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
44	HLA-C Allele 1	*07:18	*07:18	*07:18
	HLA-C Allele 2	*17:01:01	*17:01:01	*17:01:01
45	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2	*07:02:01	*07:02:01	*07:02:01
46	HLA-C Allele 1§	No PCR Product.	No PCR Product.	*04:01:01
	HLA-C Allele 2§			*07:02:01
47	HLA-C Allele 1*	*04:01:01 *04:82 *04:134 *04:217N	*04:01:01	*04:01:01
48	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
49	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2‡	*07:148	-	-
50	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
51	HLA-C Allele 1*	*04:01:01 *04:134 *04:217N	*04:01:01	*04:01:01
	HLA-C Allele 2	*06:02:01	*06:02:01	*06:02:01
52	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2	*16:01:01	*16:01:01	*16:01:01
53	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*04:01:01	*04:01:01	*04:01:01
54	HLA-C Allele 1	*02:02:02	*02:02:02	*02:02:02
	HLA-C Allele 2	*06:02:01	*06:02:01	*06:02:01

Sample	Locus	NGSengine	SeqPilot NGS	SeqPilot Sanger
55	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
56	HLA-C Allele 1	*06:02:01	*06:02:01	*06:02:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01
57	HLA-C Allele 1	*07:04:01	*07:04:01	*07:04:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
58	HLA-C Allele 1	*07:04:01	*07:04:01	*07:04:01
	HLA-C Allele 2*	*12:03:01 *12:225	*12:03:01	*12:03:01
59	HLA-C Allele 1	*04:01:01	*04:01:01	*04:01:01
	HLA-C Allele 2	*07:18	*07:18	*07:18
60	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*07:01:01	*07:01:01	*07:01:01
61	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2	*16:01:01	*16:01:01	*16:01:01
62	HLA-C Allele 1	*01:02:01	*01:02:01	*01:02:01
	HLA-C Allele 2	*04:01:01	*04:01:01	*04:01:01
63	HLA-C Allele 1	*07:02:01	*07:02:01	*07:02:01
	HLA-C Allele 2	*08:02:01	*08:02:01	*08:02:01
64	HLA-C Allele 1	*07:01:01	*07:01:01	*07:01:01
	HLA-C Allele 2	*16:01:01	*16:01:01	*16:01:01
65	HLA-C Allele 1*	*04:01:01 *04:134 *04:217N	*04:01:01	*04:01:01
	HLA-C Allele 2	*07:18	*07:18	*07:18
66	HLA-C Allele 1*	*04:01:01 *04:134 *04:217N	*04:01:01	*04:01:01
	HLA-C Allele 2	*06:02:01	*06:02:01	*06:02:01
67	HLA-C Allele 1	*03:03:01	*03:03:01	*03:03:01
	HLA-C Allele 2	*06:02:01	*06:02:01	*06:02:01

CHAPTER 5

5

A novel multiplexed 11 locus PCR assay using next generation sequencing

L. Truong[1,3], B.M. Matern[2], L. D'Orsogna[1,3], P. Martinez[1,3], M.G.J. Tilanus[2], D. De Santis[1,3]

1) Department of Clinical Immunology, PathWest, Fiona Stanley Hospital, Perth, Australia
2) Transplantation Immunology, Tissue Typing Laboratory, Maastricht University Medical Center, Maastricht, The Netherlands
3) UWA Medical School, The University of Western Australia, Perth, Australia

HLA. 2019 Oct 16. doi: 10.1111/tan.13729

Abstract

Abstract

The rapid progress of Human Leukocyte Antigen (HLA) typing techniques has contributed to improving the outcome of hematopoietic stem cell transplantation (HSCT). However, unambiguous HLA typing remains challenging. Next Generation Sequencing (NGS) has been shown to resolve the HLA typing ambiguity and simplify HLA typing workflows. The aim of this study is to develop a multiplexed full-gene PCR assay for eleven HLA loci that can be used on any NGS platform to provide additional information to the traditionally sequenced regions. The entire gene of HLA-A, HLA-B, HLA-C, DRB1, DRB3/4/5, DQB1, DQA1, DPB1 and DPA1 were amplified in four multiplexed reactions. A DNA reference panel of 47 samples representing the most common allele groups was selected to evaluate this novel assay using the Ion Torrent sequencing platform. The specificity and sensitivity of this assay was confirmed on additional 158 samples from a local Caucasian control cohort. Full gene sequences from start to stop codons including some UTR regions were obtained for all eleven HLA loci with complete gene coverage and sufficient read-depth for 3619 alleles. The whole amplicon was analysed for HLA class I genes, while only exons were analysed for class II genes. All alleles were amplified as expected with 100% concordance at full gene resolution for HLA class I and exon resolution for HLA class II loci when compared with previously used NGS or Sanger sequencing methods. In summary, the novel multiplexed PCR approach for full-gene HLA typing enabled for a large amount of genetic information to be generated in a simple and fast workflow.

1. Introduction

The human major histocompatibility complex (MHC) is known as the most complex region of the human genome. It contains several human leukocyte antigen (HLA) loci including the classical class I and class II genes namely HLA-A, HLA-B, HLA-C, DRB1, DRB3/4/5, DQB1, DQA1, DPB1 and DPA1 that play an important role in the adaptive immune responses.[1] The HLA genes encode for a series of highly polymorphic cell surface antigens and represents the fundamental ability of recognition of self-versus non-self through the presentation of class I and class II antigens to CD8+ T cells and CD4+ T cells, respectively.[2] In transplantation, this mechanism has to be modulated for successful engraftment.[3]

The allelic polymorphisms of classical HLA genes generally cluster at exons 2 and 3.[4] Variations in these exons that encode for the antigen recognition site (ARS) of HLA molecules can affect the specificity of antigen presentation and trigger T cell alloreactivity.[5] HLA disparity at the ARS induces allo-immune reactions by the engrafted donor T-cells recognizing the foreign host-derived antigens and escalating to the detrimental graft-versus-host disease (GVHD) or vice versa, the recipient immune system recognises foreign donor haematopoietic cells and leads to graft failure or rejection. Therefore, HLA genotype matching at the ARS between the donor and the recipient is critical to prevent GVHD, graft failure, rejection and to improve the overall outcome of haematopoietic stem cell transplantation (HSCT) as demonstrated by multiple large-scale studies.[4, 6-8] More recently, Mayor *et al* showed that better matching, found when typing is done at high resolution that includes exons outside the ARS, introns and untranslated regions, can significantly improve outcomes for recipients of HSCT and should be prospectively performed at donor selection.[9]

Polymorphisms are also present in other exons and regions of HLA genes, albeit at different frequency. HLA typing laboratories are obliged to distinguish null alleles such as A*24:09N, B*51:11N, C*04:09N, DRB4*01:03N and DRB5*01:08N, in which the differences to the common alleles reside outside of the ARS, for solid organ and haematopoietic transplant.[10] Furthermore, the complexity of HLA genes beyond the ARS has been shown to have an impact on the regulation of gene expression. Variants within the untranslated regions (UTRs) can regulate the level of surface expression. For example, variants in the 3'UTR of HLA-C can bind to a microRNA, has-miR-148, resulting in relatively low surface expression of alleles that bind this microRNA and high expression of HLA-C alleles that escape post-transcriptional regulation.[11] Similarly, polymorphism in the introns of HLA genes can affect the splicing process of messenger RNA (mRNA) and result in lowly expressed proteins. For example, HLA-A*24:02:01:02L allele carries a polymorphism in intron 2 position g708G>A that alters the splice site prior to exon 3 and results in low level of expression of this allele.[12] Information on the expression level of HLA proteins is

equitably important in the assessment of transplantation outcome since variations in the expression of HLA molecules can influence the strength of alloreactivity proportionally and consequently can result in direct implications for transplants with mismatched HLA genes. In 2015, Petersdorf *et al* reported an association between an expression marker, *rs9277534*, located in the 3′UTR of DPB1 and the risk of developing acute GVHD in patients who had high expression DPB1 alleles and received a DPB1 mismatched HSCT.[13] Therefore, it is necessary to detect single nucleotide polymorphism (SNP), insertion and deletions (indels) located in the expanded exons, intronic and regulatory regions of HLA genes in order to provide unambiguous information of the genomic constitution as well as gene expression.

Given 23,907 HLA alleles have been discovered since 1987 (IPD-IMGT/HLA database, release v3.37 http://www.ebi.ac.uk/imgt/hla),[14] the majority of these alleles are defined based on the sequence of exon 2 and 3 for HLA class I genes and exon 2 for HLA class II genes.[15] A HLA genotyping method that allows the full gene sequence to be determined is required to complete the reference database and to facilitate the unambiguous HLA matching of the donor and recipient for HSCT. In this study, we describe the development of 11 full gene-specific PCR assays (HLA-A, HLA-B, HLA-C, DRB1, DRB3/4/5, DQB1, DQA1, DPB1 and DPA1) that amplify the entire gene sequence contiguously from 5′ to 3′ UTRs of the genes. We define our method as full gene as it amplifies the protein coding regions of the genes (from start to stop codon) as well as all introns and parts of the untranslated regions.

2. Materials and methods

2.1 Sample selection and DNA extraction

To validate the primer specificity, a DNA reference panel of 47 samples representing the common HLA antigen groups was selected. All samples were previously well-characterised in-house by either a long range PCR assay or the commercial AllType*TM* Next Generation Sequencing (NGS) based typing method on the Ion Torrent platform. Subsequently, 158 DNA control samples from the Western Australian Busselton cohort[16], who are predominantly of Caucasian ethnicity with previously reported HLA genotypes, were included in this study to confirm primer specificity and sensitivity. DPB1 alleles were selected based on those most commonly seen and available in our local population.

All genomic DNA was extracted from peripheral white blood cells or B-cell transformed cell lines using the DNA Midi Kit (Qiagen, Germany) on the QIAsymphony SP instrument according to the vendor's protocol. The concentration and purity of extracted DNA were assessed by the optical density (OD) 260/280 ratio of 1.8-2.0. Samples were then normalized to a concentration of 25ng/mL.

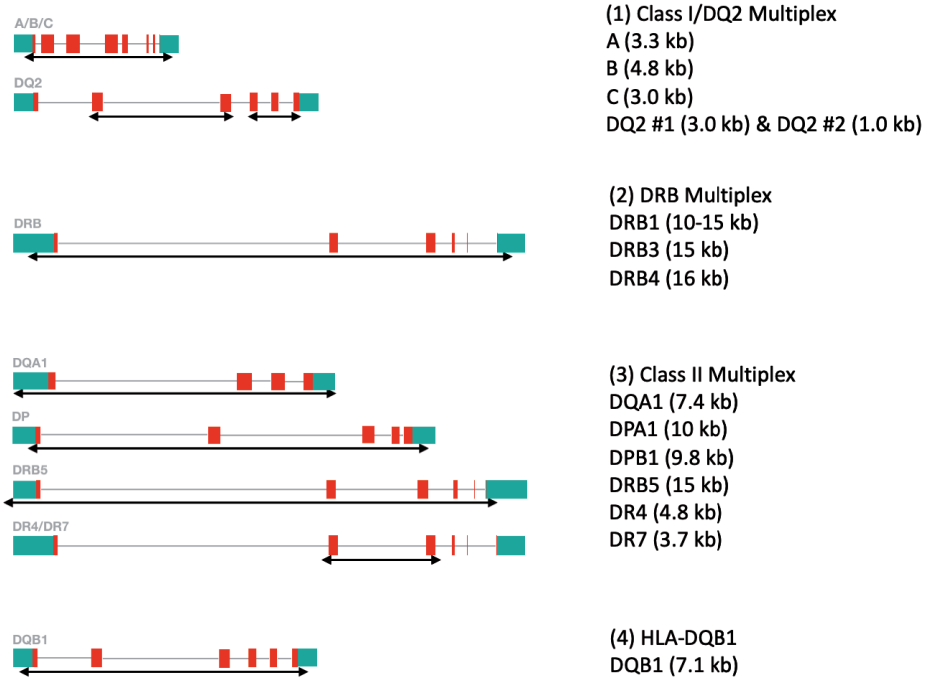


Figure 1. Long-range amplification primer locations for contiguous amplification of HLA genes.
The green and red shadings represent untranslated and coding regions, respectively. The amplified products are depicted by the arrow black lines

5

2.2 Primer design

Eight locus-specific PCR primer pairs for HLA-A, HLA-B, HLA-C, DRB1, DRB3/4/5 and DPB1 were designed to generate full-length gene amplicons from 5'UTR to 3'UTR using the genomic references from notable databases such as GenBank, dbSNP, the 1000 genomes project and IPD-IMGT/HLA databases. DQB1, DQA1 and DPA1 amplification primers were adapted from previous studies.[15, 17]

Two specific long-range PCR primer pairs for DQB1*02 alleles were designed to increase the coverage spanning from exon 2 to exon 3 (DQB1*02 pair #1) and exon 4 to exon 6 (DQB1*02 pair #2) in heterozygous samples. In addition, one specific PCR primer pair for DRB1*04 and DRB1*07 alleles spanning from exon 2 to exon 3 was also designed to address issues of preferential amplification against these alleles in heterozygous individuals. The primer sequences and location of HLA-A, HLA-B, HLA-C, DRB1, DRB3/4/5, DPB1, DQB1*02, DRB1*04, DRB1*07 are available upon request to the corresponding author under a non-disclosure agreement. A schematic representation of the gene coverage is shown in Figure 1.

2.3 Polymerase chain reaction (PCR) for the full gene method

The eleven genes were amplified in four multiplexed reactions. For the amplification of the HLA class I genes and sub-segments of the DQB1*02 gene, a multiplexed PCR method was applied. The sub-segment amplification of DQB1*02 allele was kept separate from the full gene DQB1 PCR to prevent formation of non-specific products in the reaction. From here on, this multiplexed amplification is known as 'Class I/DQ2 multiplex'. The amplification of the eight HLA class II genes was performed in three separate reactions. The first included a single-locus amplification of DQB1, the second, a multiplex reaction that includes DRB1, DRB3 and DRB4 genes (DRB multiplex) and the final reaction, a multiplex PCR that includes the sub-segment of DRB1*04, DRB1*07 alleles, and full gene products of DQA1, DPA1, DPB1, DRB5 genes (Class II multiplex).

2.3.1 Class I/DQ2 multiplexed amplification:

HLA-A, HLA-B, HLA-C and DQB1*02 genes were amplified in 25mL reaction volume consisting of 50ng of genomic DNA, 1.25U of the high fidelity PrimeSTAR GXL DNA polymerase (Takara Bio Inc, Shiga, Japan), 1X PrimeSTAR GXL buffer (Takara Bio Inc, Shiga, Japan), 200mM dNTPs mix (Takara Bio Inc, Shiga, Japan), 0.1mM of HLA-A primers, 1mM of HLA-B primers, 0.1mM of HLA-C, and 0.2mM of DQB1*02 primers. Amplification was performed on a Eppendorf Mastercycler Pro (ThermoFisher Scientific, Massachusetts,

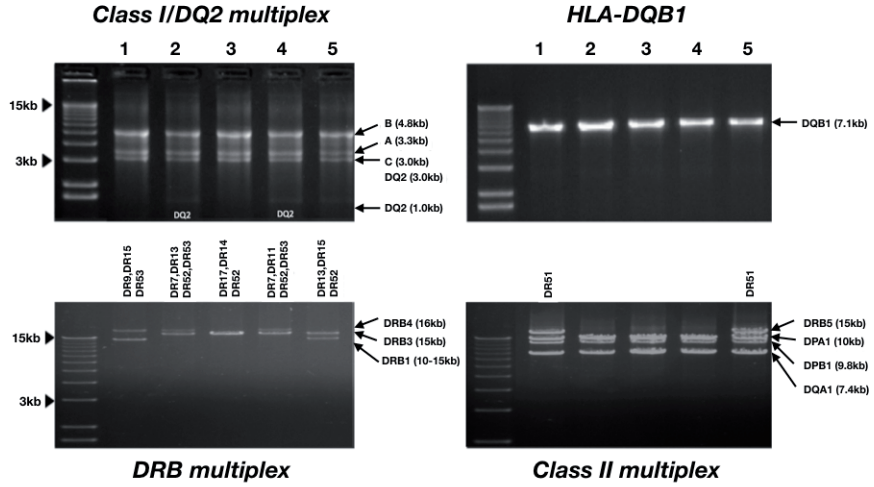


Figure 2. Electrophoresis image on 0.7% gel of amplified products from five DNA samples using locus specific primers. Number 1 to 5 above the lanes represent samples R97-0900270, R05-0114117, Q94-0050546, Q94-0053272, Q94-0055912, respectively. The first lane in each gel presents bands of the 1 kb plus DNA ladder. The size of the bands and the corresponding HLA loci are indicated on the right side of the figure. The 1 kb band in class I/DQ2 multiplex was designed to be at low concentration, therefore, the band was not clearly visible on the agarose gel

USA) using the following rapid 2-step PCR cycling conditions: primary denaturation at 94°C for 2 min, followed by 30 cycles of 98°C for 10 s and 68°C for 3 min.

2.3.2 DQB1 amplification:

DQB1 was amplified in 25mL reaction volume consisting of 50ng of genomic DNA, 1.25U of PrimeSTAR GXL DNA polymerase (Takara Bio Inc, Shiga, Japan), 1X PrimeSTAR GXL buffer (Takara Bio Inc, Shiga, Japan), 200mM dNTPs mix (Takara Bio Inc, Shiga, Japan), and 0.2mM of each primer. Amplification was performed on the Veriti thermal cycler (ThermoFisher Scientific, Massachusetts, USA) with adjusted ramp speed to simulate the GeneAmp 9600 thermal cycler as follows: primary denaturation at 94°C for 2 min, followed by 30 cycles of 98°C for 10 s and 68°C for 3 min.

2.3.3 DRB multiplexed amplification:

DRB1/DRB3/DRB4 were amplified in 25mL reaction volume consisting of 50ng of genomic DNA, 1.25U of PrimeSTAR GXL DNA polymerase (Takara Bio Inc, Shiga, Japan), 1X PrimeSTAR GXL buffer (Takara Bio Inc, Shiga, Japan), 200mM dNTPs mix (Takara Bio Inc, Shiga, Japan), and 0.2 – 0.4mM of each primer. Amplification was performed on a Eppendorf Mastercycler Pro (ThermoFisher Scientific, Massachusetts, USA) using the following rapid 2-step PCR cycling conditions; primary denaturation at 94°C for 2 min, followed by 30 cycles of 98°C for 10 s and 68°C for 4 min.

2.3.4 Class II multiplexed amplification:

In the Class II multiplex reaction, 50ng of genomic DNA was amplified with the GoTaq Long PCR Master Mix (Promega, Wisconsin, USA) and 0.02 – 0.2mM of each primer pair. Amplification was performed on a Eppendorf Mastercycler Pro (ThermoFisher Scientific, Massachusetts, USA) using the following PCR cycling conditions; primary denaturation at 94°C for 2 min, followed by 30 cycles of 94°C for 30 s, 60°C for 30 s, 68°C for 9 min, and a final extension at 72°C for 10 min.

The amplification products were quantitated by loading 2mL of the PCR product on a 0.7% agarose gel. Once the positive bands of the correct size were confirmed, the amplified products were subsequently pooled at a fixed-volume ratio (determined empirically) (Class I/DQ2 multiplex [20ml], DQB1 [23ml], DRB multiplex [23ml], and Class II multiplex [10ml]) prior to PCR purification. The amplicon pool was then purified using 0.6X Agencourt AMPure beads (Beckman Coulter, USA) on the automated liquid handler Microlab STAR Line (Hamilton, Nevada, USA) and eluted in a final volume of the 80ml which is sufficient to perform library preparation twice if required. The purified amplicon pool was then ready for library preparation for Next-Generation Sequencing (NGS) on the Ion Torrent platform.

2.4 AllType*TM* PCR method

The 47 samples from the DNA panel were genotyped for all 11 loci using the One Lambda AllType*TM* protocol in a single multiplexed reaction. The amplification was performed as per manufacturer's recommendation. The amplified products were purified using 0.6X Agencourt AMPure beads (Beckman Coulter, USA) on the automated liquid handler Microlab STAR Line (Hamilton, Nevada, USA) and eluted in 25ml of molecular grade water (Sigma-Aldrich, Missouri, USA). The purified AllType*TM* products were then ready for library preparation and Ion Torrent sequencing.

2.5 In-house long-range PCR method

The 158 samples from the WA Busselton cohort were HLA typed for HLA-A, HLA-B, HLA-C, DRB1, DPB1, DQB1, DQA1 and DPA1 using the in-house long range PCR on the Ion Torrent NGS platform. DRB3, DRB4 and DRB5 loci for this cohort were typed using an in-house developed PCR and Sanger Sequencing-Based Typing (SBT).

The in-house long-range PCR amplified the promoter to 3'UTR of class I genes, DQA1 and DPA1, and the ARS (exon 2 to exon 3) of DRB1, DPB1 and DQB1. Briefly, HLA-A, HLA-B, and HLA-C were amplified in separate PCR in 25mL reaction volume consisting of 50ng of genomic DNA, 1U of Elongase enzyme Mix (Invitrogen, California, USA), 5X buffer B (Invitrogen, California, USA), 200mM dNTPs mix (Invitrogen, California, USA), 1mM of $MgCl_2$, 5% DMSO (Sigma Aldrich, Missouri, USA), and 0.2mM of HLA-A primers, 0.2mM

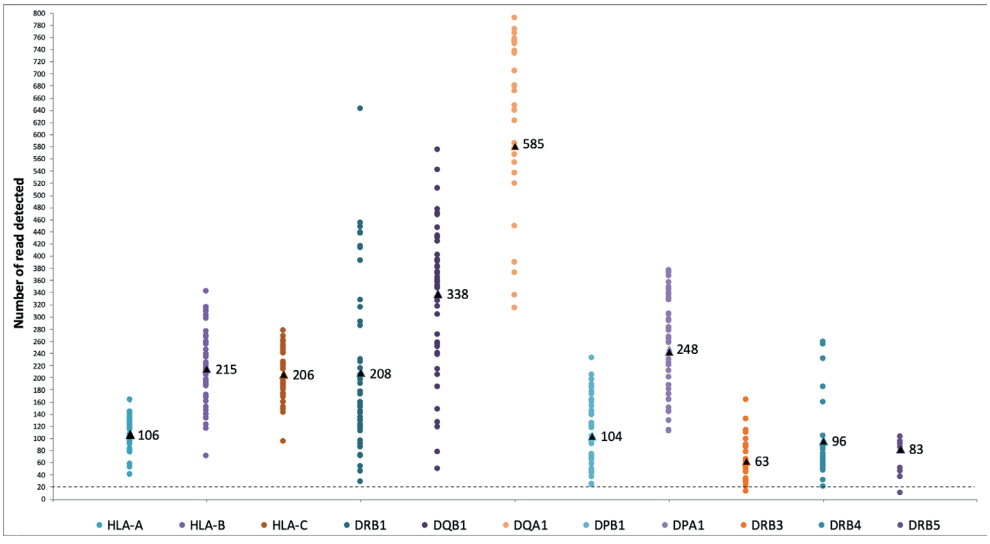


Figure 3 The minimum read-depth detected across the amplicon in each HLA locus. The round and triangle dots specify the minimum read depth in each DNA sample and the average depth, respectively. The black line represents the threshold of 20 reads

of HLA-B primers, and 0.2mM of HLA-C, respectively. The PCR cycling conditions were as follows: primary denaturation at 94°C for 2 min, followed by 30 cycles of 94°C for 30 s, 62°C for 30 s, and 68°C for 3 min.

For single-locus amplification of DQB1, DQA1 and DPA1 gene, the target amplicon was amplified in 25mL reaction volume consisting of 50ng of genomic DNA, 1.25U of PrimeSTAR GXL DNA polymerase (Takara Bio Inc, Shiga, Japan), 1X PrimeSTAR GXL buffer Takara Bio Inc, Shiga, Japan), 200mM dNTPs mix Takara Bio Inc, Shiga, Japan), and 0.2mM of each primer. The PCR cycling conditions were as follows: primary denaturation at 94°C for 2 min, followed by 30 cycles of 98°C for 10 s, 62°C for 30 s and 68°C for 3 min.

Lastly, for singleplex amplification of DRB1 and DPB1, 50ng of genomic DNA was amplified with the GoTaq Long PCR Master Mix (Promega, Wisconsin, USA) and 0.2mM of each primer pair. The PCR cycling conditions were as follows: primary denaturation at 94°C for 2 min, followed by 30 cycles of 94°C for 30 s, 60°C for 30 s, 68°C for 3 min, and a final extension at 72°C for 10 min. The amplified products were subsequently pooled into a single tube, ready for library preparation and Ion Torrent sequencing.

2.6 Library preparation for Ion Torrent Sequencing

The library preparation workflow was identical for all HLA amplicons generated by either, the novel full gene, AllType*TM* or in-house long range PCR methods. Library preparation was adapted from the manufacture's protocol for Ion Xpress Plus Fragment Library Kit and Ion Xpress Barcode Adapters Kit (ThermoFisher Scientific, Massachusetts, USA) to be fully automated on the liquid handler Microlab STAR Line (Hamilton, Nevada, USA) and to reduce total library preparation time.

Briefly, 35ml of the amplicon pool was enzymatically sheared, ligated with the dual adapters containing unique indexed sequence on one end and a P1 adapter on the other end, and size-selected using a dual bead-based protocol. The fragment size and concentration of each library was quantitated using the LABChip DNA High Sensitivity assay (PerkinElmer, Massachusetts, USA). Each barcoded library was then normalized and subsequently pooled at equimolar concentration into a single tube. A maximum of 48 libraries were pooled for Ion Torrent sequencing to ensure a minimal of 20X coverage across approximately 90 kb of nucleotide bases per sample. In the secondary amplification, the barcoded library pool was amplified using primers complementary to ligated adaptors and KAPA HiFi HotStar ReadyMix (Roche, Basel, Switzerland) for 8 cycles. The concentration and size of the multiplexed library pool was quantitated using the QuBit dsDNA High Sensitivity assay (ThermoFisher Scientific, Massachusetts, USA) and diluted to 130pmol/ml for template preparation on the Ion Chef System (ThermoFisher Scientific, Massachusetts, USA). The template preparation for 400 to 600 base-pair libraries

was fully automated on the Ion Chef system using Ion 520 & Ion 530 ExT kit (ThermoFisher Scientific, Massachusetts, USA) and sequencing was performed on the Ion GeneStudio S5XL system using Ion S5 EXT Sequencing kit on Ion 530 chip (ThermoFisher Scientific, Massachusetts, USA).

2.7 Data analysis and HLA allele assignment

The signal processing, base calling, trimming and de-multiplexing of raw data were performed by the Torrent Suite 5.10.1 (ThermoFisher Scientific). The quality-filtered sequences are sorted according to the index sequence of Ion Xpress barcodes and FASTQ files are generated ready for HLA allele analysis.

The HLA allele assignment was performed using GenDX NGSengine software version 2.13 with IMGT/HLA database version 3.35.0 (GenDx, Utrecht, The Netherlands). The amplicon sequence was analysed for HLA class I genes, while only the exons were analysed for class II genes. Intronic analysis of class II genes was also attempted using NGSengine (GenDx, Utrecht, The Netherlands), however, hindered by various issues. Firstly, the analysis of intron 1 and intron 2 of DRB1 by the NGSengine software resulted in several thousand base-calling errors due to the mis-mapping of short reads in highly repetitive regions which resulted in the requirement of a large number of manual edits to rectify. This amount of manual curation was laborious and significantly increased the analysis time of the DRB1 gene. Additionally, the inclusion of DPB1 introns in some cases resulted in incorrect assignment of phase by the NGSengine software due to the lack of reference sequence, and the inability to link polymorphisms with short reads. The software seemed to favour alleles that were found to be the best match regardless of the length of reference sequence available. The sequencing of these samples with long read chemistry confirmed that the allele call provided using the short read technology was in fact incorrectly phased. HLA data from all methods were stored in the laboratory information management system of the Department of Clinical Immunology according to the standard operating procedures and available for genotype concordance comparison. HLA allele concordance results were compared manually.

2.8 Assessment of PCR and Sequencing quality

The full gene assay was validated against the following PCR acceptance criteria: i) the amplicon was of correct band size within ± 10% of the theoretical size of the amplicon, ii) the absence of HLA allele dropout, iii) allelic balance should be between 20% to 80%, but not less than 10%, iv) the results obtained with the novel full gene assay must provide equal or higher resolution typing than the current typing assay in 100% samples tested, v) the full gene assay should be sufficiently robust to eliminate any repeat testing of routine samples, therefore, the failure rate should be less than 5%.

The sequencing results of an Ion 530 chip on the Ion S5/XL system were also assessed against established acceptance criteria: i) the number of reads and total bases should be between > 9 million reads and > 1.5 Gbases, respectively, ii) the minimum loading density of ISP is > 60%, poor loading results in reduced sequencing data output, iii) the mean and mode of read-length is 260bp ± 10% and 350bp ± 10%, shorter read lengths may result in a higher number of genotyping ambiguities due to the inability to resolve cis/trans ambiguities, iv) polyclonality is < 40%, high number of polyclonal ISPs will reduce the sequencing data output, iv) primer dimer sequence < 2%, which ensures that there is efficient removal of small fragments, v) enrichment percentage and key signal score have to meet a minimum of 99% and 40, respectively.

3. Results

3.1 HLA full gene amplification

After the optimal PCR conditions (such as annealing temperature of primers, concentration of Mg2+ and dNTPs, cycling conditions with associated ramp speed) were determined, specific full-gene amplified products of HLA genes were found in all tested samples. The end result was three multiplex and one single locus PCR reaction that amplified all eleven HLA classical genes with specificity and sensitivity.

In the class I/DQ2 multiplex, three bands were observed at approximately 3 kb, 3.3 kb and 4.8 kb, which were concordant with the theoretical fragment sizes according to the genomic reference sequence and IPD-IMGT/HLA database for HLA-C, HLA-A and HLA-B, respectively. The specific enriched sub-segment of the DQB1*02 allele spanning from exon 2 to exon 3 overlapped with the full-length amplified products of HLA-C in size at 3 kb. Therefore, the two bands could not be distinguishable using gel agarose electrophoresis. The sub-segment of DQB1*02 allele spanning from exon 4 to exon 6 was designed to be at low concentration to prevent over representation of this fragment and therefore the band was not clearly visible on the agarose gel (Figure 2).

A single band at 7.1 kb was detected in the single gene amplification for DQB1 locus. The PCR products from DRB1 gene varied in size depending on the DRB1 sub-type ranging from 10 kb to 15 kb, while DRB3 and DRB4 genes were detected at 15 kb and 16 kb, respectively. In the class II multiplex, three to four bands were observed at 7.4 kb (DQA1), 9.8 kb (DPB1), 10 kb (DPA1) and 15 kb (DRB5), the presence of the DRB5 gene is dependent on the DRB haplotype of the sample (Figure 2).

As part of the optimisation for DRB1, additional specific primer pairs for DRB1*04/07 was included in the class II multiplex reaction to increase the coverage over exon 2 and

3 for the two allele groups. The primer mixture of specific primer pairs for DRB1*04/07 and locus specific primer pairs for DQA1/DPA1/DPB1/DRB5 was titrated empirically to ensure adequate representation of all products and to avoid PCR bias towards the smaller amplicon. The PCR products of the DRB1*04 and DRB1*07 specific primer pairs were observed at 4.8 kb and 3.7 kb, respectively, in the presence of the other HLA class II genes (See Figure 1 in the supplementary data).

3.2 HLA sequencing on the Ion S5 system

The full gene sequencing of eleven classical HLA loci was performed for 47 samples in a single run on the Ion GeneStudio S5 system. The assessment of sequencing metrics against the NGS acceptance criteria are shown in Table 1. A total of 14.7 million reads and 4.43 Gigabases of sequencing data were available for downstream analysis following quality filtering. The number of reads per sample ranged from 150,271 to 446,303 and averaged at 300,954 ± 57537 reads. Overall, all sequencing metrics met the acceptance criteria, except for the polyclonality filter which was higher than the threshold of 40%. High polyclonality reduces the useable data, however the number of usable reads was above the acceptance criteria for an Ion 530 chip and therefore the sequencing run was suitable for HLA allele analysis.

The analysis of sequence read length distribution from a representative data file was performed using FastQC tool (Babraham Bioinformatics, U.K.) in order to determine the proportion of 400 to 600 base-read fragments in Ion Torrent sequencing as per specification by the manufacturer. There was approximately 20% of all detected reads longer than 400 base pairs and ranging up to 700 base pairs (44,707 of 216,639 sequence reads). Sequence reads ranging between 300 to 400 base pairs made up 37% (80,591 of 216,639) of all the reads and the mode of all sequences were detected at 360 base pairs in length (35,181 of 216,639 reads).

Sequencing Criteria	Acceptance limit for Ion 530 chip	Sequencing Run	Pass limit
Number of reads	9 – 12 million sequence reads	14.7 million reads	Yes
Total bases	1.5 – 4.5 Gbases	4.43 Gbases	Yes
ISP loading %	> 60%	91%	Yes
Polyclonality %	< 40%	41.9%	No
Read-length	Mean of 260 bp ± 10% Mode of 350 bp ± 10%	Mean of 301 bp Mode of 360 bp	Yes Yes
Low Quality Filter	< 25%	14.8%	Yes
Enrichment	[3] 99%	100%	Yes
Key Signal	> 40	76	Yes
Internal control	Correct allele assigned	Yes	Yes

Table 1: Assessment of the sequencing metrics against the acceptance limit for an Ion GeneStudio S5XL run using an Ion 530 chip.

3.3 Evaluation of multiplex full gene PCR method

3.3.1 Minimal read-depth assessment

To evaluate the novel full gene PCR method, we assessed the minimal read-depth, and allelic balance in all eleven HLA loci derived from the mappable sequence reads defined by NGSengine software. A minimal read-depth of 20 reads was assigned for all loci. Of the 824 HLA sequenced alleles (47 samples x 11 loci), insufficient coverage (less than 20 reads in the ARS) excluded allele assignment for one DRB3*01:01:02 and one DRB5*02:02 allele in sample Q94-0056288 and Q95-0051198, respectively (Figure 3). Overall, the failure rate for not meeting the minimal read-depth was 0.24% (2 of 824).

On further investigation, the poor read coverage in exon 2 of DRB3 and DRB5 genes was absent when reads from the FASTQ files of Q94-0056288 and Q95-0051198 were mapped to the human genome reference sequence using LAST aligner for parallel confirmation (Figure 2 in supplementary data). The minimal depth measured in exon 2 for DRB3*01:01:02 and DRB5*02:02 allele was 64 and 105, respectively. This indicated that the discrepancy in the allele analysis was not related to the efficiency of PCR amplification of these two alleles, but due to the mapping algorithms used to map and align the sequence reads in NGSengine.

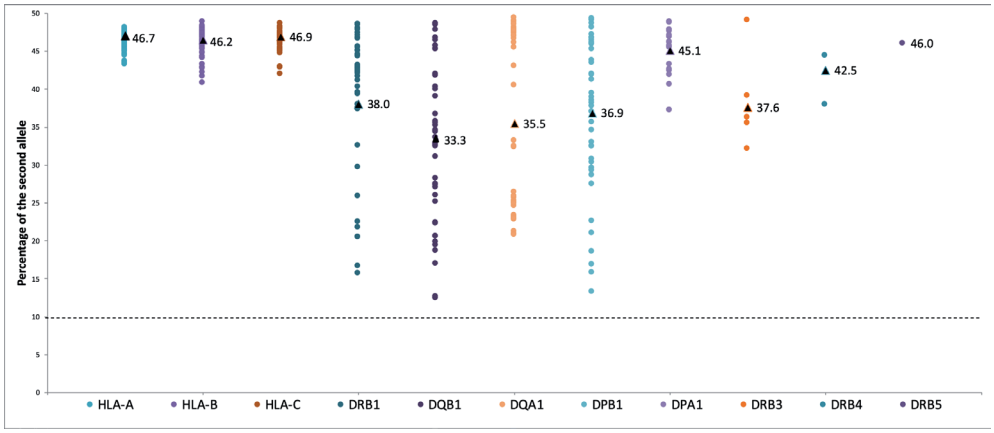


Figure 4. The allelic balance in heterozygous HLA loci defined by NGSengine software. The round and triangle dots specify the percentage of the second allele detected in each DNA sample and the average depth, respectively. The black line represents the threshold of 10%. There was no average data for heterozygosity in HLA-DRB5 as there was only one heterozygous HLA-DRB5 individual in the DNA reference panel

3.3.2 Assessment of allelic balance

There was no evidence of allele dropout in any of the tested samples. To validate that the HLA alleles from both chromosomes were amplified uniformly, the estimated percentage of the minor allele across the whole gene in a heterozygous sample was calculated by the NGSengine software and illustrated in Figure 4. Of the 448 results for HLA class I genes, DPA1, DRB3/4/5 loci, the minor or second allele was detected at a minimum of 35% across all heterozygous positions, suggesting that both of the HLA alleles were amplified and represented sufficiently for accurate HLA typing. The allelic balance for the amplification of DQA1 was above 15% in 188 results. On the other hand, there was evidence of allelic imbalance in DRB1, DQB1 and DPB1 genes, however, all samples met the pre-determined threshold of 10% using additional locus specific primers in the amplification mix.

3.4 HLA genotyping resolution of the evaluation DNA panel

The previous genotyping results for HLA-A, HLA-B, HLA-C, DRB1, DRB3/4/5, DQB1, DQA1, DPB1 and DPA1 obtained using the AllType*TM* NGS typing method were available for comparison with the novel full gene PCR assay. As a result of differing typing resolution due to the amplicon coverage, concordance was determine by defining whether the reported allele was within the string of alleles of the ambiguous result.

Of the 47 samples in the DNA panel, the novel full gene typing results were 100% concordant to the 3*rd* field for all HLA loci compared with the AllType*TM* method. With the addition of exon 1 in all class II genes, the full-gene assay provided higher resolution than the AllType*TM* method for DRB1 (5 of 94), DQB1 (49 of 94) and DPB1 (6 of 94) loci (Table 2 and 4). Ambiguous results were observed in 2/94 (2.1%) DQB1 and 24/94 (25.5%) DPB1 alleles, using the novel full gene method, due to the inability to resolve phase ambiguities with short read technology in heterozygous samples (Table 3 and 4). Additionally, the novel full-gene assay is unable to amplify the DQB1*03:276N allele, an allele that contains a 3.7kbp deletion upstream of intron 1 which is the location of the 5′ DQB1 primer binding site. Therefore, in an individual who is homozygous for DQB1*03:01:01, it is not possible to exclude the presence of DQB1*03:276N on the other chromosome. As the result, one DQB1 result was reported with allele ambiguity of DQB1*03:01:01/03:276N (Table 3 and 4).

At the 4*th* field resolution, the full gene assay provided equal resolution for HLA-A and HLA-B compared to AllType*TM* PCR method. There were 24 of 94 HLA-C results with intronic ambiguities, which were fully resolved by AllType*TM* amplicons, due to the presence of polymorphisms in the UTR that are not covered by the full gene assay (See Table 1 in the supplementary data). The primer design of HLA-B was extended to include 1500bp downstream of the stop codon to resolve the 4*th* field resolution of alleles such as B*51:01:01 and B*40:01:02. Unfortunately, the NGSengine software does not allow the alignment of the 3′ end of HLA-B genes resulting in ambiguous typing results for these

alleles. For this evaluation, these alleles were resolved through manual curation using LAST aligner and IGV software.

In summary, the novel full gene HLA genotyping assay provided equal resolution in the 3*rd* field for class I genes compared to the AllType*TM* method and better resolution for class II genes due to the addition of exon 1 of HLA genes. However, the presence of DQB1*03:276N allele cannot be excluded in individuals with only one copy of DQB1*03:01:01 allele and there are still several HLA-C alleles that cannot be resolved to 4*th* field resolution due to polymorphisms in the UTR regions outside of the region amplified.

3.5 HLA genotyping resolution of the WA Busselton control cohort

To confirm primer specificity and sensitivity, 158 samples of the WA Busselton control[16] cohort were tested. A total of 2795 alleles (158 individuals x 8 loci x 2 alleles, 120 DRB3 alleles, 105 DRB4 alleles and 42 DRB5 alleles) were identified and compared to genotyping results obtained with the Sanger SBT method (for DRB3/4/5) or in-house long range NGS method (for all other HLA loci). The novel full gene and the previous in-house methods were 100% concordant to the 3*rd* field for all HLA class I genes, 99.7% for DRB1 (315 of 316 results), 99.7% for DQB1 (315 of 316 results), 98.1% for DPB1 (310 of 316 results), 95.2% for DRB5 (40 of 42 results) and 100% for all other loci. The disparity at 3*rd* field were due to multiple reasons including: (1) novel polymorphism detected in the additional exons, which were not typed in the former method, and resulted in different allele name or (2) homozygous genotype in the ARS but heterozygous type in other exons (See Table 2 in the supplementary data).

As observed previously in the panel validation, the full gene PCR method provided significantly higher resolution compared to the partial gene typing method. The full gene assay provided higher resolution than the previous partial gene in-house typing method by 8.2% for DRB1 (26 of 316), 49.7% for DQB1 (157 of 316), 17.7% for DPB1 (56 of 316), 66.7% for DRB3 (80 of 120), 100% for DRB4 (105 of 105) and 83.3% for DRB5 (35 of 42) due to the extension beyond the ARS for all class II genes (Table 2). However, there were still several DQB1and DPB1 results that could not be resolved to the allelic level due to the DQB1*03:01:01/03:276N ambiguity and the inability to phase heterozygous DPB1 alleles as mentioned previously (Table 3 and 4).

Locus	Ambiguity	Full gene result	Number of alleles (DNA reference panel)	Number of alleles (WA Busselton panel)	Comment
DRB1	04:07:01 / 04:92	04:07:01	0	6	Resolved at codon 207 in exon 4
	03:01:01 / 03:01:08	03:01:01	0	1	Resolved at codon 112 in exon 3
	09:01:02 / 09:31	09:01:02	1	7	Resolved at codon (-21) in exon 1
	12:01:01 / 12:10	12:01:01	2	9	Resolved at codon (-16) in exon 1
	15:02:01/15:140/ 15:149	15:02:01	2	3	Resolved at codon (-2) and (2) in exon 1
DQB1	02:02:01 / 02:02:06	02:02:01	11	22	Resolved at codon (-21) in exon 1
	03:01:01/ 03:01:41/03:01:43/ 03:276N/ 03:297	03:01:01	16	65	Resolved by the presence of exon 1
	03:02:01 / 03:289	03:02:01	4	30	Resolved at codon (-12) in exon 1
	03:02:01 / 03:289	03:289	1	2	Resolved at codon (-12) in exon 1
	05:01:01 / 05:01:24	05:01:01	11	35	Resolved at codon (-6) in exon 1
	06:01:01 / 06:01:15	06:01:01	6	3	Resolved at codon (-11) in exon 1
DPB1	01:01:01 / 417:01	01:01:01	0	18	Resolved at codon 194 in exon 4
	03:01:01 / 104:01:01	03:01:01	0	12	Resolved at codon 205 in exon 4
	03:01:01 / 104:01:01	104:01:01	0	3	Resolved at codon 205 in exon 4
	05:01:01 / 135:01	05:01:01	0	6	Resolved at codon 205 in exon 4
	105:01:01 / 665:01	105:01:01	4	4	Resolved at codon (-22) and (-14) in exon 1
	13:01:01 / 107:01	13:01:01	2	13	Resolved at codon (-22) and (-14) in exon 1

Table 2: HLA ambiguity resolved by the full gene assay but not the AllType*TM* or the in-house long-range assay.

Locus	Ambiguity cannot be resolved	Number of alleles (DNA reference panel)	Number of alleles (Busselton Panel)	Comment
DQB1*	03:01/03:276N	1	6	The presence of DQB1*03:276N cannot be excluded in this sample, which only had one copy of DQB1*03:01:01 detected.
	06:02/06:84, 06:04/06:39	2	0	Unable to determine phase due to short read-length
DPB1*	04:01/702:01, 104:01/124:01	4	4	Unable to determine phase due to short read-length
	03:01/351:01, 04:02/463:01	8	8	Unable to determine phase due to short read-length
	03:01/124:01, 04:01/350:01	2	22	Unable to determine phase due to short read-length
	01:01/162:01, 02:01/461:01	2	4	Unable to determine phase due to short read-length
	02:01/416:01, 04:02/105:01	4	6	Unable to determine phase due to short read-length
	04:01/126:01, 04:02/105:01	2	24	Unable to determine phase due to short read-length
	02:01:02/414:01, 19:01/106:01	2	0	Unable to determine phase due to short read-length
	03:01/104:01, 05:01/135:01	0	4	Unable to determine phase due to short read-length

Table 3: Allelic ambiguities that cannot be resolved by the novel full-gene PCR method.

	DNA Reference Panel			Busselton Panel		
Locus	**Number of alleles**	**AlltypeTM**	**Full-gene**	**Number of alleles**	**In-house**	**Full-gene**
A	94	100% (94/94)	100% (94/94)	316	100% (316/316)	100% (316/316)
B	94	100% (94/94)	100% (94/94)	316	100% (316/316)	100% (316/316)
C	94	100% (94/94)	100% (94/94)	316	99% (314/316)	100% (316/316)
DRB1	94	95% (89/94)	100% (94/94)	316	92% (290/316)	100% (316/316)
DRB3	34	100% (34/34)	100% (34/34)	120	33% (40/120)	100% (120/120)
DRB4	28	100% (28/28)	100% (28/28)	105	0% (0/105)	100% (105/105)
DRB5	10	100% (10/10)	100% (10/10)	42	16% (7/42)	100% (42/42)
DQB1	94	45% (42/94)	97% (91/94)	316	48% (153/316)	98% (310/316)
DQA1	94	100% (94/94)	100% (94/94)	316	100% (316/316)	100% (316/316)
DPB1	94	68% (64/94)	74% (70/94)	316	60% (190/316)	78% (246/316)
DPA1	94	100% (94/94)	100% (94/94)	316	100% (316/316)	100% (316/316)

Table 4. HLA genotype results that are resolved to the third-field resolution by the full-gene PCR assay compared with AllType and in-house long-ranged method

4. Discussion

The main goal of this study was to develop a method that could provide full-length gene characterisation for all eleven HLA genes without sacrificing the processing time or typing cost for high-resolution HLA results. This full gene PCR method enables the sequencing of the protein coding region of the genes including all exons, introns and some of the UTR regions. There are several NGS based HLA typing methods described in the current literature however these methods amplify exon 2 and 3 of the class II genes mostly and exclude other parts of the genes. These approaches lead to ambiguous results and potentially omit the identification of novel nucleotides that are located outside of the ARS including those that could result in a non-expressed allele[18-21]. Furthermore, the full gene assay described here is the first to include full length gene amplification for all 11 loci including the lower expression genes such as DRB3/4/5, DQA1 and DPA1 in multiplexed reactions whereby all 11 loci are amplified in only 4 amplification reactions. Other full gene methods[15, 22, 23] amplify each of the HLA genes separately, requiring more reactions per individual and therefore increasing the HLA typing cost, and downstream processing time. The described NGS based method using contiguous multiplexed full-length PCR products can therefore circumvent these issues.

The current study allows for a large amount of genetic information in a streamlined workflow. The entire workflow from DNA extraction, PCR set-up, PCR purification, quantitation, library preparation, template preparation and sequencing was fully automated on robotic liquid handling systems with minimal hand-on requirement. From

the extraction of genomic DNA to the completion of sequencing analysis of up to 48 samples, the complete workflow was performed in 4-days, which is reasonable processing time to generate high-resolution allele definition for clinical HLA genotyping. Together with the multiplexed reaction approach, the amplicons were pooled prior to purification using SPRI AMPure beads in order to simplify the workflow and reagent cost. Lastly, allele assignment using the NGSengine software for 48 samples can be completed in 3 hours with minimal manual manipulation.

Validation of the novel full-gene HLA genotyping assay was conducted against several criteria from specificity of PCR products, sufficient sequence read-depth, allelic balance and genotype resolution. Specific amplified products were observed on the electrophoresis image in all four reactions. It is also important to note that the DNA quality and purity are crucially important in long ranged PCR as the amplified products can be as large as 16 kb (DRB4) and therefore required auxiliary care and investment in a high quality DNA extraction method. Sufficient read-depth was obtained to accurately assign HLA alleles in all 47 samples of the reference panel. The low coverage in the 3′ end of exon 3 in DRB3, and the 3′UTR in HLA-B was not related to the efficiency of long-range PCR, but due to the limitation in alignment algorithm of GenDx NGSengine. This highlights that in addition to optimising genotyping method, it is extremely crucial to apply high stringency for bioinformatics analysis tool to ensure accurate HLA genotypes.

There was no allele dropout observed in this study. However, there was evidence of preferential amplification against DQB1*02, DRB1*04 and DRB1*07 alleles in the presence of DQB1*06 or DRB1*01, respectively. This issue was rectified by the addition of specific allele group amplification of DQB1*02, DRB1*04 and DRB1*07 in which the amplified products spanning from exon 2 to exon 3 of the genes. The predicament of preferential amplification observed in DRB1 gene could potentially due to the limitation of multiplexed PCR in amplifying multiple targets at different size in a single reaction. According to the IPD-IMGT/HLA database, DRB1*07:01:01 allele can extend more than 16,000 bp in length compared to DRB1*01:01:01 allele which is approximately 11,000 bp in length. There is a noticeable difference in length between the two alleles. Physical properties of target sequence such as target length, GC content or secondary structures were shown to inherently influence the PCR bias and result in preferential amplification as reported in previous studies. [24, 25]

The advantages of full-length genotyping method were highlighted by the identification of novel positions in the exons and regions that were not included in the current partial gene typing methods. The full gene typing assay had 100% concordant to the the 3*rd* field for all class I genes compared to the AllType*TM* and in-house long range NGS typing method, and significantly higher resolution for several class II genes with the addition of exon 1.

Furthermore, three novel alleles were identified including one DQB1*03, one DPA1*02 and one DRB3*02 in which the novel alleles differ in one nucleotide position in exon 1 of DQB1*03:01:01, exon 4 of DPA1*02:01:01 and exon 1 of DRB3*02:02:01, respectively. This full gene assay therefore provides an opportunity for HLA typing laboratories to complete the IPD-IMGT/HLA sequence database with full length sequence, identify novel alleles and allow studies into the impact of high resolution HLA typing on HSCT outcomes.

The full gene assay was limited in the ability to resolve all 4*th* field ambiguities due to the incomplete coverage of the UTR regions in several of the genes. HLA sequences, particularly the UTR regions, are being continually extended, making primer design that encompasses these new sequences unmanageable. Therefore, the cost benefits of continuingly redesigning primers beyond the protein coding regions of the gene needs to be considered. Our full gene method covers all exons and introns but is limited by the inability to completely resolve 4*th* field ambiguities in the UTR regions of class I genes, the inability to amplify the null allele DQB1*03:276N and the lack of appropriate software for analysis of introns in class II genes. To our knowledge there is no commercially available typing method that is able to distinguish DQB1*03:276N allele from DQB1*03:01:01 and therefore resolve this ambiguity. Although we were able to demonstrate the amplification of full gene products for the class II genes, the sequence analysis of these genes was restricted due to factors not related to PCR amplification. The introns of class II genes were omitted from analysis due to the prevalence of large numbers of repetitive sequence in the introns, mainly observed in the DRB and DPB1 genes, which resulted in a large number of erroneous base calls. In addition, when the analysis of DPB1 included the introns of the gene, the NGSengine software assigned phase incorrectly in the absence of reference sequence and the inability to link polymorphisms in one or more of the ambiguous alleles. The exclusion of the introns in DPB1 resulted in accurate allele calling albeit with increased allele ambiguity. The sequence data for the complete amplicons are however stored and could be re-analysed for contiguous full-gene sequence to further improve the genotyping resolution of these genes when further improvements in bioinformatics tools are made.

Many of the ambiguities reported using this assay were the result of the inability to phase due to short read chemistry. Even though this assay was validated and tested using the Ion Torrent platform, the full gene PCR assay has been applied to single molecule sequencing on the Oxford Nanopore MinION and PacBio sequencing platforms for completely phased-defined sequencing. With the advancements in the field of single molecule sequencing, bioinformatics and the collaborative international effort to extend the HLA reference database by the International HLA & Immunogenetics Workshop, it may be feasible to obtain 4*th* field typing resolution for the class II genes in the future. Nevertheless, the

newly developed assay has the capability to make several contributions to the efficiency of HLA genotyping in the routine diagnostic HLA laboratory.

Acknowledgements

We are grateful to all the colleagues at the Department of Clinical Immunology, PathWest and the Department of Transplantation Immunology, Maastricht University Medical Center for their technical assistance and troubleshooting. In particular, the team of Dr Mathijs Groeneweg and Ms Laila Gizzarelli (PathWest).

References

1. Shaw BE, Arguello R, Garcia-Sepulveda CA, Madrigal JA. The impact of HLA genotyping on survival following unrelated donor haematopoietic stem cell transplantation. British journal of haematology. 2010;150:251-8.
2. Bodmer WF. The HLA system: structure and function. Journal of Clinical Pathology. 1987;40:948-58.
3. Howell WM, Carter V, Clark B. The HLA system: immunobiology, HLA typing, antibody screening and crossmatching techniques. J Clin Pathol. 2010;63:387-90.
4. Tiercy JM. How to select the best available related or unrelated donor of hematopoietic stem cells? Haematologica. 2016;101:680-7.
5. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. Journal of human genetics. 2009;54:15-39.
6. Arora M, Weisdorf DJ, Spellman SR, Haagenson MD, Klein JP, Hurley CK, *et al*. HLA-identical sibling compared with 8/8 matched and mismatched unrelated donor bone marrow transplant for chronic phase chronic myeloid leukemia. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2009;27:1644-52.
7. Flomenberg N, Baxter-Lowe LA, Confer D, Fernandez-Vina M, Filipovich A, Horowitz M, *et al*. Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. Blood. 2004;104:1923-30.
8. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, *et al*. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood. 2007;110:4576-83.
9. Mayor NP, Hayhurst JD, Turner TR, Szydlo RM, Shaw BE, Bultitude WP, *et al*. Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2019;25:443-50.
10. ASHI. Standards for Accredited Laboratories. 2016.
11. Vince N, Li H, Ramsuran V, Naranbhai V, Duh FM, Fairfax BP, *et al*. HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region. American journal of human genetics. 2016;99:1353-8.
12. Laforet M, Froelich N, Parissiadis A, Bausinger H, Pfeiffer B, Tongio MM. An intronic mutation responsible for a low level of expression of an HLA-A*24 allele. Tissue antigens. 1997;50:340-6.
13. Petersdorf EW, Malkki M, O'hUigin C, Carrington M, Gooley T, Haagenson MD, *et al*. High HLA-DP Expression and Graft-versus-Host Disease. New England Journal of Medicine. 2015;373:599-609.
14. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. Nucleic acids research. 2015;43:D423-D31.

15. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, *et al*. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue antigens. 2012;80:305-16.
16. Hooper B, Whittingham S, Mathews JD, Mackay IR, Curnow DH. Autoimmunity in a rural community. Clin Exp Immunol. 1972;12:79-87.
17. Fae I, Wenda S, R. Mayr W, Fischer G. DQB1typing with Next Generation Sequencing: Detection of two new alleles. Tissue antigens. 2013:349.
18. Ozaki Y, Suzuki S, Kashiwase K, Shigenari A, Okudaira Y, Ito S, *et al*. Cost-efficient multiplex PCR for routine genotyping of up to nine classical HLA loci in a single analytical run of multiple samples by next generation sequencing. BMC genomics. 2015;16:318.
19. Smith AG, Pyo CW, Nelson W, Gow E, Wang R, Shen S, *et al*. Next generation sequencing to determine HLA class II genotypes in a cohort of hematopoietic cell transplant patients and donors. Hum Immunol. 2014;75:1040-6.
20. Zhou M, Gao D, Chai X, Liu J, Lan Z, Liu Q, *et al*. Application of high-throughput, high-resolution and cost-effective next generation sequencing-based large-scale HLA typing in donor registry. Tissue antigens. 2015;85:20-8.
21. Danzer M, Niklas N, Stabentheiner S, Hofer K, Proll J, Stuckler C, *et al*. Rapid, scalable and highly automated HLA genotyping using next-generation sequencing: a transition from research to diagnostics. BMC genomics. 2013;14:221.
22. Ehrenberg PK, Geretz A, Sindhu RK, Vayntrub T, Fernandez Vina MA, Apps R, *et al*. High-throughput next-generation sequencing to genotype six classical HLA loci from 96 donors in a single MiSeq run. Hla. 2017;90:284-91.
23. Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q. Impact of three Illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol. 2015;76:166-75.
24. Walsh PS, Erlich HA, Higuchi R. Preferential PCR amplification of alleles: mechanisms and solutions. PCR methods and applications. 1992;1:241-50.
25. Wagner A, Blackstone N, Cartwright P, Dick M, Misof B, Snow P, *et al*. Surveys of Gene Families Using Polymerase Chain Reaction: PCR Selection and PCR Drift. Systematic Biology. 1994;43:250-61.

Part 2.
What's in a Haplotype?

CHAPTER 6

6

Insights into the polymorphism in HLA-DRA and its evolutionary relationship with HLA haplotypes

B.M. Matern, T.I. Olieslagers, C.E.M. Voorter, M. Groeneweg, M.G.J. Tilanus

Transplantation Immunology, Tissue Typing Laboratory, Maastricht University Medical Center, Maastricht, The Netherlands

HLA. 2019 Oct 16. doi: 10.1111/tan.13730

Abstract

Abstract

HLA-DRA encodes the alpha chain of the HLA-DR protein, one of the classical HLA class II molecules. Reported polymorphism within HLA-DRA is currently limited compared with other HLA genes, as only a single polymorphism encodes an amino acid difference in the translated protein. Since this SNP (rs7192, HLA00662.1:g.4276G>T p.Val217Leu) lies within exon 4, in the region encoding the cytoplasmic tail, the resulting protein is effectively monomorphic. For this reason, in-depth studies on HLA-DRA and its function have been limited. However, analysis of sequences from the 1000 Genomes Project and preliminary data from our lab reveals unrepresented polymorphism within HLA-DRA, suggesting a more complex role within the MHC than previously assumed. This study focuses on elucidating the extent of HLA-DRA polymorphism, and extending our understanding of the gene's role in HLA-DR~HLA-DQ haplotypes. 98 samples were sequenced for full-length HLA-DRA, and from this analysis, we identified 20 novel SNP positions in the intronic sequences within the 5711 bp region represented in IPD-IMGT/HLA. This polymorphism gives rise to at least 22 novel HLA-DRA alleles, and the patterns of intronic and 3′ UTR polymorphism correspond to HLA-DRA~HLA-DRB345~HLA-DRB1~HLA-DQB1 haplotypes. The current understanding of the organization of the genes within the HLA-DR region assumes a single lineage for the HLA-DRA gene, as opposed to multiple gene lineages, such as in HLA-DRB. This study suggests that the intron and 3′ UTR polymorphism of HLA-DRA indicates different lineages, and represents the HLA-DRA~HLA-DRB345~HLA-DRB1~HLA-DQB1 haplotypes.

Introduction:

HLA

Human Leukocyte Antigen (HLA) is the name given to the Major Histocompatibility Complex (MHC) in humans. The primary function of class II HLA is the presentation of extracellular peptides on the cell surface. When combined with an antigen, class II HLA creates a complex on the cell surface that acts as a ligand for T-cell receptors. T-cells interact with presented antigens to distinguish "self" and "non-self" peptides, and they can subsequently trigger an immunogenic response, giving HLA a critical role in the adaptive immune system.

The HLA molecule can present a variety of peptides from self or non-self antigens, providing individuals and populations with the capability to respond to a variety of pathogens. The variability is provided by the polymorphism in the HLA proteins, especially in the subunits responsible for antigen presentation. The protein polymorphism is reflected in the nucleotide sequence polymorphism, making the MHC on chromosome 6 the most hyperpolymorphic region of the human genome.[1] Polymorphism within HLA genes provides it's capability to detect and respond to foreign tissue, which can create a complication for solid organ and stem cell transplantations (SCT). Matching of HLA genes reduces immunogenic response and the subsequent transplant-related complications, such as organ rejection and graft-versus-host disease. In addition to genotype matching, matching for SCT based on phased haplotypes can also affect both GVHD and disease recurrence.[2] Although HLA polymorphism is vast and complex,[3] some patterns of polymorphism can be observed with the help of publicly available databases, such as the 1000 Genomes Project[4] or IPD-IMGT/HLA.[5]

Due to its critical role in immune response, HLA is associated with a large number of human diseases.[6] GWAS studies have found, just within the HLA-DR region, associations with Multiple Sclerosis,[7] Alzheimer's disease,[8] penicillin allergy,[9] Ulcerative Colitis,[10] and dry eye disease,[11] among many others. The biological mechanisms that bring about a disease can vary widely,[6] and statistical correlations to individual SNPs may not imply a functional role for the polymorphism, but may be linked to a different locus, meaning the SNP serves as a proxy to functional polymorphism. Studies of the polymorphism within HLA are necessary to clarify the mechanisms and patterns of inheritance of disease associations within this region.

Polymorphism in non-coding regions of the HLA genes, such as introns and UTRs, is not expressed in the protein, but it can have significant effects on biological function, such as changes in expression levels and expression regulation. The promoter region is contained in the UTR sequence upstream of the HLA coding regions, and HLA is rich

with microRNA encoded sequences,[12] including some encoded within the introns of HLA-DRA.[13] Polymorphism within either the encoding microRNA sequences or within the expressed messenger RNA can affect HLA expression, or may affect a variety of functions in human biology.

HLA-DR

HLA-DR is one of the class II HLA proteins. Like the other class II molecules, HLA-DR comprises a heterodimeric alpha and beta chain, which are encoded by five genes located within the human chromosome 6. HLA-DRA encodes the HLA-DR alpha chain, and the beta chain can be encoded by one of four distinct genes, most commonly HLA-DRB1. The beta chain can also be encoded by one of the HLA-DRB3, HLA-DRB4, or HLA-DRB5, genes, which are generally expressed at lower levels than HLA-DRB1,[14] and may or may not be present in a given haplotype, depending on the associated HLA-DRB1 gene.[15] HLA-DR molecules encoded by the different HLA-DRB genes generally present overlapping sets of peptides, but in some cases distinct HLA-DR proteins have been shown to present distinct, or even complementary peptide repertoires.[16] The alpha chain's ability to interact with multiple beta chains allows the formation of multiple potential protein heterodimers, and therefore ensures the expression of at least one HLA-DR molecule.

HLA-DR can be contrasted with the other class II proteins in that the polymorphism is determined primarily by the beta chain. This lack of known polymorphism is exhibited in the IPD-IMGT/HLA database, a repository for HLA reference sequences which serves as the standard knowledge base for HLA research.[5] Recent releases of the IPD-IMGT/HLA database (up to 3.37.0, July 2019) contain 7 HLA-DRA reference allele sequences, a count which is unchanged since release 3.4.0 (April 2011), and which is notably fewer than the 132 alleles for HLA-DPA1, 183 for HLA-DQA1, or 3,171 HLA-DRB alleles (3.37.0). Six of the seven HLA-DRA sequences are provided as full-length sequences, from 5′ UTR to 3′ UTR, including introns.[17] There are four known SNPs within the HLA-DRA exons, two each in exons 3 and 4. Three of the four exon SNPs are silent mutations, and therefore do not change the encoded protein. The sole amino acid polymorphism in HLA-DRA (rs7192, HLA00662.1:g.4276G>T p.Val217Leu)[18] is located in the region of exon 4, which encodes the cytoplasmic tail of the HLA-DR alpha chain.[19] This SNP defines the difference between the two HLA-DRA proteins, HLA-DRA*01:01 and HLA-DRA*01:02.[20] This polymorphism results in two distinct proteins, but since the functional change is in a region unrelated to antigen presentation, the encoded protein is considered monomorphic.[21] HLA-DRA is, however, marked by notable intronic polymorphism. In contrast to the nearly monomorphic exons, IPD-IMGT/HLA (release 3.37.0) shows 60 SNP positions and two major deletions in the HLA-DRA introns.

HLA Gene Organization

Within humans, the HLA-DRA gene is situated within one of the known HLA haplotype patterns, which are defined by the ubiquitous HLA-DRB1 gene. Haplotype patterns vary in which HLA-DRB1, HLA-DQA1 and HLA-DQB1 alleles are present, and also determine the presence or absence of protein-coding HLA-DRB3,4,5 genes. The haplotypes are also marked by non-coding pseudogenes, including HLA-DRB2,6,7,8, and 9. These pseudogenes are products of historical duplication and recombination events,[22] and although their biological function, if any, is unknown, they can provide valuable evolutionary perspective.

Data from non-human primates exhibit much more polymorphism within the HLA-DRA gene than has been defined within humans.[23,24] In addition, haplotype patterns with distinct organizations of HLA-DRB genes and pseudogenes have been identified within non-human primates, suggesting that ancestral haplotype patterns have extensive variation.[14]

Evidence from the 1000 Genomes Project database combined with observed patterns in non-human primates suggests that the HLA-DRA gene as seen in IPD-IMGT/HLA represents an underestimation of the total polymorphism that exists in the gene. This study seeks to define the polymorphism within HLA-DRA using the MinION single-molecule nanopore sequencing approach, which provides an appealing platform for full-length HLA-DRA sequencing and phasing. Furthermore, this study explores whether the HLA-DRA polymorphism defines general sequence patterns, and aims to elucidate associations and it's patterns of inheritance with HLA-DRB3/4/5, HLA-DRB1, and HLA-DQB1.

Material and Methods:

Initial analysis was performed on the HLA-DRA region of the 1000 Genomes Project database.[4] The ensembl database provides variant frequency data from the 1000 Genomes Project,[25] which was downloaded in the form of a table of SNP information in the region near the HLA-DRA locus. SNP locations were filtered by a minimum minor allele frequency (MAF) of 1%, and correlated with corresponding positions within IPD-IMGT/HLA. SNPs were categorized by genomic feature, and by whether they are also represented in IPD-IMGT/HLA.

A panel of samples was selected for subsequent MinION sequencing of HLA-DRA. 98 Samples were chosen based on their availability in the laboratory, and stratified based on the known HLA-DRB types. The panel was selected so that each HLA-DRB1 allele group, as distinguished by the first field of standard HLA nomenclature,[20] was included for complete

HLA-DRB1 coverage. Each sample was typed at the HLA-DRB1, HLA-DRB3/4/5 and HLA-DQB1 loci using Sanger SBT (SSBT).[26]

DNA was isolated using the QiaAMP DNA Blood mini kit (Qiagen, Hilden, Germany) from 5 ml EDTA blood. In short, the blood was incubated in ammonium chloride for erythrocyte lysis and washed with PBS. The cell pellet was incubated with Qiagen protease at 56°C for 10 minutes, and subsequently applied to the QiaAMP column. The columns were then used according to the manufacturer's instructions. DNA concentration and quality was assessed using spectrophotometry with the DeNovix DS-11 FX, (Denovix, Wilmington, DE, USA) with a minimum requirement of 100 ng of DNA at a concentration of 100-200 ng/µl.

Primers were designed to amplify the full-length HLA-DRA gene (Supplementary Table 3) to form an amplicon 8.7 kb in length. The amplicon spans the 5.7 kb HLA-DRA gene as represented in IPD-IMGT/HLA, with an additional 1 kb of sequence on the 5′ end and 1.9 kb of the downstream 3′ sequence. Samples were amplified using the Expand Long Template PCR System (Roche, Basel, Switzerland) according to protocols outlined in Voorter *et al*.[26]

The PCR amplicons were prepared for sequencing using standard Oxford Nanopore 1D^2 protocols (Oxford Nanopore, Oxford, UK). The amplicons were cleaned using 1:1 CleanPCR beads (GCbiotech, Waddinxveen, the Netherlands). The Oxford Nanopore 96x barcoding kit was used to affix barcode sequences by annealing to the universal MinION barcode tag sequence included in our MinION amplification primers (Supplementary Table 3). After a secondary cleanup using CleanPCR beads, the barcoded samples were pooled, and the ends of the double-stranded DNA were repaired, and prepared for sequencing by affixing an adenine nucleotide. Barcoded amplicons were further ligated with 1D^2 sequencing adapters and sequenced using Oxford Nanopore 1D^2 Sequencing.

The 1D^2 reads were basecalled and paired using Albacore (v2.3.3, Oxford Nanopore, Oxford, UK) . The nanopore reads are demultiplexed, and barcode sequences were removed using PoreChop (v0.2.3).[27] Nanopore Prospector[28] is a set of scripts developed for various analysis of nanopore reads in the context of HLA, which was used for subsequent SNP analysis of the HLA-DRA reads. Reads were filtered based on expected amplicon length and homology with the expected HLA-DRA sequence, to exclude any off-target reads. From an alignment of the 1D^2 nanopore reads against a known HLA-DRA*01:01:01:01 reference from IPD-IMGT/HLA, heterozygous positions were identified and used in combination with the k-means clustering algorithm[29,30] to phase the reads into two pools, each representing one of an individual's HLA-DRA alleles. The pool of reads for each allele was realigned to the same reference, and SNPs were identified and categorized based on which genomic feature (exon, intron, or UTR) they lie within.

From the analysis of full-length nanopore reads, sequence patterns were identified within each genomic feature, providing a basis to separate sequences into clusters based on major patterns of polymorphism. Sequence motifs were defined based on observations of general SNP patterns, and these motifs were used to cluster the sequences. Inclusion within a sequence pattern cluster was determined based on a rule that a sequence can have only a single nucleotide difference from a pattern cluster's motif, allowing us to observe general polymorphism patterns while allowing some minor polymorphism. There are two homopolymer regions, at IPD-IMGT/HLA genomic positions 848 ("A" homopolymer, 8-11 bases) and 1828 ("A" homopolymer, 9-10 bases), which present a challenge for analysis and interpretation using only MinION reads. These positions were disregarded in the pattern clusters, but were confirmed via SSBT for novel alleles. Individual sequence pattern clusters were combined into a full-length sequence pattern for each sequence. This allowed clustering of the HLA-DRA alleles by general patterns of polymorphism, and these sequence clusters were used in subsequent haplotype analysis.

The sequence patterns were analyzed in the context of the corresponding HLA-DRB1, HLA-DRB3,4,5, and HLA-DQB1 genes. Pypop 0.7.0[31] was used to identify haplotype patterns based on the expectation maximization algorithm, and sample haplotype frequencies were used to phase the HLA-DRA polymorphism against the corresponding HLA-DRB and HLA-DQB1 alleles.

Sequences containing novel polymorphisms were validated and submitted to EMBL-ENA, and subsequently to IPD-IMGT/HLA. Allele names were assigned by the WHO Nomenclature Committee in August 2019. This follows the agreed policy that, subject to the conditions stated in the most recent Nomenclature Report,[20] names will be assigned to new sequences as they are identified. HLA allele names and ENA accession numbers are shown in Table 2. Confirmatory sequencing was performed on the HLA-DRA gene using the Sanger-based allele-specific sequencing approach previously described by Voorter *et al*.[26] using primers shown in Supplementary Table 3. Sequence polymorphisms were identified using Lasergene 15 (Version 15.2.0, DNASTAR. Madison, WI. USA.) and compared against the results from MinION sequencing and known polymorphism within IPD-IMGT/HLA. Secondary analysis and confirmation of the 1D^2 reads was performed using GenDx NGSEngine (Version 2.11.0.11444, GenDx, Utrecht, The Netherlands)

Results

A schematic overview of the known polymorphism found by analysis of the 1000 Genomes Project data is presented in Figure 1. In this analysis we were able to confirm the 62 polymorphisms that are represented in IPD-IMGT/HLA, and suggest an additional 120 polymorphic positions that are not represented. Every polymorphism that is found in either IPD-IMGT/HLA, or our panel of MinION sequences were also found in the 1000 Genomes Project data. Each SNP in 1000 Genomes Project data includes a measure of the Minor Allele Frequency (MAF), which represents the percentage of sequences which possess the second-most common allele, and provides an estimation of how polymorphic a position is. The 1000 Genomes Project represents sequences of over 2500 individuals, and SNPs can have a variety of population frequencies. Some rare SNPs have MAF values that are rounded to 0% and thus SNPs with MAF < 1% were disregarded in our analysis. Within exons 3 and 4, no additional common SNPs were identified. Interestingly, a single SNP (rs16822586, HLA00662.1:g.405G>C p.Val-10Leu) with a MAF of 3.9% was reported within Exon 1, but this SNP was not found in either IPD-IMGT/HLA or in the sequence analysis of our sample panel. At this position the minor C allele results in a Valine -> Leucine amino acid change in the leader sequence of the protein, where it may affect translocation of the HLA protein to the cell surface. The minor allele is rarely found in the European population, and is more common (3-7%) in the African and Eastern Asian populations.

HLA-DRA Exon Polymorphism

From our HLA-DRA full-length sequencing panel, known exon polymorphisms from IPD-IMGT/HLA were confirmed, but no novel exon SNPs were identified. Furthermore, there are no novel combinations of these four SNPs; every sequence follows one of five patterns, as seen in Table 1. Since we did not encounter the HLA-DRA*01:02:01 pattern, every HLA-DRA sequence identified in our panel matches one of four exon patterns. For the purposes of haplotype analysis, the HLA-DRA sequences were clustered based on matching exon sequences known in IPD-IMGT/HLA, for example, HLA-DRA*01:02:03-like sequences.

The Pypop haplotype analysis was used to phase known HLA-DRA polymorphisms against nearby HLA-DRB and HLA-DQB1 genes. Pypop provided sample haplotype frequency estimates based on sample population genotypes, which were used to phase the HLA alleles in the HLA-DR region. A comparison of the HLA-DRA exons with the nearby HLA-DRB1 can be seen in Figure 2. Some patterns become apparent when looking closely at the exon patterns and the nearby genes. A notable example is the green pattern of HLA-DRA*01:02:03-like sequences, which is only found in the HLA-DRB1*15 allele-group. Each of the 14 sequences is found to be in a haplotype with HLA-DRB5*01:01 and HLA-DRB1*15:01:01, demonstrating a strong correlation between HLA-DRA*01:02:03-like sequences and HLA-DRB1*15:01:01.

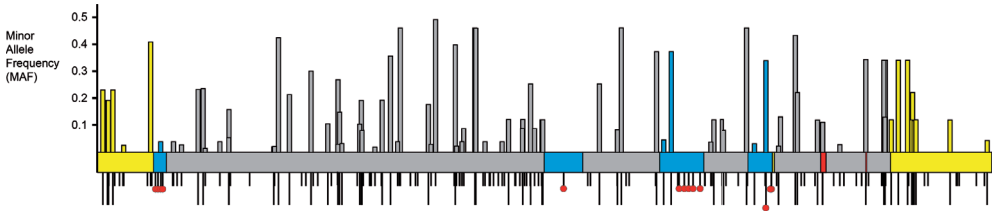


Figure 1 : Schematic overview of polymorphism within the HLA-DRA region
A representation of HLA-DRA polymorphism as seen in the 1000 Genomes Project. The black vertical bars below the figure represent all polymorphism as represented in the 1000 Genomes Project. SNPs that are represented in IPD-IMGT/HLA (up to release 3.37.0) have a longer vertical bar than the SNPs without IPD-IMGT/HLA representation. SNPs with a Minor Allele Frequency(MAF) > .01 are also represented with a vertical bar above the figure, with a height reflecting the MAF. Blue regions are exons, and yellow regions represent UTRs. Some HLA-DRA alleles feature one or both of the two major intron 4 deletions, shown in red. Missense SNPs are indicated by a red circle below the SNP marker.

	EX3: 3617	EX3: 3665	EX4: 4203	EX4: 4276
HLA-DRA*01:01:01	**G**	**C**	**C**	**G**
HLA-DRA*01:01:02	**G**	**A**	**T**	**G**
HLA-DRA*01:02:01	**G**	**C**	**C**	**T**
HLA-DRA*01:02:02	**G**	**A**	**C**	**T**
HLA-DRA*01:02:03	**A**	**A**	**C**	**T**

Table 1. Exon Polymorphism PatternsHLA-DRA has four exonic SNPs, contained within exon 3 and 4, which are found in one of five distinct SNP patterns. These SNPs are shown with their gDNA positions as seen in genomic alignments from IPD-IMGT/HLA**.** EX4: 4276 is the only SNP which encodes an amino acid difference. This SNP lies in the cytoplasmic tail region of exon 4, and it defines the difference between HLA-DRA*01:01 and HLA-DRA*01:02 proteins. Neither HLA-DRA*01:02:01, nor any novel exon polymorphisms or combinations of known SNPs were identified within our sequencing panel.

Another distinct pattern can be seen in haplotypes containing HLA-DRB1*07 alleles. Based on the HLA-DRA exon patterns, the HLA-DRB1*07 sequences are split into two haplotypes with distinct HLA-DQB1 and HLA-DRB4 types. Every sequence with HLA-DRA*01:01:01-like exons (Red) is correlated with HLA-DQB1*02:02. This haplotype also has a HLA-DRB4*01 sequence, but the second field cannot be determined from the HLA-DRA exon pattern alone. Every sequence with HLA-DRA*01:01:02-like exons (Yellow) forms a haplotype with HLA-DQB1*03:03, and a HLA-DRB4*01:03:01:02N, demonstrating another correlation pattern between the HLA-DRA exons and a null HLA-DRB4 allele. The HLA-DRB1*04 column has only a single sequence with the HLA-DRA*01:01:02 exon pattern, and this is also correlated with a HLA-DRB4*01:03:01:02N allele, strengthening the connection between HLA-DRA and HLA-DRB4.

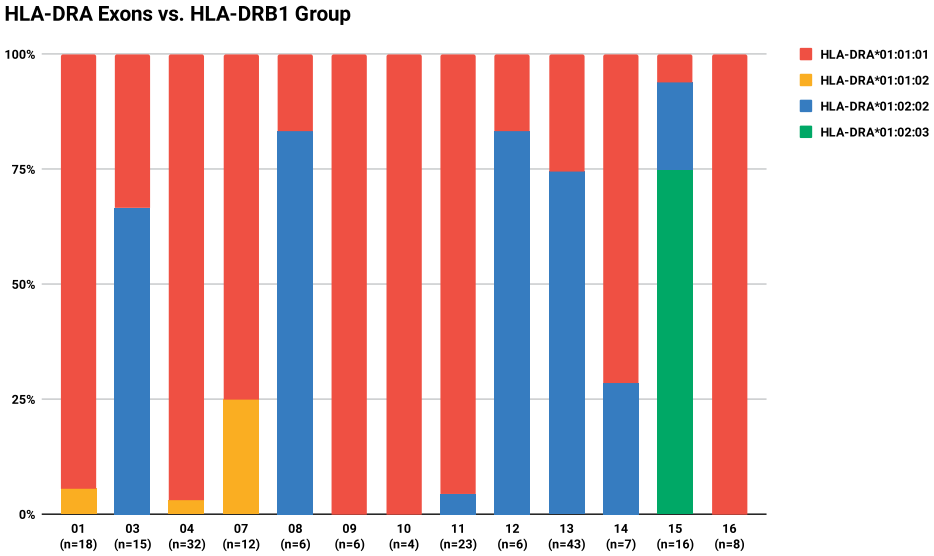


Figure 2. HLA-DRA Exon patterns vs HLA-DRB1 groups.
This figure shows the relationship of the exon sequences from our sequencing panel, and their linked HLA-DRB1 alleles. HLA-DRB1 groups are represented as vertical columns, and the four colors represent the proportion of haplotypes that have one of the known HLA-DRA exon patterns. A fifth exon pattern, which matches the HLA-DRA*01:02:01 allele, is listed in IPD-IMGT/HLA but was not identified in our panel.

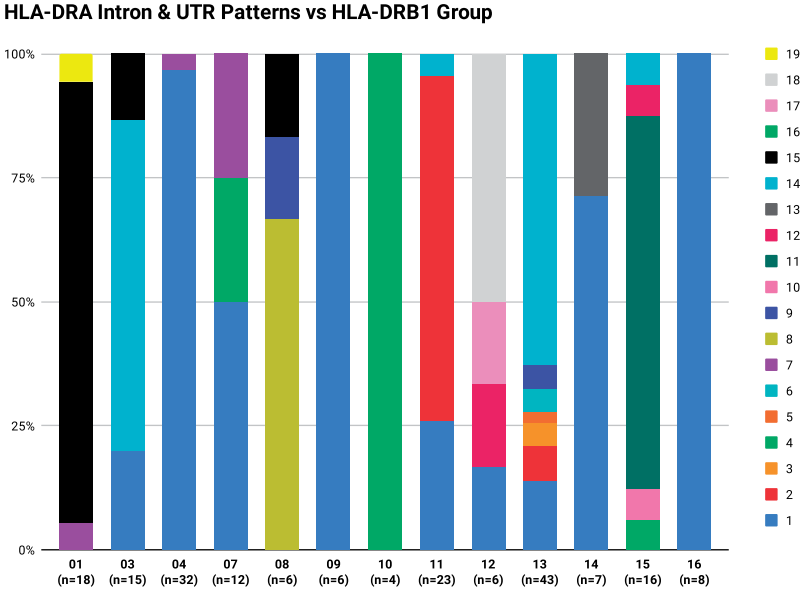


Figure 3. Patterns of HLA-DRA Intron & UTR polymorphism forms distinct haplotype patterns with HLA-DRB1.
The HLA-DRA intron & UTR patterns are represented by the distinct colors, and each column represents a HLA-DRB1 group. In many cases, such as in HLA-DRB1*03 and HLA-DRB1*07, the HLA-DRA polymorphism pattern indicates a proxy for determining the HLA-DRB345~HLA-DRB1~HLA-DQB1 haplotype. See Supplementary Table 2 for a summary of the SNP patterns and haplotypes.

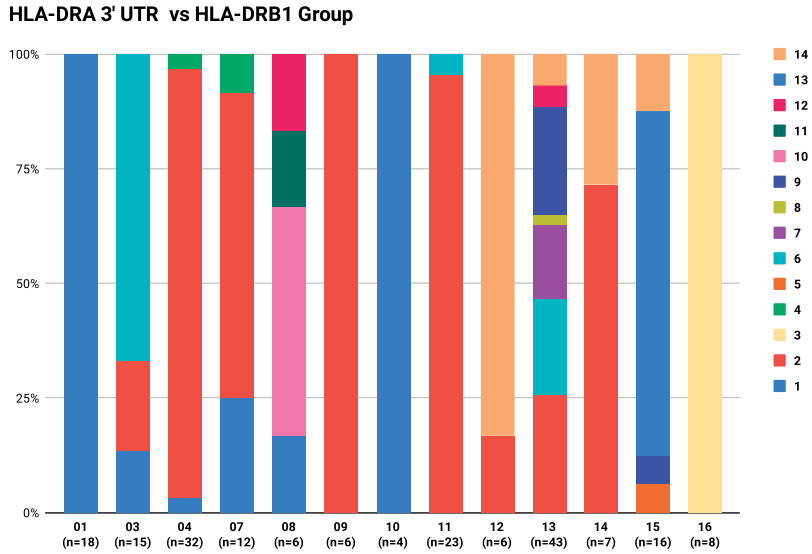


Figure 4. HLA-DRA 3′ UTR compared with HLA-DRB1.
This shows the relationship between the 3′ UTR of HLA-DRA with HLA-DRB1. Distinctions between the 3′ UTR sequences is defined by polymorphism outside the region represented by IPD-IMGT/HLA, which is reflected in the pattern names. Haplotype patterns between HLA-DRA and HLA-DRB~HLA-DQB are apparent, especially in the case of HLA-DRB1*13.

HLA-DRA Intronic Polymorphism vs HLA-DRB1 genes.

The known HLA-DRA sequence in IPD-IMGT/HLA is represented as a region 5711 base pairs in length, and few studies have analyzed the sequence outside this region. Our PCR amplicon covers a region about 3000 bases further than the IPD-IMGT/HLA region, and while the patterns of polymorphism within these regions were used for the purpose of pattern identification, the identification of novel alleles is restricted to the intron and UTR regions as represented in IPD-IMGT/HLA. Within this region, 20 novel SNP positions were identified within the introns, which combined with the 60 currently known intronic polymorphic positions in IPD-IMGT/HLA brings the total to 80. Every polymorphism identified is also found within the 1000 Genomes Project data. These 20 novel SNPs form the basis for 22 novel alleles, shown in Table 2.[32]

Clustering analysis of the intron patterns revealed distinct patterns of polymorphism within the introns and UTRs. The sequence clustering identified the major patterns of inheritance, while allowing some minor polymorphism between allele sequences. The population haplotype analysis resulted in phasing of the HLA-DRA polymorphism against HLA-DRB1 and HLA-DRB345 genes, which revealed distinct HLA-DRA~HLA-DRB345~HLA-DRB1~HLA-DQB1 haplotype patterns. Inheritance patterns become apparent when the HLA-DRA introns, represented by the different colors, are compared with HLA-DRB1,

as summarized in Figure 3. It is clear that some HLA-DRA intron patterns, especially pattern #1, are correlated with many different HLA-DRB1 groups, but looking closely at the HLA-DRB1 groups reveals distinct patterns. One notable pattern can be seen in the allele group HLA-DRB1*10. In this column, each HLA-DRA allele has a light-green polymorphism pattern (#16), which is not found on the haplotype of any other HLA-DRB1* group. Furthermore, each of these sequences is observed to be present on a HLA-DRA*01:01~HLA-DRB1*10:01~HLA-DQB1*05:01 haplotype.

Further haplotype patterns can be observed by closer analysis of the individual HLA-DRB1 groups. For instance, HLA-DRB1*03 correspond to three HLA-DRA intron patterns, representing three distinct haplotypes. HLA-DRA intron pattern 1 (Blue) corresponds to the haplotype HLA-DRB3*02:02~HLA-DRB1*03:01~HLA-DQB1*02:01. Likewise, HLA-DRA intron pattern 14 (Cyan), represents the haplotype HLA-DRB3*01:01~HLA-DRB1*03:01~HLA-DQB1*02:01. The remaining intron pattern 15 (Black) is two sequences, one of which is a HLA-DRB3*01:01~HLA-DRB1*03:02~HLA-DQB1*04:02 haplotype, and the other matches the pattern 14 haplotype, possibly indicating a recombination.

HLA-DRB1*07 shows three patterns in the HLA-DRA introns, which support the patterns observed in the HLA-DRA exons. In this case, HLA-DRA intron pattern 1 (Blue) represents a haplotype HLA-DRB4*01:03~HLA-DRB1*07:01~HLA-DQB1*02:02. Intron pattern 4 (Green) indicates the HLA-DRB4*01:01~HLA-DRB1*07:01~HLA-DQB1*02:02 haplotype, while pattern 7 (Purple) represents the distinct haplotype HLA-DRB4*01:03:01:02N~HLA-DRB1*07:01~HLA-DQB1*03:03, demonstrating a distinction of the HLA-DRB4*01 alleles based on the HLA-DRA intron sequences, and providing more evidence on the correlation between the introns of HLA-DRA, and the HLA-DRB1 and HLA-DRB4 types.

Closer comparisons of the HLA-DRA 3′ UTR with the phased HLA-DRB and HLA-DQB1 genes elucidated more specific haplotype patterns, as seen in Figure 4. This is most notable within the HLA-DRB1*13 column, which is correlated with 7 HLA-DRA 3′ patterns. Each HLA-DRA pattern defines one of 5 distinct haplotypes. For example, pattern 2 indicates HLA-DRB3*02:02~HLA-DRB1*13:01~HLA-DQB1*06:03, patterns 6 and 8 represents HLA-DRB3*03:01~HLA-DRB1*13:02~HLA-DQB1*06:04, while patterns 7 and 9 represent HLA-DRB3*01:01~HLA-DRB1*13:01~HLA-DQB1*06:03. Similar patterns between the HLA-DRA 3′ UTR and HLA-DR~HLA-DQ haplotypes can be identified in the other HLA-DRB1 groups, see Supplementary Table 1 for HLA-DRA SNPs and haplotype information.

Discussion:

Our sequencing analysis reveals considerable polymorphism within HLA-DRA, indicating that this gene is more polymorphic than what is commonly assumed. Newly identified polymorphism is located in the Intron and UTR sequences, and no novel polymorphism or SNP patterns were identified in the exons. Full-length single molecule nanopore sequencing has enabled unambiguous phasing of the polymorphism across the gene, an achievement which may not be possible by traditional short-read NGS analysis of an HLA gene with little available reference sequence and relatively low heterozygosity.

SampleID	Allele Local Name	EMBL-ENA Accession#	IPD-IMGT/HLA Allele Name
19583	HLA-DRA_MUMC_01	LR606129	HLA-DRA*01:02:02:05
19583	HLA-DRA_MUMC_02	LR606128	HLA-DRA*01:02:02:06
2271	HLA-DRA_MUMC_03	LR606130	HLA-DRA*01:02:02:03
17728	HLA-DRA_MUMC_04	LR606275	HLA-DRA*01:01:01:05
20910	HLA-DRA_MUMC_05	LR606131	HLA-DRA*01:01:01:07
8394	HLA-DRA_MUMC_06	LR606132	HLA-DRA*01:02:02:04
22611	HLA-DRA_MUMC_07	LR606127	HLA-DRA*01:02:02:08
47732	HLA-DRA_MUMC_08	LR606122	HLA-DRA*01:02:02:14
20492	HLA-DRA_MUMC_09	LR606124	HLA-DRA*01:02:02:07
98	HLA-DRA_MUMC_11	LR606274	HLA-DRA*01:01:01:04
152	HLA-DRA_MUMC_12	LR606123	HLA-DRA*01:02:02:02
25945	HLA-DRA_MUMC_13	LR606277	HLA-DRA*01:02:02:10
25945	HLA-DRA_MUMC_14	LR606276	HLA-DRA*01:01:01:10
25689	HLA-DRA_MUMC_16	LR606278	HLA-DRA*01:02:02:09
23030	HLA-DRA_MUMC_17	LR606279	HLA-DRA*01:01:01:08
17728	HLA-DRA_MUMC_18	LR606125	HLA-DRA*01:01:01:06
39183	HLA-DRA_MUMC_20	LR606281	HLA-DRA*01:02:02:12
47506	HLA-DRA_MUMC_21	LR606280	HLA-DRA*01:01:01:12
45817	HLA-DRA_MUMC_22	LR606282	HLA-DRA*01:02:02:13
24673	HLA-DRA_MUMC_23	LR606283	HLA-DRA*01:01:01:09
47506	HLA-DRA_MUMC_24	LR606285	HLA-DRA*01:01:01:11
39183	HLA-DRA_MUMC_25	LR606126	HLA-DRA*01:02:02:11

Table 2 - Novel HLA-DRA Alleles. Alleles with novel polymorphism were confirmed using SSBT. These alleles were submitted to EMBL-ENA using Saddlebags software,[32] and subsequently to IPD-IMGT/HLA. The allele names have been officially assigned by the WHO Nomenclature Committee in August 2019.[20]

No novel polymorphic positions were identified within the exons of HLA-DRA, every sample in our panel matches one of the known exon patterns. Interestingly, the exons showed no novel combinations of the SNPs, every combination of our exon SNPs matches one of four previously known exon sequences. The fifth exon pattern represented in IPD-IMGT/HLA corresponds to HLA-DRA*01:02:01, and this pattern was not found in our panel. The conserved exons suggest a strong selective pressure on the expressed alpha subunit of the HLA-DR protein, which could contrast the classic understanding of HLA, which maintains that the polymorphism of HLA proteins provides the flexibility of an individual's immune system. This widespread variability provides a species the ability to present a wide variety of peptides, enabling the defense against many threats.

The conserved exons within HLA-DRA suggest a different selective pressure. In class II HLA proteins, the alpha subunit typically presents a lower level of polymorphism than the corresponding beta subunit, which is most apparent in HLA-DR; HLA-DRA expresses 2 distinct proteins in humans, which can be paired with any of the expressed HLA-DRB genes, encompassing at least 2,226 proteins. Although the HLA-DRB genes are expressed at varying levels,[33] HLA-DRA is always expressed at slightly higher levels.[34] This, combined with the absence of any known null HLA-DRA alleles, suggests that HLA-DRA is essential for HLA-DR expression. The variety in expressed HLA-DRB genes and alleles leads to variability in the structure of the beta protein. It is conceivable that selective pressures keep the alpha subunit well conserved, as variability within HLA-DRA could lead to incompatible alpha and beta subunits.

Although there are only a few known SNPs within the exons of HLA-DRA, this polymorphism can still affect the behavior and function of the HLA-DRA protein. In addition to the coding SNP found in exon 4 (rs7192, HLA00662.1:g.4276G>T p.Val217Leu), one functional difference can be observed in rs8084 (HLA00662.1:g.3665C>A p.Ile109=), a C>A SNP within exon 3 of HLA-DRA. Although the polymorphism results in a synonymous amino acid sequence, an A nucleotide at this position creates an AG sequence, which is a known splice acceptor sequence.[35] In the case of HLA-DRA, this polymorphism can produce an alternative splice variant, where the mRNA is missing the first 75 bases of exon 3. We have preliminarily observed the existence of this splice variant using amplification primers specific to cDNA containing the splice variant, but the viability of the alternatively spliced variant is not fully understood. Future studies may elucidate the role of this polymorphism on splicing and the function of the HLA-DR protein, it's interaction with T-Cell receptors, and potential explanations for mechanisms of disease associations.

While not as polymorphic as the other class II alpha genes (HLA-DQA1 and HLA-DPA1), there is extensive polymorphism within the intron and UTR sequences in HLA-DRA. The function of the variation in the introns, and potential off-target effects are not yet known.

The Ventor group has extensively explored the general patterns in polymorphism within the non-coding regions within the human genome, by comparing the known pathogenic variants in the non-coding and regulatory elements.[36] They found that regions of the genome which contain regulatory sequences are more often "constrained", meaning they have less variability in the patterns of nucleotides, demonstrating that observations of the patterns of polymorphism within non-coding regions may help identify or understand the regulatory behavior of non-coding regions.

The 3′ UTR sequence of some class II HLA genes has been shown to be characteristic for the gene and surrounding sequence,[37] and this insight informed our further haplotype analysis focused on the HLA-DRA 3′ UTR sequences. Specific analysis of the 3′ UTR provides more refined clusters than clusters based on whole-gene intron patterns, and allows a better understanding of how HLA-DRA polymorphism can be used as an indicator for identifying HLA-DR~HLA-DQ haplotypes. IPD-IMGT/HLA represents HLA-DRA alleles as a 5.7kb region, including 1.4kb of sequence representing the intron 4 and the 3′ UTR of the gene. Our amplicon and sequencing results cover an additional 1.9kb of 3′ sequence beyond the IPD-IMGT/HLA region. Interestingly, several of our identified alleles (HLA-DRA*01:01:01:12, HLA-DRA*01:02:02:08, HLA-DRA*01:01:01:01, HLA-DRA*01:01:01:02, HLA-DRA*01:01:01:03), while identical within the IPD-IMGT/HLA region, showed significant polymorphism in the downstream 3′ region. This polymorphism is not reflected in the IPD-IMGT/HLA allele sequences and HLA allele names, but does represent divisions in the HLA-DR~HLA-DQ haplotypes (Figure 4). This reinforces that the 3′ region is significant to identification of genomic context, and that it is important to collect as much sequence as feasible, to expand HLA databases and our understanding of HLA haplotypes.

SNPs within the introns and UTR sequences of HLA-DRA have been correlated with a variety of human diseases.[7-11] These polymorphisms may have direct influence on the behavior of the HLA protein. Polymorphism within promoter or microRNA-encoding sequences may change expression of a protein, and polymorphism within introns or exons can introduce alternative splice variants, perhaps changing the way HLA interacts with T-Cells.[6] HLA-DRA exon polymorphism shows an association of HLA-DRA*01:02:03 with HLA-DRB1*15:01. HLA-DRB1*15:01 has well known associations with Multiple Sclerosis,[38] and the biological mechanisms of the disease are being explored. This association is most likely not the result of a single gene or SNP, but the disease is related to haplotype effects[39] and the combined function of multiple genes within the MHC. Polymorphism can often not be assigned a distinct function, but it acts as a proxy to disease-related functionality elsewhere. Elucidating the polymorphism across the whole HLA region enabled more refined disease associations, and may give clues to the potential disease function.

HLA-DRB1 alleles evolved from a series of duplications and recombinations.[22] Historical recombinations and duplications can cause similar genes to relocate, which in the case of HLA-DRB results in multiple genes, which have some functional homology but have distinct genomic context. Recombination can also result in a non-functioning pseudogene. The recombined genes and pseudogenes are inherited through evolution, which allows some allele-level polymorphism, but generally maintains the haplotype organization. The concepts of recombination and duplication also apply to other loci, such as HLA-DRA. We propose that loci that are currently defined as a single HLA gene may actually represent distinct gene lineages.

The traditional view of gene organization within the HLA-DR region is the existence of well-defined, and distinct genes, which exist in a small number of distinct organizations. HLA-DRA within non-human primates displays more variability[23,40] than what is known in humans, they have a similar HLA-DRA gene with distinct patterns of inheritance. We have shown that, like what has been observed in cynomolgus monkeys,[41] SNP patterns within HLA-DRA can present divisions in the haplotypes that were thought to represent a single lineage, demonstrating that these haplotype patterns represent their own distinct lineages. Future studies will explore the nature of these haplotype lineages, and how the haplotype frequencies relate to human ethnicities.

While HLA typing is the major consideration for SCT, Petersdorf *et al*. have shown that it is important to consider haplotypes in addition to individual HLA genotypes.[42] In retrospective studies, transplantations where haplotypes were matched in addition to genotypes, resulted in reduced numbers of cases of GVHD. A clinician must find a balance between disease recurrence and graft vs host disease, and knowledge of haplotypes gives important clues. This study has shown that SNPs within the introns and UTR of HLA-DRA can, in some cases, give clues towards the haplotypes of the related genes. The effects of haplotype matching on transplantation outcome are not fully understood, and analysis of the entire haplotype region allows more detailed understanding of the haplotype and it's relation to outcomes.

A long term goal for many HLA laboratories is to analyze the entire HLA region as a single haplotype sequence, rather than individual genes. Techniques to extract sequence from entire regions are being developed,[43] but present a new set of challenges in implementation. They depend on region-specific probes, and rely on secondary molecules, such as biotin and streptavidin, which can affect the behavior and sequencing results, especially in a nanopore sequencing setting. Interpretation of the region presents another challenge. HLA genes share significant homology, the HLA region presents many distinct patterns of gene organization due to recombinations.

Nevertheless, the HLA field is moving towards interpreting HLA as an inherited haplotype. Except in cases of recombinations and duplications, the unit of inheritance is not distinct genes, but an entire haplotype region. Interpreting HLA as a collection of randomly assembled, unrelated genes leaves gaps in our analysis of haplotype effects related to gene organization, transplantation, and disease associations. As an HLA community, it is valuable to represent, store and analyze the HLA region as a haplotype, allowing more refined and specific analysis of the interactions between genes and misunderstood haplotype effects, with the aim of improving our understanding of HLA and how it can affect a human's health and transplantation outcomes.

Acknowledgements

The authors thank Christel Meertens for her assistance in validation and submission of novel alleles. Thanks to Tom Jansen for assistance in sequencing and analysis and Laura de Rooij for the sequencing and enlightening 1000 Genomes Project analysis. Thanks to Professor Steven GE Marsh and James Robinson for insights on HLA-DRA and IPD-IMGT/HLA. Thanks to Martin Maiers for his feedback and discussion on the manuscript, and thanks to Diana van Bakel for her contributions to manuscript submission.

References

1. Leffler EM, Gao Z, Pfeifer S, *et al*. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science.* 2013;339(6127):1578.
2. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Med.* 2007;4(1):e8.
3. Robinson J, Guethlein LA, Cereb N, *et al.* Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genet.* 2017;13(6):e1006862.
4. The Genomes Project C, Auton A, Abecasis GR, *et al.* A global reference for human genetic variation. *Nature.* 2015;526:68.
5. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research.* 2015;43(Database issue):D423-D431.
6. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. *Nature Reviews Immunology.* 2018;18:325.
7. Morrison BA, Ucisik-Akkaya E, Flores H, Alaez C, Gorodezky C, Dorak MT. Multiple sclerosis risk markers in HLA-DRA, HLA-C, and IFNG genes are associated with sex-specific childhood leukemia risk. *Autoimmunity.* 2010;43(8):690-697.
8. Mansouri L, Messalmani M, Klai S, *et al*. Association of HLA-DR/DQ polymorphism with Alzheimer's disease. *Am J Med Sci.* 2015;349(4):334-337.
9. Gueant JL, Romano A, Cornejo-Garcia JA, *et al.* HLA-DRA variants predict penicillin allergy in genome-wide fine-mapping genotyping. *The Journal of allergy and clinical immunology.* 2015;135(1):253-259.
10. Lee HS, Yang SK, Hong M, *et al.* An intergenic variant rs9268877 between HLA-DRA and HLA-DRB contributes to the clinical course and long-term outcome of ulcerative colitis. *J Crohns Colitis.* 2018.
11. Kessal K, Liang H, Rabut G, *et al*. Conjunctival Inflammatory Gene Expression Profiling in Dry Eye Disease: Correlations With HLA-DRA and HLA-DRB1. *Frontiers in Immunology.* 2018;9(2271).
12. Clark PM, Chitnis N, Shieh M, Kamoun M, Johnson FB, Monos D. Novel and Haplotype Specific MicroRNAs Encoded by the Major Histocompatibility Complex. *Scientific reports.* 2018;8(1):3832-3832.
13. Clark PM, Monos DS. OR21 The novel functional role of HLA DRA and DRB5 encoded mirna transcripts. *Human immunology.* 2017;78:22.
14. Bontrop RE, Otting N, de Groot NG, Doxiadis GG. Major histocompatibility complex class II polymorphisms in primates. *Immunol Rev.* 1999;167:339-350.
15. Andersson G. Evolution of the human HLA-DR region. *Front Biosci.* 1998;3:d739-745.
16. Scholz EM, Marcilla M, Daura X, Arribas-Layton D, James EA, Alvarez I. Human Leukocyte Antigen (HLA)-DRB1*15:01 and HLA-DRB5*01:01 Present Complementary Peptide Repertoires. *Frontiers in Immunology.* 2017;8(984).
17. Mack SJ. A GENE FEATURE ENUMERATION APPROACH FOR DESCRIBING HLA ALLELE POLYMORPHISM. *Human immunology.* 2015;76(12):975-981.

18. den Dunnen JT, Dalgleish R, Maglott DR, *et al*. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation.* 2016;37(6):564-569.
19. Harton J, Jin L, Hahn A, Drake J. Immunological Functions of the Membrane Proximal Region of MHC Class II Molecules [version 1; peer review: 3 approved]. *F1000Research.* 2016;5(368).
20. Marsh SGE, Albert ED, Bodmer WF, *et al.* Nomenclature for factors of the HLA system, 2010. *Tissue Antigens.* 2010;75(4):291-455.
21. Marsh SGE, Parham P, Barber LD. 5 - HLA Class II Antigens and Alleles: Workshops and Nomenclature. In: Marsh SGE, Parham P, Barber LD, eds. *The HLA FactsBook*. London: Academic Press; 2000:26-36.
22. Doxiadis GGM, Hoof I, de Groot N, Bontrop RE. Evolution of HLA-DRB genes. *Molecular biology and evolution.* 2012;29(12):3843-3853.
23. Doxiadis GG, de Vos-Rouweler AJ, de Groot N, Otting N, Bontrop RE. DR haplotype diversity of the cynomolgus macaque as defined by its transcriptome. *Immunogenetics.* 2012;64(1):31-37.
24. Aarnink A, Estrade L, Apoil PA, *et al*. Study of cynomolgus monkey (Macaca fascicularis) DRA polymorphism in four populations. *Immunogenetics.* 2010;62(3):123-136.
25. Zerbino DR, Achuthan P, Akanni W, *et al*. Ensembl 2018. *Nucleic Acids Research.* 2018;46(D1):D754-D761.
26. Voorter CEM, Palusci F, Tilanus MGJ. Sequence-Based Typing of HLA: An Improved Group-Specific Full-Length Gene Sequencing Approach. In: Beksaç M, ed. *Bone Marrow and Stem Cell Transplantation*. New York, NY: Springer New York; 2014:101-114.
27. Wick R. Porechop Github. https://github.com/rrwick/Porechop. Accessed Sept 1, 2019.
28. Matern B. Nanopore Prospector Github. https://github.com/transplantation-immunology-maastricht/nanopore-prospector. Accessed Sept 1, 2019.
29. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002;24(7):881-892.
30. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011;12:2825-2830.
31. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update--a software pipeline for large-scale multilocus population genomics. *Tissue Antigens.* 2007;69 Suppl 1:192-197.
32. Matern BM, Groeneweg M, Voorter CEM, Tilanus MGJ. Saddlebags: A software interface for submitting full-length HLA allele sequences to the EMBL-ENA nucleotide database. *HLA.* 2018;91(1):29-35.
33. Vincent R, Louis P, Gongora C, Papa I, Clot J, Eliaou JF. Quantitative analysis of the expression of the HLA-DRB genes at the transcriptional level by competitive polymerase chain reaction. *The Journal of Immunology.* 1996;156(2):603-610.
34. Boegel S, Lower M, Bukur T, Sorn P, Castle JC, Sahin U. HLA and proteasome expression body map. *BMC Med Genomics.* 2018;11(1):36.
35. Voorter CE, Gerritsen KE, Groeneweg M, Wieten L, Tilanus MG. The role of gene polymorphism in HLA class I splicing. *Int J Immunogenet.* 2016;43(2):65-78.

36. di Iulio J, Bartha I, Wong EHM, *et al.* The human noncoding genome defined by genetic diversity. *Nat Genet*. 2018;50(3):333-337.
37. Hongming F, Tilanus M, Eggermond Mv, Giphart M. Reduced complexity of RFLP for HLA-DR typing by the use of a DRβ3'cDNA probe. *Tissue Antigens.* 1986;28(3):129-135.
38. Hollenbach JA, Oksenberg JR. The immunogenetics of multiple sclerosis: A comprehensive review. *J Autoimmun.* 2015;64:13-25.
39. Mack SJ, Udell J, Cohen F, *et al.* High resolution HLA analysis reveals independent class I haplotypes and amino-acid motifs protective for multiple sclerosis. *Genes Immun*. 2018.
40. Aarnink A, Estrade L, Apoil P-A, *et al*. Study of cynomolgus monkey (Macaca fascicularis) DRA polymorphism in four populations. *Immunogenetics.* 2010;62(3):123-136.
41. Blancher A, Aarnink A, Tanaka K, *et al*. Study of cynomolgus monkey (Macaca fascicularis) Mhc DRB gene polymorphism in four populations. *Immunogenetics.* 2012;64(8):605-614.
42. Petersdorf EW, Malkki M, Horowitz MM, Spellman SR, Haagenson MD, Wang T. Mapping MHC haplotype effects in unrelated donor hematopoietic cell transplantation. *Blood.* 2013;121(10):1896-1905.
43. Dapprich J, Ferriola D, Mackiewicz K, *et al*. The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC genomics.* 2016;17:486-486.

CHAPTER 7

7

Division of HLA-DRB1*13 haplotypes by extended HLA-DRA 3′ UTR polymorphism refines HLA-DRB1*13~HLA-DRB3~HLA-DQB1 haplotypes and gives clues to HLA-DR13 immunogenicity

B.M. Matern, T.I. Olieslagers, M. Groeneweg, M.G.J. Tilanus

Transplantation Immunology, Tissue Typing Laboratory, Maastricht University Medical Center, Maastricht, The Netherlands

Manuscript In Preparation

Abstract

Abstract

The IPD-IMGT/HLA database lists 2,696 HLA-DRB1 alleles, of which 380 (14%) are assigned as HLA-DRB1*13. HLA-DRB1*13 exists on haplotypes which have well-defined linkages with specific HLA-DRB3 and HLA-DQB1 alleles. Serotyping of HLA-DRB1*13 antigens is made difficult by a lack of distinguishing monospecific antibodies. HLA-DRB1*13 haplotypes have been previously shown to have reduced immunogenicity, providing a protective effect against both autoimmune disease and recurrent miscarriage. Analysis of amino acid polymorphism reveal that HLA-DRB1*13 has no amino acid patterns that uniquely characterize the allele group. The conservation of the exons of HLA-DRB1*13 is likely due to a selective pressure to maintain alleles with reduced immunogenicity. The divisions of the haplotypes by the HLA-DRA 3′ UTR polymorphism suggest that the HLA-DRB1*13 allele group is not completely understood, and genes and haplotypes that are currently considered to be individual lineages likely represent multiple distinct lineages. This study explores how HLA-DQB1~HLA-DRB1*13~HLA-DRB3~HLA-DRA haplotypes are extended and redefined by sequence polymorphism in the 3′ UTR of HLA-DRA, and how this haplotype-wide polymorphism relates to epitopes that define HLA-DR13 subtypes and the subsequent immunogenicity of the HLA-DR13 molecule.

1. Introduction

1.1 HLA

Human Leukocyte Antigen (HLA) is a set of genes on chromosome 6 which make up the human Major Histocompatibility Complex (MHC). HLA molecules have a critical role in the adaptive immune system, and matching of HLA alleles can reduce adverse side effects in stem cell[1] and solid organ[2, 3] transplantations. In standard HLA nomenclature,[4] HLA alleles are given specific and meaningful names. The names are divided into four colon-delimited fields. As a general rule, the first field indicates a distinction in the allele's serological equivalent, while the second, third, and fourth fields indicate differences in the allele's amino acid (2^{nd} field) or non-coding nucleotide sequence(3^{rd} and 4^{th} field). This nomenclature is useful for defining and grouping alleles by their serological or molecular characteristics, but HLA is rife with exceptions to these general rules. Counterexamples can be found in nearly all HLA loci,[5] but the most notable exception is in the HLA-DPB1 locus, where the lack of distinguishing serological typing means that individual HLA-DPB1 alleles are often assigned to their own individual allele group, with sequentially assigned group numbers. More pertinently, the HLA-DRB1*13 allele group contains individual allele sequences which share sequence homology, but do not share identical serological typing. Defining alleles based on sequence homology illustrates a shift away from serological typing towards molecular typing, and laboratories often discontinue serological typing once molecular typing techniques became available.[6]

The IPD-IMGT/HLA database[7] (release 3.38.0) currently lists 2,696 HLA-DRB1 alleles, making it the most polymorphic HLA class II locus. Of these alleles, 380 (14%) have been categorized as HLA-DRB1*13, the second highest allele count of the 13 HLA-DRB1 allele groups (HLA-DRB1*04 comprises 417 alleles). The nucleotide polymorphism is reflected in the amino acid polymorphism, and HLA-DRB1*13 contains at least 295 distinct protein sequences. Although they share sequence homology, specific HLA-DR13 alleles are known to have separate patterns of evolution due to historical recombinations.[8] In most cases these allele groups are meant to represent alleles with similar serological behavior, they allow for significant polymorphism in both the nucleotide and amino acid sequences, and a single variation in amino acid sequence can have strong effects on peptide binding.[9]

HLA-DR13 is known to be difficult to type serologically, due to a lack of monospecific antibodies and distinguishing antisera.[10] Distinguishing DR13 subtypes by serology is only feasible using multiple antibodies, and can therefore only be detected by comparing reactivities from a variety of antisera. Laboratories often use intricate panels of sera to test for the presence of expressed HLA-DR13 antigens. Even with well-defined serology panels, it can be a challenge to identify which expressed HLA molecule, or more specifically which epitope, is the target of an antibody. Attempts have been made to correlate HLA-DR13

serological typing with the polymorphism within the nucleotide sequence,[11] which has revealed some correlations with the sequence patterns and serological and cellular types. In addition, there have been attempts to assign serological specificities of specific HLA alleles based on their peptide sequences in silico.[12] The use of a neural network, combined with the known serology of well-categorized allele sequences has shown some success. Unfortunately, since new alleles with unknown serological types continue to be discovered, and since prediction models are trained based on known serological typings, neural network-based models have not been accepted as a complete model for determining serological types.

1.2 HLA Haplotypes

The MHC has evolved through a long history of recombinations and gene duplications.[13] An individual HLA gene may be formed by a translocation or duplication, or even a recombination of two separate HLA Loci.[8] These recombinations and duplications are reflected in the offspring,[14] creating an MHC region which contains distinct homologous genes, and polymorphism in the copy numbers of homologous genes. The separate genes are subjected to separate selective pressures, and therefore develop their own polymorphism, specific to their distinct evolutionary lineages.[15] This is especially apparent in the HLA-DR region.[16] Haplotypes which contain HLA-DRB1*13 are, like HLA-DRB1*03,11,12, and 14, marked by the presence of a HLA-DRB3 gene. HLA-DRB3 serves a similar function to HLA-DRB1 in that they both encode the beta subunit of the HLA-DR protein. Similarly, haplotypes with HLA-DRB1*04,07 or 09 alleles are marked by an additional HLA-DRB4 gene, haplotypes with HLA-DRB1*15 and 16 have an additional HLA-DRB5 gene, but haplotypes with HLA-DRB1*01, 08, and 10 do not have an additional expressed HLA-DRB gene. All HLA haplotypes are further marked by the presence of HLA-DRB pseudogenes, which have over time lost their function as expressed proteins, but share some evolutionary history with the expressed HLA-DRB genes. The presence of multiple copies of the beta subunit provides more variability in both the expressed HLA, and in the MHC haplotype as a whole.

Specific haplotype patterns have been conserved over a long evolutionary history,[17] resulting in alleles with longstanding linkages to alleles at another locus. The haplotype associations vary by ethnicity, but alleles or groups have well-known associations with nearby HLA genes. HLA-DRB1*13 has well-documented linkages to HLA-DQB1 and HLA-DRB3, for example 3.6% of haplotypes in caucasians carry HLA-DQB1*06:03~HLA-DRB1*13:01~HLA-DRB3*01:01 alleles.[18] These associations imply that many class II proteins are commonly expressed together. Due to the strongly linked inherited haplotypes, results from serological tests can be difficult to specifically define, due to ambiguities in which loci or epitope is responsible for the antibody interactions.

We recently included HLA-DRA in studies of haplotypes in the HLA-DR~HLA-DQ region, which showed that common haplotype patterns are refined, and in some cases split by HLA-DRA sequences.[19] This is especially apparent in the HLA-DRB1*13 haplotypes, which demonstrated the most drastic divisions in haplotype patterns. This study explores and redefines HLA-DQB1~HLA-DRB1*13~HLA-DRB3~HLA-DRA extended haplotypes defined by sequence polymorphism in the 3′ UTR of HLA-DRA, even the region downstream of the gene as represented in IPD-IMGT/HLA, and explores how this haplotype-wide polymorphism relates to epitopes that define HLA-DR13 subtypes and how they affect the immunogenicity of the HLA-DR13 molecule.

2. Material and Methods

2.1 Sequencing and Haplotype Analysis

Sequencing and analysis of HLA-DRA has been described previously.[19] For specific analysis of HLA-DRB1*13 haplotypes, 37 samples which had available DNA and which had previously typed positive for HLA-DRB1*13 were selected. The samples were previously typed to at least 2 fields for HLA-DRB1, HLA-DRB3, and HLA-DQB1 by Sanger SBT.[20] HLA-DRA was amplified using primers that span an 8.7kb region, covering the 5.7kb HLA-DRA region represented in IPD-IMGT/HLA, and a 2.9kb of extended UTR sequence. These amplicons were full-length sequenced using 1D^2 MinION sequencing. Novel HLA-DRA polymorphism within the IPD-IMGT/HLA region was confirmed using allele-specific Sanger SBT. Haplotype patterns were elucidated by phasing the polymorphism of HLA-DRA against HLA-DRB1, HLA-DRB3,4,5, and HLA-DQB1 with the expectation maximization algorithm[21] as implemented in Pypop (v0.7.0).[22]

2.2 Epitope Analysis

Alignment of amino acid sequences were obtained from the IPD-IMGT/HLA database[7] (release 3.38.0). Alleles from each HLA-DRB1 group were selected for comparison, (Table 1) based on the availability of distinct full-length sequences. Class II HLA allele sequences in IPD-IMGT/HLA must have an available exon 2 sequence,[4] but full-length amino acid sequences are not available for every allele, with many alleles missing sequence from exons 1,4,5,or 6. The first four allele sequences with nearly full length sequences were selected for analysis.

HLA-DRB1*13 subtype amino acid differences were determined using alignments of the amino acid sequences from the distinct HLA-DRB1*13 amino acid sequences from IPD-IMGT/HLA. These alignments were analyzed with specific focus on amino acid differences within the HLA-DRB1*13 allele group. Differences in the amino acid sequences were identified and summarized, and were phased against the HLA-DQB1~HLA-DRB1*13~HLA-DRB3~HLA-DRA haplotype patterns shown previously.[19]

2.3 Polymorphic Index

Polymorphic positions were identified from alignments of genomic sequence for all available HLA-DRB1 sequences in IPD-IMGT/HLA (release 3.38.0). Polymorphic index[23] (PI) is calculated by dividing the count of polymorphic positions within a region by the length of a region, and serves as an estimation of how polymorphic a region is. PI was calculated for alleles within all HLA-DRB1 groups, both for exon 2 alone and for the combined intron and UTR sequences.

01:01:01:01	**07:04**	**10:19**	**13:04**
01:02:01:01	**07:09**	**10:25**	**14:01:01**
01:03:01	**07:27**	**11:01:01:01**	**14:02:01:01**
01:13	**08:01:01**	**11:02:01:01**	**14:03:01**
03:01:01:01	**08:02:01:01**	**11:03:01**	**14:04:01**
03:02:01	**08:03:02:01**	**11:04:01**	**15:01:01:01**
03:04:01	**08:04:01**	**12:01:01:01**	**15:02:01:01**
03:06	**09:01:02:01**	**12:02:01:01**	**15:03:01:01**
04:01:01:01	**09:21**	**12:10**	**15:04**
04:02:01	**09:31**	**12:17**	**16:01:01**
04:03:01:01	**09:32**	**13:01:01:01**	**16:02:01:01**
04:05:01:01	**10:01:01:01**	**13:02:01:01**	**16:04:01**
07:01:01:01	**10:03**	**13:03:01**	**16:07**

Table 1. HLA-DRB1 allele sequences which were included for determination of characteristic epitopes. For each allele group, four alleles were selected and aligned using IPD-IMGT/HLA[7] web interface. The second field of HLA nomenclature[4] represents amino acid differences, and alleles with unique amino acid sequence were selected, starting with the lowest second-field identifiers. Some alleles were not used in analysis based on unavailable amino acid sequence data.

HLA-DRB1 Allele Group	Characteristic AAs #	Polymorphic Index (Exon 2)	Polymorphic Index (Introns & UTRs)
01	**3**	0.36	0.00087
03	**3**	0.45	0.00084
04	**5**	0.53	0.00198
07	**6**	0.33	0.00072
08	**3**	0.29	0.00544
09	**5**	0.19	0.00059
10	**7**	0.14	0.00024
11	**1**	0.52	0.00191
12	**4**	0.26	0.01943
13	**0**	0.45	0.00617
14	**0**	0.36	0.00868
15	**2**	0.49	0.00165
16	**0**	0.23	0.00048

Table 2. Characteristic amino acids and Polymorphic Index of HLA-DRB1 allele groups.
Characteristic amino acids are polymorphic positions that consistently show the same amino acid for all alleles within a group, but are not found in another allele group. Amino acid polymorphism may function as epitope differences, which may determine differences in amino acid interaction and differences in serological behavior. Characteristic amino acids are shown alongside calculated Polymorphic Index, which gives a rough estimate of the polymorphism of a gene feature region. HLA-DRB1*13 is not noticeably different than other allele groups, and does not fully explain the enigmatic serology and immunogenicity of HLA-DRB1*13.

HLA-DQB1	HLA-DRB1	32	37	47	57	71	86	HLA-DRB3	HLA-DRA	N
06:03	13:01	H	N	F	D	E	V	3*01:01	01:02:02:05/08	**16**
06:03	13:01	H	N	F	D	E	V	3*02:02	01:01:01:03/06/07	**11**
06:04	13:01	H	N	F	D	E	V	3*01:01	01:02:02:08	**1**
06:03	13:01	H	N	F	D	E	V	3*03:01	01:02:02:08	**1**
06:04	13:02	H	N	F	D	E	G	3*03:01	01:02:02:01/03/08	**10**
06:09	13:02	H	N	F	D	E	G	3*03:01	01:02:02:04	**2**
03:01	13:03	Y	Y	Y	S	K	G	3*01:01	01:02:02:06	**2**

Table 3. HLA-DQB1~HLA-DRB1*13~HLA-DRB3~HLA-DRA Haplotype Patterns
The general HLA-DRB1 haplotype patterns identified in Matern *et al.*[19] The HLA-DRB1*13 subtype haplotypes are marked by differences in the associated HLA-DRA and HLA-DRB3 alleles. Multiple HLA-DRA alleles exist on the same haplotypes, which is indicated by a / in the fourth-field typing. HLA-DRB1*13 subtypes are distinguished by polymorphic amino acid positions, also shown. Each of these polymorphic positions are located within HLA-DRB1 exon 2, and no other region was found to contain amino acid differences. These amino acid polymorphisms indicate epitope differences which may affect antibody interaction and immunogenicity.

3. Results and Discussion

3.1 Polymorphism in HLA-DRB1*13

Characteristic amino acids were defined as polymorphic amino acid positions that have a consistent amino acid for every common allele within a HLA-DRB1 group (*e.g.* HLA-DRB1*13), which is not found in other groups at the same locus. Characteristic amino acids were assigned to their corresponding HLA-DRB1 group (Supplementary Table 1) and counted and summarized (Table 2). Characteristic amino acids were identified for 10 of the 13 HLA-DRB1 allele groups, but none were identified for HLA-DRB1*13,14, and 16. Every amino acid polymorphism in molecules within these three allele groups can also be identified within another HLA-DRB1 allele group. Epitopes, the immunogenic components of a molecule that act as recognizable targets of specific antibodies,[24] have structure which is determined by the amino acid sequence. The lack of characteristic amino acid sequences which define these allele group therefore indicates a lack of characteristic epitopes. Comparisons of the subtypes of HLA-DRB1*13 did however reveal several amino acid differences between individual HLA-DRB1*13 alleles (Table 3). Every identified allele-distinguishing amino acid lies within exon 2, which encodes the part of the HLA molecule responsible for antigen presentation, suggesting epitope differences that may cause distinct serological behavior between alleles within the same group. These findings should be interpreted in the context of the available data. As a general rule, all HLA class II sequences in IPD-IMGT/HLA must have an available exon 2 sequence, but several groups did not have four available full-length amino acid sequence. Certain alleles (*e.g.* 07:27, 10:03, 10:19, 10:25) were missing sequence representing some combination of exons 1,4,5,6, but none are missing exons 2 or 3.

Polymorphic index was measured in order to compare the polymorphism of HLA-DRB1*13 relative to the other HLA-DRB1 allele groups, as shown in Table 2. The comparisons are focused both on exon 2, and for the combined intron and UTR sequences. The polymorphic index for the HLA-DRB1*13 allele group was not noticeably higher or lower than other groups. This suggests that polymorphism within the HLA-DRB1*13 allele group does not alone explain the differences in serological behavior, but more insight can be derived from looking at alleles in the context of HLA-DR~HLA-DQ haplotypes.

HLA-DRB1*13 has been correlated with protection from autoimmune disease[25, 26] and recurrent miscarriage[27]. The mechanisms of the protective effect have not been completely elucidated. The correlation is not specific to a single haplotype or ethnicity, suggesting that the presence of the specific HLA-DR13 molecule, rather than the associated haplotype, provides the effect. It may be that the HLA-DR13 molecule interacts preferentially with T-regulatory cells, providing an immunosuppressant effect. [25] Bettencourt *et. al* suggest the effect could be explained by a preferential deletion

of cells expressing the HLA-DR13 molecules during thymic selection, resulting in a less immunogenic repertoire of T-cells.[26] Some studies suggest that HLA-DR13 may not express the serological determinants necessary to start an immune response,[11] implying a lack of a specific and consistent epitope to interact with an antibody. Regardless of the specific mechanism, the protective effect is related to the HLA-DR13 molecule, which is determined predominantly by the amino acid sequence. The lack of characteristic HLA-DRB1*13 epitope sequences, combined with the polymorphism in the antigen presentation epitopes likely explain the lack of monospecific HLA-DRB1*13 antibodies and antisera, and gives clues about the reduced immunogenicity and protection from autoimmune disease.

3.2 3′ UTR indicates Haplotypes and Expression

The 3′ untranslated region of HLA genes is increasingly found to indicate haplotype and genomic context information, and polymorphism within this region has been used to identify the presence or absence of genes.[28, 29] The 3′ UTR also contains polymorphism which is indicative of expression of HLA molecules, whether soluble[30] or cell-surface HLA. The well-known rs9277534 SNP in the 3′ UTR of HLA-DPB1 has been correlated with expression levels of the DP molecule on the cell surface,[31-33] and is subsequently becoming a consideration in stem cell transplantations. The 3′ UTR also contains one or more polyadenylation sites,[34] and polymorphism within this region can alter the structure of mRNA transcripts. 3′ polymorphism has also been linked to disease prevalence[35] and is increasingly being studied to determine links with infection or parasite susceptibility.[36]

The 3′ UTR of HLA-DRA is marked by interesting sequence features, splicing behavior, and patterns of polymorphism. The 3′ UTR region represented in IPD-IMGT/HLA[7] (release 3.38.0) is 1394 bp in length, and contains at least 19 SNPs and 2 major sequence deletions. These two deletions have intriguing sequence patterns. (Figure 1) The 34bp deletion sequence (HLA00662.1:g.4625_4658del)[37] is located within a sequence motif repeat, and the deleted sequence are one in a series of three repeats. These repeat sequences are flanked by several regions of 4-5 base A-nucleotide repeats, which displays patterns reminiscent of a historical sequence duplication or ALU region insertion.[38] Furthermore, the smaller 8bp deletion (HLA00662.1:g.4910_4917del) forms a perfect reverse-complement with the sequence 7 bases downstream, possibly indicating a hairpin loop structure or sites of microRNA interaction. Interestingly, both of these sequence deletions lie within an intron region, which is spliced out from the HLA-DRA 3′ UTR mRNA sequence. The region also contains at least two polyadenylation sites,[39] which may provide alternative polyadenylation behavior. In addition to the polymorphism represented in IPD-IMGT/HLA, there are at least 28 additional SNPs in the 1.9 kb of downstream 3′

sequence.[19] These observations indicate unexplored sequence function within the 3′ UTR of HLA-DRA, and demonstrate that this region is significant, both in its function and role in identifying haplotypes and genomic context.

3.3 HLA-DRB1*13 Haplotypes

Different HLA-DRB1*13 alleles exist within distinct haplotype patterns, shown in Table 3. The haplotype patterns are shown phased against the HLA-DRB1*13 amino acid differences, suggesting that epitope differences are linked with haplotype patterns. HLA-DRB1*13 haplotypes are marked by the presence of an additional expressed HLA-DRB3 gene,[40] and they possess characteristic HLA-DRA, HLA-DRB3 and HLA-DQB1 alleles. Polymorphism within the 3′ UTR of HLA-DRA divides and extends the known haplotypes, providing refinement in our understanding of the HLA-DR~HLA-DQ region. A notable example is HLA-DQB1*06:03~HLA-DRB1*13:01~HLA-DRB3*01:01 haplotype, which contains multiple HLA-DRA alleles (HLA-DRA*01:02:02:05 and HLA-DRA*01:02:02:08). Although these alleles differ by only a single 5′ UTR SNP in the 5.7kb region represented in IPD-IMGT/HLA, they differ significantly in the gene's downstream 3′ UTR region.[19]

Polymorphism that is outside the represented gene sequence can affect the expression of HLA molecules,[32, 36] and in the case of HLA-DRA affect the haplotype definition. HLA-DRA*01:02:02:08 lies on two haplotypes with distinct HLA-DRB1, HLA-DRB3 and HLA-DQB1 alleles, which cannot be distinguished by the 5711bp HLA-DRA sequence region alone. A C/T SNP (rs3129890), approximately 1200bp downstream of the IPD-IMGT/HLA region, is correlated with the haplotype divisions, where the T allele indicates the HLA-DRB1*13:01 haplotype, and the C allele indicates the HLA-DRB1*13:02 haplotype. (*i.e.* HLA-DRA*01:02:02:08~**rs3129890C**~HLA-DRB3*03:01~HLA-DRB1*13:02~HLA-DQB1*06:04 and HLA-DRA*01:02:02:08~**rs3129890T**~HLA-DRB3*01:01~HLA-DRB1*13:01~HLA-DQB1*06:03) Future studies will explore the strength of this linkage, and how it relates to population differences, but this provides further evidence that there is value in analysis of the extended full length gene and surrounding sequence, for HLA-DRA as well as the entire MHC.

3.4 HLA-DRB1*13 Gene Definition

Most definitions of a gene are based on a protein-encoding region at a specific locus within a chromosome. The locus is described both by it's relative position within a chromosome, and by how it relates to nearby genes and polymorphism. Determining the cutoff between allelic and gene distinctions is a controversial topic,[41] but it is valuable to define gene loci based on the gene's function and context within chromosomal haplotypes. Divisions of HLA-DRB1*13 alleles onto distinct haplotypes supports the hypothesis that what is currently known as HLA-DRB1*13 represent different genes, which have over time evolved in their own separate lineages.

HLA genes in this region have high levels of polymorphism, in allelic polymorphisms within specific gene loci, and in the gene copy number variations, and in the context of surrounding intergenic polymorphism. This demonstrates that the HLA-DR~HLA-DQ region is more flexible than is commonly known. Newly identified divisions in the known haplotypes of the HLA-DR~HLA-DQ region indicate a need for a more flexible understanding of the gene organization within this region. The identification of HLA-DRB3,4, and 5 as separate genes is generally accepted, but there is value in reanalysis of the definitions of the genes, which may need to be further extended. This study observes distinct divisions in HLA-DRB1 genes, but HLA-DQB1, HLA-DRA and other loci are inherited on distinct haplotypes as well, and may also represent multiple distinct genes with conserved sequence patterns. If this is indeed the case, our understanding of HLA haplotypes, and our assumptions about the full-length MHC and the intergenic sequence polymorphism between the HLA genes, require reinterpretation and further elucidation. First attempts to address the full-length MHC regions have been realized already by capture technology[42] and will continue to improve and give insight into patterns of full MHC polymorphism and HLA gene organization.

3.5 Haplotypes in SCT

In addition to elucidating evolutionary lineages, defining MHC haplotypes provides clinical value. Matching of high-resolution HLA alleles is the most important consideration for improving outcomes for stem cell transplantations(SCT).[1, 43] A major focus is on class II matching based on HLA-DRB1, but lower-expressed[44] loci such as HLA-DQB1, HLA-DRB3,4,5,[45, 46] or HLA-DPB1[31] are also important considerations. In additional to individual loci, considering phased haplotypes of HLA loci provides further outcome improvements,[47, 48] possibly due to the heterodimer interactions of multiple HLA loci within the MHC, or additional untyped genes or SNPs that are linked to the typed sequence polymorphism, or behavior of microRNA sequences encoded on the same haplotype.[49] The importance of donor selection based on phased haplotypes is further confirmed by the emergence of haploidentical transplantations,[50-52] where a single phased HLA haplotype is matched between patient and donor, while disregarding mismatches on the second haplotype. The mechanism of the haplotype effect is not understood, but is related to interactions between both HLA and non-HLA factors. Regardless of the mechanism, outcomes involving matched haplotypes provide more favorable outcomes than those without haplotype matching.

This study has shown that 3′ UTR polymorphism of HLA-DRA extends currently extend known HLA-DR~HLA-DQ haplotypes. These newly defined extended MHC haplotype patterns are shown to be more flexible than commonly assumed, and may have important considerations in matching for transplantations. The value of furthering our understanding of HLA haplotypes is apparent in haploidentical transplantation, and confirms that there

is added value in sequencing and studying the intergenic regions and extending the regions that are represented in standard databases. The lack of characteristic epitopes in HLA-DRB1*13 clarifies the allele group's enigmatic behavior and serological typing and gives clues to the reduced immunogenicity of HLA-DRB1*13. This evidence indicates a need to reconsider current understandings of HLA genes and their relationship with MHC haplotypes.

Acknowledgements

Thanks to the Christel Meertens, Dominique Pellaers, and Simone van der Linden for assistance with sequencing and typing of HLA-DRB3. Thanks also to the secretarial staff for assistance with manuscript submission.

References

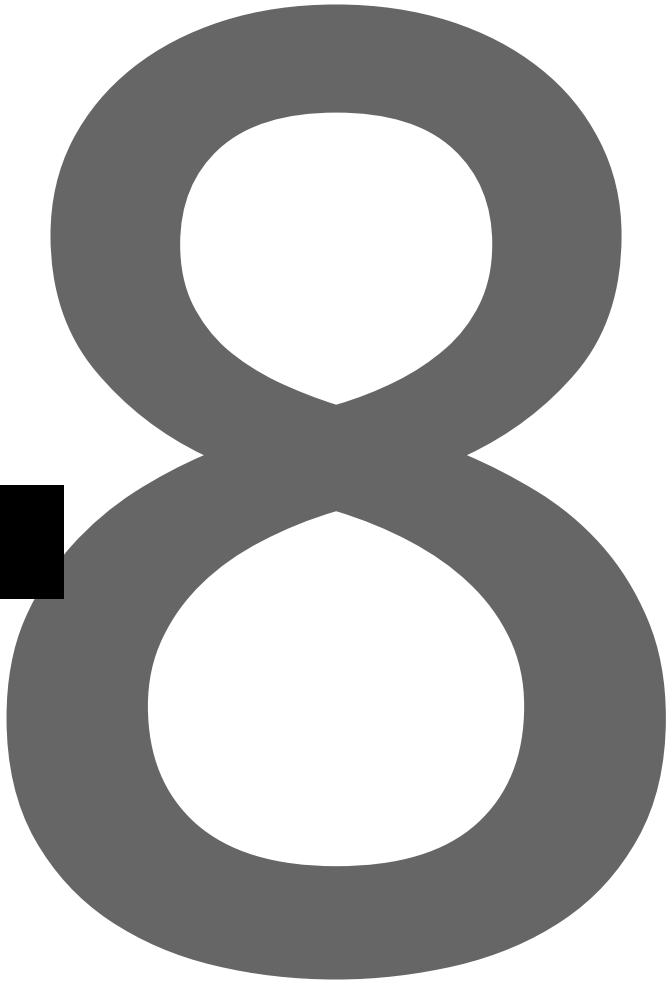
1. Mayor NP, Hayhurst JD, Turner TR, *et al*. Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. *Biology of Blood and Marrow Transplantation.* 2019;25(3):443-450.
2. Susal C, Opelz G. Current role of human leukocyte antigen matching in kidney transplantation. *Curr Opin Organ Transplant.* 2013;18(4):438-444.
3. Oertel M, Berr F, Schroder S, *et al*. Acute rejection of hepatic allografts from HLA-DR13 (Allele DRB1(*)1301)-positive donors. *Liver Transpl.* 2000;6(6):728-733.
4. Marsh SGE, Albert ED, Bodmer WF, *et al*. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens.* 2010;75(4):291-455.
5. Holdsworth R, Hurley CK, Marsh SGE, *et al*. The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. *Tissue Antigens.* 2009;73(2):95-170.
6. Erlich HA, Opelz G, Hansen J. HLA DNA typing and transplantation. *Immunity.* 2001;14(4):347-356.
7. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research.* 2015;43(D1):D423-D431.
8. Lee KW, Johnson AH, Hurley CK. Two divergent routes of evolution gave rise to the DRw13 haplotypes. *J Immunol.* 1990;145(9):3119-3125.
9. Hurley CK, Steiner N. Differences in peptide binding of DR11 and DR13 microvariants demonstrate the power of minor variation in generating DR functional diversity. *Human Immunology.* 1995;43(2):101-112.
10. Schreuder GMT, Gebuhrer L, Lepage V, *et al*. Antigen Society #25 Report DRw6, DRw13, DRw14). Paper presented at: Immunobiology of HLA; 1989//, 1989; New York, NY.
11. Bosch ML, Tilanus MGJ, Giphart MJ. HLA-DRw6: A Molecular Approach. Paper presented at: Histocompatibility Testing 1984; 1984//, 1984; Berlin, Heidelberg.
12. Maiers M, Schreuder GMT, Lau M, *et al*. Use of a neural network to assign serologic specificities to HLA-A, -B and -DRB1 allelic products. *Tissue Antigens.* 2003;62(1):21-47.
13. Cullen M, Noble J, Erlich H, *et al*. Characterization of recombination in the HLA class II region. *Am J Hum Genet.* 1997;60(2):397-407.
14. Askar M, Madbouly A, Zhrebker L, *et al*. HLA Haplotypes In 250 Families: The Baylor Laboratory Results And A Perspective On A Core NGS Testing Model For The 17th International HLA And Immunogenetics Workshop. *Human Immunology.* 2019;80(11):897-905.
15. Doxiadis GGM, Hoof I, de Groot N, Bontrop RE. Evolution of HLA-DRB genes. *Molecular biology and evolution.* 2012;29(12):3843-3853.
16. Andersson G. Evolution of the human HLA-DR region. *Front Biosci.* 1998;3:d739-745.
17. Degli-Esposti MA, Leaver AL, Christiansen FT, Witt CS, Abraham LJ, Dawkins RL. Ancestral haplotypes: conserved population MHC haplotypes. *Human Immunology.* 1992;34(4):242-252.

18. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human Immunology.* 2013;74(10):1313-1320.
19. Matern BM, Olieslagers TI, Voorter CEM, Groeneweg M, Tilanus MGJ. Insights into the polymorphism in HLA-DRA and its evolutionary relationship with HLA haplotypes. *HLA.* 2019.
20. Voorter CEM, Palusci F, Tilanus MGJ. Sequence-Based Typing of HLA: An Improved Group-Specific Full-Length Gene Sequencing Approach. In: Beksaç M, ed. *Bone Marrow and Stem Cell Transplantation.* New York, NY: Springer New York; 2014:101-114.
21. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological).* 1977;39(1):1-22.
22. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update – a software pipeline for large-scale multilocus population genomics. *Tissue Antigens.* 2007;69(s1):192-197.
23. Gerritsen KE, Groeneweg M, Meertens CM, Voorter CE, Tilanus MG. Full-length HLA-DRB1 coding sequences generated by a hemizygous RNA-SBT approach. *Tissue Antigens.* 2015;86(5):333-342.
24. Duquesnoy RJ. Human leukocyte antigen epitope antigenicity and immunogenicity. *Curr Opin Organ Transplant.* 2014;19(4):428-435.
25. Furukawa H, Oka S, Tsuchiya N, *et al*. The role of common protective alleles HLA-DRB1*13 among systemic autoimmune diseases. *Genes Immun.* 2017;18(1):1-7.
26. Bettencourt A, Carvalho C, Leal B, *et al*. The Protective Role of HLA-DRB1*13 in Autoimmune Diseases. *Journal of immunology research.* 2015;2015:948723-948723.
27. Meuleman T, Lashley LE, Dekkers OM, van Lith JM, Claas FH, Bloemenkamp KW. HLA associations and HLA sharing in recurrent miscarriage: A systematic review and meta-analysis. *Hum Immunol.* 2015;76(5):362-373.
28. Hongming F, Tilanus M, Eggermond Mv, Giphart M. Reduced complexity of RFLP for HLA-DR typing by the use of a DRβ3′cDNA probe. *Tissue Antigens.* 1986;28(3):129-135.
29. Bontrop RE, Broos LAM, Pham K, Bakas RM, Otting N, Jonker M. The chimpanzee major histocompatibility complex class II DR subregion contains an unexpectedly high number of beta-chain genes. *Immunogenetics.* 1990;32(4):272-280.
30. Craenmehr MHC, Haasnoot GW, Drabbels JJM, *et al*. Soluble HLA-G levels in seminal plasma are associated with HLA-G 3′ UTR genotypes and haplotypes. *HLA.* 2019;94(4):339-346.
31. Petersdorf EW, Malkki M, O'hUigin C, *et al*. High HLA-DP Expression and Graft-versus-Host Disease. *New England Journal of Medicine.* 2015;373(7):599-609.
32. Morishima S, Shiina T, Suzuki S, *et al*. Evolutionary basis of HLA-DPB1 alleles affects acute GVHD in unrelated donor stem cell transplantation. *Blood.* 2018;131(7):808-817.
33. Klasberg S, Lang K, Gunther M, *et al*. Patterns of non-ARD variation in more than 300 full-length HLA-DPB1 alleles. *Hum Immunol.* 2019;80(1):44-52.
34. Briata P, Radka SF, Sartoris S, Lee JS. Alternative splicing of HLA-DQB transcripts and secretion of HLA-DQ beta-chain proteins: allelic polymorphism in splicing and polyadenylation sites. *Proceedings of the National Academy of Sciences of the United States of America.* 1989;86(3):1003-1007.

35. Svendsen SG, Nilsson LL, Djurisic S, *et al*. Extended HLA-G haplotypes in patients with age-related macular degeneration. *HLA*. 2018.
36. Sonon P, Gomes RG, Brelaz-de-Castro MCA, *et al*. Human leukocyte antigen-G 3′ untranslated region polymorphism +3142G/C (rs1063320) and haplotypes are associated with manifestations of the American Tegumentary Leishmaniasis in a Northeastern Brazilian population. *Human Immunology*. 2019;80(11):908-916.
37. den Dunnen JT, Dalgleish R, Maglott DR, *et al*. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*. 2016;37(6):564-569.
38. Deininger P. Alu elements: know the SINEs. *Genome Biology*. 2011;12(12):236.
39. Liu H, Han H, Li J, Wong L. An in-silico method for prediction of polyadenylation signals in human sequences. *Genome Inform*. 2003;14:84-93.
40. Marsh SGE, Parham P, Barber LD. 3 - The Organization of HLA Genes Within the HLA Complex. In: Marsh SGE, Parham P, Barber LD, eds. *The HLA FactsBook*. London: Academic Press; 2000:7-13.
41. Andersson G, Andersson L, Larhammar D, Rask L, Sigurdardottir S. Simplifying genetic locus assignment of HLA-DRB genes. *Immunol Today*. 1994;15(2):58-62.
42. Dapprich J, Ferriola D, Mackiewicz K, *et al*. The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC genomics*. 2016;17:486-486.
43. Dehn J, Spellman S, Hurley CK, *et al*. Selection of unrelated donors and cord blood units for hematopoietic cell transplantation: guidelines from the NMDP/CIBMTR. *Blood*. 2019;134(12):924-934.
44. Boegel S, Lower M, Bukur T, Sorn P, Castle JC, Sahin U. HLA and proteasome expression body map. *BMC Med Genomics*. 2018;11(1):36.
45. Fernández-Viña MA, Klein JP, Haagenson M, *et al*. Multiple mismatches at the low expression HLA loci DP, DQ, and DRB3/4/5 associate with adverse outcomes in hematopoietic stem cell transplantation. *Blood*. 2013;121(22):4603-4610.
46. Ducreux S, Dubois V, Amokrane K, *et al*. *HLA-DRB3/4/5 mismatches are associated with increased risk of acute GVHD in 10/10 matched unrelated donor hematopoietic cell transplantation*. Vol 932018.
47. Petersdorf EW, Malkki M, Horowitz MM, Spellman SR, Haagenson MD, Wang T. Mapping MHC haplotype effects in unrelated donor hematopoietic cell transplantation. *Blood*. 2013;121(10):1896-1905.
48. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Med*. 2007;4(1):e8.
49. Clark PM, Chitnis N, Shieh M, Kamoun M, Johnson FB, Monos D. Novel and Haplotype Specific MicroRNAs Encoded by the Major Histocompatibility Complex. *Scientific reports*. 2018;8(1):3832-3832.
50. Fabricius WA, Ramanathan M. Review on Haploidentical Hematopoietic Cell Transplantation in Patients with Hematologic Malignancies. *Adv Hematol*. 2016;2016:5726132.

51. Gao L, Zhang C, Gao L, *et al*. Favorable outcome of haploidentical hematopoietic stem cell transplantation in Philadelphia chromosome-positive acute lymphoblastic leukemia: a multicenter study in Southwest China. *J Hematol Oncol.* 2015;8:90.
52. Wang Z, Zheng X, Yan H, Li D, Wang H. Good outcome of haploidentical hematopoietic SCT as a salvage therapy in children and adolescents with acquired severe aplastic anemia. *Bone Marrow Transplant*. 2014;49(12):1481-1485.

CHAPTER 8



Polymorphism clustering of the 21.5kb DPA-promoter-DPB region reveals novel extended full length haplotypes.

L. Truong[1,2], B.M. Matern[3], M. Groeneweg[3], L. D'Orsogna[1,2], P. Martinez[1,2], M.G.J. Tilanus[3], D. De Santis[1,2]

[1] Clinical Immunology, PathWest, Fiona Stanley Hospital, Perth, Australia
[2] School of Medicine, The University of Western Australia, Perth, Australia
[3] Transplantation Immunology, Tissue Typing Laboratory, Maastricht University Medical Center, Maastricht, the Netherlands

Manuscript Submitted (HLA)

Abstract

Abstract

DPB1 and DPA1 genes share the same promoter region. Single-nucleotide polymorphisms (SNPs) within the regulatory regions of DP have been reported. This study hypothesizes that by including the SNPs in the promoter region of DP, extended haplotypes are defined and promoter polymorphism is more extensive than what is currently reported.

In order to identify the SNPs in the region of interest, the entire DP region spanning 21.5 kb was amplified in 3 separate long-ranged PCR reactions. A DNA panel consisting of 100 samples representing a broad range of DPB1 alleles were amplified and sequenced using a dual sequencing strategy. BAM alignments were generated and the mapped sequence alignments were analysed using IGV.

A total of 76 SNPs were identified and SNPs were clustered into 12 SNP-linked haplotypes. Multiple sequence alignment of promoter sequences indicated four distinct lineages within the connective region (CR) between two genes. The relationship between DPA1, CR, DPB1 and amino acid motifs was strictly related to HV1 and HV6 with minor exceptions. Of 12 promoter haplotypes, DPB1 alleles observed with ProDP-4 were in complete linkage with HV1/2/5/6, rs9277534G SNP, and highly immunogenic TCE group. Multiple extended haplotypes of different intronic subtypes of the same DPB1 alleles were also unravelled.

This new view on the full region haplotype shows the relation of polymorphism, genes and alleles, and provides a basis for future functionality related nomenclature. Lastly, the novel clustering of the DP extended haplotype warrants for future investigations of DP haplotype matching in the outcome of HSCT.

1. Introduction

Since the discovery of white blood cell antigen by Professor Jean Dausset, Professor Jon van Rood and Rose Payne in 1954,[1-3] there has been extensive scientific research into the functionality of genetic diversity in the HLA (HU-1 and LA) system. The HLA gene family is known as the most polymorphic system not only in the major histocompatibility complex (MHC) but also in the entire human genome.[4] The extreme polymorphism in HLA system is illustrated by an extraordinary number of 25,756 alleles that have been submitted to the IPD-IMGT/HLA database (release v3.38 – http://www.ebi.ac.uk/imgt/hla).[5]

DPA1 and DPB1 are part of HLA class II genes. Together, they encode for DP αβ-chain heterodimers that are expressed on the cell surface of antigen presenting cells. DPA1 and DPB1 have low linkage disequilibrium (LD) with other class II loci, however, the LD between DPA1 and DPB1 loci is strong.[6] Interestingly, DPB1 and DPA1 are oriented in opposite directions.[7] Therefore, the region between the start codons, which is approximately 2.5 kb in length, contains the promoter and transcription regulatory regions of both DPA1 and DPB1 genes. Early investigations into genetic evidence for the regulation of DP expression demonstrated that DPA1 and DPB1 are inducible by interferon ɣ (IFN-ɣ) and the responsible region has been mapped to its promoter. In 1990, Nezu *et al* assessed the region between -152 and -126 of the upstream sequence of DPB1 gene and concluded that this region of 25 bp contains a critical IFN-ɣ-responsive element.[8] Sugawara *et al* (1992) reported the corresponding region for IFN-ɣ inducibility in DPA1 was localized to 27 bp between -55 and -81 including the Y-box element of DPA1 promoter.[9]

There is previous evidence of polymorphism in the flanking region between DP loci. Varney *et al* (1999) identified 8 single nucleotide polymorphisms (SNPs) 402 bp upstream from exon 1 of DPB1 gene.[10] In one particular SNP haplotype named DP-PRO4, there were three G/A transitions in the highly conserved X_1, Y and W′ box within the promoter of DPB1 gene. A following study by Liu *et al* (2005) characterized 760 bp upstream from exon 1 of DPA1 gene and identified 21 SNPs in five Chinese ethnic populations.[11] The data from this study unveiled a SNP density of one SNP per 36 bp in this region, which was much denser than the average level of one SNP per 1,900 bp in the human genome.[12] These studies highlight the presence of polymorphism in the promoter region at high density and in LD with DPA1 and DPB1 alleles, however, the pattern of inheritance have not yet been fully elucidated.

As of October 2019, there are over 1,519 DPB1 and 161 DPA1 alleles described in the IPD-IMGT/HLA database and more are being identified given that the two genes are not traditionally typed or included in donor selection strategies for haematopoietic stem cell transplantation (HSCT). Several studies had attempted to investigate the clinical significance of DPB1 disparity in the HSCT and three models were proposed for DPB1 matching. Zino *et al* (2004)

provided the first description of structural T-cell epitope (TCE) matching model,[13] while Thus *et al* (2016) described a DPB1 matching model based on the numbers of Predicted Indirectly Recognizable HLA Epitopes (PIRCHE).[14] Petersdorf *et al* (2015) proposed an expression model that considers the difference in DPB1 allele specific cell surface expression level.[15] Petersdorf and colleagues reported that mismatching of DPB1 at the expression level influences the incidence and severity of GVHD. Until now, there was only one single SNP found within the 3′ UTR of DPB1 that may serve as a marker for DPB1 expression even though the SNP itself is likely not involved directly in the regulation of expression.

Given the extensive evidence of polymorphism in the intergenic region shared between DPA1 and DPB1 loci, we envisage that the presence of SNPs may have an important role in the regulation of cell surface expression. Information on the expression level of HLA proteins is equitably important in the outcome of transplantation since the expression of these proteins influence the strength of immune response.[16] It is necessary to characterize the SNPs in the promoter region of different DPA1 and DPB1 alleles in comparison to SNP pattern observed in the literature.[10, 11, 17]

In this study, we describe a dual sequencing strategy using shotgun sequencing on Ion Torrent platform and single molecule sequencing on a MinION platform to characterize polymorphism in the DP region spanning over 21.5 kbp (including DPA1, DPB1 and the intergenic region between two loci). We also present the allele and genotype results for both DPA1 and DPB1 loci as part of haplotype analysis of a 100 selected samples.

2. Materials and Methods

2.1 Sample selection

A panel consisting of 100 samples including 12 cell lines from the 10th International Histocompatibility Workshop (IHWS) and 88 control samples were included in this study. The homozygous B-lymphoblastoid cell lines were chosen as it is known to be consanguineous and have previously undergone intensive sequence analysis for the entire MHC region. The control samples were selected to include as many available DPB1 alleles in our tested populations as possible.

All genomic DNA was extracted from peripheral white blood cells or B-cell transformed cell lines using the DNA Midi Kit (Qiagen, Germany) on the QIAsymphony SP instrument according to the vendor's protocol. The concentration and purity of extracted DNA were assessed by the optical density (OD) 260/280 ratio of 1.8-2.0. Samples were then normalized to a concentration of 25 ng/ml.

Figure 1: Schematic diagram represents the primer locations for the long-range amplifications of DP region, covered by three separate PCR fragments.
The green and red boxes depict untranslated and coding regions of the gene, respectively, as defined in the IPD-IMGT/HLA database. DPA1 and DPB1 loci were enveloped in a 10 kb and 9.8 kb, respectively. The 4.7 kb amplicon included the 5′ UTR, transcription regulatory region and intergenic region of both DP genes.

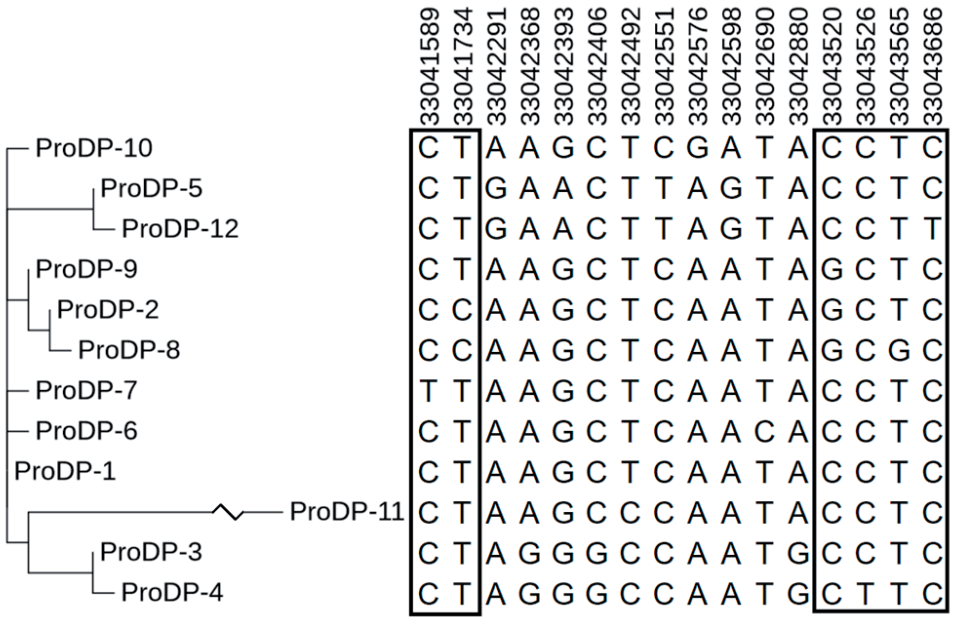


Figure 2: Phylogenetic tree of the DP Promoter sequences.
This tree shows the phylogenetic relationships of the 12 promoter sequences. The promoter sequences are distinguished by 16 SNPs, which are shown with their positions in the GRCh37 reference assembly. ProDP-11 is significantly different from the other eleven promoter sequences, it differs by an additional 60 SNPs, shown in Supplementary Table 1. The first two SNPs, shown within a box, are within the regulatory region of DPA1, while the last four SNPs, also boxed, are within the regulatory region of DPB1. The ten SNPs between the DPA1 and DPB1 regulatory regions define the DP connective region (CR).

2.2 Polymerase chain reaction (PCR)

The entire DP region spanning for 21.5 kbp was enveloped in three separate long-ranged amplification as shown in the schematic diagram in Figure 1. Methods for full gene amplification for shotgun sequencing including primer location, constituents of the master mix and cycling conditions were adapted from previous studies.[17-19] Additionally, the primers for multiplexed MinION sequencing was extended in the 5′ end with the universal sequences from Oxford Nanopore Technologies (ONT) and indices from the PCR Barcoding Expansion 96 kit (EXP-PBC096). One microliter of each product was screened on a 0.7% agarose gel. Once the positive bands of the correct size were confirmed, the amplified products were purified with 0.6X Agencourt AMPure beads (Beckman Coulter, USA) on an automated liquid handler Microlab STAR line (Hamilton, USA).

2.3 Library preparation for Ion Torrent and Nanopore sequencing

Library preparation for Ion Torrent sequencing was adapted from the manufacturer's protocol for Ion Xpress Plus Fragment Library Kit and Ion Xpress Barcode Adapters Kit (ThermoFisher Scientific, USA) to be fully automated on the liquid handler Microlab STAR Line (Hamilton, USA). Briefly, 35 ml of the amplicon pool was enzymatically sheared, ligated with the dual adapters containing unique indexed sequence, and size-selected using a dual bead-based protocol. The multiplexed library pool was diluted to 110 pmol/ml for template preparation on the Ion Chef System (ThermoFisher Scientific, USA). The sequencing was performed on the Ion GeneStudio S5XL system using Ion S5 EXT Sequencing kit on Ion 530 chip (ThermoFisher Scientific, USA).

Library preparation for Nanopore sequencing was performed following the recommended workflow for ligation sequencing SQK-LSK109 chemistry consisting of firstly an end-repair/dA-tailing step to repair blunt ends and add an "Adenine" base to the 3′ end of the amplicon, following by adapter and motor protein ligation onto the prepared ends using NEB ligase enzyme (New England Biolabs, USA), and finally a bead-based purification step to enrich the adapter-ligated fragments and remove unused nucleotides and enzymes. The final prepared library pool was loaded into the MinION R9.4 flow cell as per recommendation from ONT. The sequencing data was collected for 24 hours using the default running script for SQK-LSK109 chemistry by the MinKNOW data acquisition software.

2.4 Data analysis

The signal processing, base calling, trimming and demultiplexing of shotgun sequencing raw data were performed by the Torrent Suite 5.10.1 (ThermoFisher Scientific, USA). The quality-filtered sequences were sorted according to the indices of Ion Xpress barcodes and FASTQ files were generated ready for polymorphism analysis. The Fast5 files from Nanopore sequencing were converted to FastQ files by Guppy base-caller and demultiplexed by Porechop (https://

github.com/rrwick/Porechop). The FASTQ read data was further filtered by a minimum length of 2 kbp and minimum Q-score of 7 using NanoFilt (https://github.com/wdecoster/nanofilt).

The generated nucleotide sequences in FASTQ file format from Ion Torrent and Nanopore sequencing platforms were mapped to hg19 (the human assembly GRCh37) reference to create BAM alignments. The Integrative Genomics Viewer (IGV) browser were used to review sequence alignments generated and to detect the presence of SNPs. Full gene analysis of DPA1 and DPB1 was performed by GenDX NGSengine software 2.15 with IPD-IMGT/HLA library version 3.37 (GenDx, The Netherlands). NGSengine software created a consensus alignment to the IPD-IMGT/HLA reference sequence of the genotyped allele. For alleles with incomplete full length sequence, such as the introns of DPB1, the software were able to identify the HLA allele best matched in the available exon sequences when all distant polymorphisms were linked by the use of single molecule sequencing. The microsatellite region in intron 2 was excluded from the analysis (from gDNA position 8806 to gDNA position 8807) due to the limitation of both technologies to characterize this segment. Where novel sequences were identified by shotgun sequencing and confirmed by MinION sequencing, assignment of a HLA type was based on aligning the novel sequence at the gDNA, cDNA, and protein level to identify the most similar known HLA allele.

2.5 Multiple sequence alignment and phylogenetic tree analysis

Consensus sequences of the 2.5kb promoter sequences were aligned using ClustalW.[20] From this multiple sequence alignment, a phylogenetic tree was calculated using the neighbour joining algorithm.[21] The tree was visualized using The Interactive Tree of Life as shown in Figure 2.[22]

2.6 Haplotype analysis

DPA1~Promoter~DPB1 haplotypes were identified using one of two methods. In many cases, the promoter sequences could be physically phased by using the full length MinION reads. Heterozygous polymorphisms in the overlapping region of PCR amplicons allows polymorphism to be assigned to both the promoter, as well as one of the two DP genes, resulting in direct physical phasing of the promoter haplotypes. In cases where there was no heterozygosity in the overlapping regions and phasing polymorphism physically was not possible, the haplotypes were phased using comparisons to homozygous cell lines and inferred from the statistical haplotype analysis.

Haplotype analysis was performed using PyPop,[23] which implements the expectation algorithm.[24] Genotypes for each sample at DPA1, DPB1, and the promoter were correlated with the TCE groups, as imputed using the functional distance calculations by Crivello *et al*,[25] and the rs9277534 proxy for expression, imputed by Schone *et al*[26] based on DPB1 exon 3 sequences. Pypop reported Hardy Weinberg Equilibrium constants and sample haplotype frequencies, which were used to identify the DPA1~Promoter~DPB1 haplotypes.

Pos	SNP ID	ProDP 1	ProDP 2	ProDP 3	ProDP 4	ProDP 5	ProDP 6	ProDP 7	ProDP 8	ProDP 9	ProDP 10	ProDP 11	ProDP 12
33,041,589	*rs2051548*	C	C	C	C	C	**T**	C	C	C	C	C	C
33,041,734	*rs2856830*	T	**C**	T	T	T	T	**C**	T	T	T	T	T
33,042,291	*rs9380340*	A	A	A	A	**G**	A	A	A	A	A	A	**G**
33,042,368	*rs9469344*	A	A	**G**	**G**	A	A	A	A	A	A	A	A
33,042,393	*rs9394130*	G	G	G	G	**A**	G	G	G	G	G	G	**A**
33,042,406	*rs9469345*	C	C	**G**	**G**	C	C	C	C	C	C	C	C
33,042,492	*rs60349783*	T	T	**C**	**C**	T	T	T	T	T	**C**	T	T
33,042,551	*rs9296073*	C	C	C	C	**T**	C	C	C	C	C	C	**T**
33,042,576	*rs138949603*	A	A	A	A	A	A	A	A	**G**	A	A	A
33,042,598	*rs9296074*	A	A	A	**G**	A	A	A	A	A	A	A	**G**
33,042,690	*rs74341050*	T	T	T	T	**C**	T	T	T	T	T	T	T
33,042,880	*rs9878870*	A	A	**G**	**G**	A	A	A	A	A	A	A	A
33,043,520	*rs2071349*	C	**G**	C	C	C	C	**G**	**G**	C	C	C	C
33,043,526	*rs2071350*	C	C	C	**T**	C	C	C	C	C	C	C	C
33,043,565		T	T	T	T	T	T	**G**	T	T	T	T	T
33,043,686	*rs140559351*	C	C	C	C	C	C	C	C	C	C	C	**T**

Table 1. Polymorphism located in the regulatory region between DPA1 and DPB1. ProDP-11 is significantly different from the other eleven promoter sequences, it differs by an additional 60 SNPs, shown in Supplementary Table 1.

Promoter	N	DPA1	DPB1	HV1	HV2	HV3	HV4	HV5	HV6	CR	TCE	Expr SNP
ProDP_1	1	01:03	02:01	LFQG	EEFV	DEE	ILEEE	M	GGPM	1	3	A
ProDP_1	30	01:03	04:01	LFQG	EEFA	AAE	ILEEK	M	GGPM	1	3	A
ProDP_1	1	01:04	04:01	LFQG	EEFA	AAE	ILEEK	M	GGPM	1	3	A
ProDP_1	10	01:03	04:02	LFQG	EEFV	DEE	ILEEK	M	GGPM	1	3	A
ProDP_1	1	01:03	23:01	LFQG	EEFV	AAE	ILEEK	M	GGPM	1	3	A
ProDP_1	1	01:03	33:01	LFQG	EEFA	AAE	ILEEE	M	GGPM	1	3	A
ProDP_1	1	01:03	80:01	LFQG	EEFV	DED	ILEEK	M	GGPM	1	3	A
ProDP_1	5	03:01	105:01	LFQG	EEFV	DEE	ILEEK	M	GGPM	1	3	A
ProDP_1	1	01:03	126:01	LFQG	EEFA	AAE	ILEEK	M	GGPM	1	3	A
ProDP_1	1	01:03	128:01	LFQG	EEFA	AAE	ILEEK	M	GGPM	1	3	A
ProDP_1	3	01:03	138:01	LFQG	EEFV	AAE	ILEEK	M	GGPM	1	3	A
ProDP_7	4	01:03	04:02	LFQG	EEFV	DEE	ILEEK	M	GGPM	1	3	A
ProDP_10	1	01:03	04:01	LFQG	EEFA	AAE	ILEEK	M	GGPM	1A	3	A
ProDP_1	13	01:03	03:01	VYQL	EEFV	DED	LLEEK	V	DEAV	1	2	G
ProDP_1	4	01:03	06:01	VYQL	EEFV	DED	LLEEE	M	DEAV	1	2	G
ProDP_1	1	01:03	11:01	VYQL	QEYA	AAE	LLEER	M	DEAV	1	3	G
ProDP_1	1	01:03	20:01	VYQL	EEFV	DED	LLEEK	M	DEAV	1	3	G
ProDP_1	3	01:03	21:01	VYQL	EELV	EAE	ILEEE	M	DEAV	1	2	G
ProDP_1	1	01:03	36:01	VYQL	EELV	EAE	ILEEK	M	DEAV	1	3	G
ProDP_1	7	01:03	104:01	VYQL	EEFV	DED	LLEEK	V	DEAV	1	2	G
ProDP_1	1	01:03	124:01	VYQL	EEFV	DED	LLEEK	V	DEAV	1	2	A
ProDP_1	1	01:03	130:01	VYQL	EEFV	DED	LLEEK	M	DEAV	1	3	G
ProDP_1	1	01:03	05:01*	LFQG	EELV	EAE	ILEEK	M	DEAV	1	3	G
ProDP_6	4	02:07	19:01	LFQG	EEFV	EAE	ILEEE	I	DEAV	1B	2	G
ProDP_6	1	02:12	85:01	VYQL	EEYA	AAE	ILEEK	M	DEAV	1B	3	G
ProDP_1	2	01:03	15:01	VYQG	QEYA	AAE	LLEER	M	VGPM	1	3	G
ProDP_1	1	01:04	15:01	VYQG	QEYA	AAE	LLEER	M	VGPM	1	3	G
ProDP_1	3	01:03	18:01	VYQG	EEFV	DEE	ILEEK	M	VGPM	1	3	G
ProDP_1	1	03:01	18:01	VYQG	EEFV	DEE	ILEEK	M	VGPM	1	3	G
ProDP_2	16	01:03	02:01	LFQG	EEFV	DEE	ILEEE	M	GGPM	1	3	A
ProDP_2	1	01:03	02:02	LFQG	EELV	EAE	ILEEE	M	GGPM	1	3	A
ProDP_2	1	01:03	04:01*	LFQG	EEFA	AAE	ILEEK	M	GGPM	1	3	A
ProDP_2	1	01:03	416:01	LFQG	EEFV	DEE	ILEEE	M	GGPM	1	3	A
ProDP_2	1	01:03	47:01	LFQG	EEFV	EAE	ILEEE	M	GGPM	1	3	A
ProDP_2	1	01:03	81:01	LFQG	EEFA	DEE	ILEEE	M	GGPM	1	3	A
ProDP_8	1	01:03	02:01	LFQG	EEFV	DEE	ILEEE	M	GGPM	1	3	A
ProDP_9	3	01:03	02:01	LFQG	EEFV	DEE	ILEEE	M	GGPM	1	3	A
ProDP_2	2	01:03	16:01	LFQG	EEFV	DEE	ILEEE	M	DEAV	1	3	G

Promoter	N	DPA1	DPB1	HV1	HV2	HV3	HV4	HV5	HV6	CR	TCE	Expr SNP
ProDP_2	1	01:03	652:01	LFQG	EEFV	DEE	ILEEE	M	DEAV	1	3	G
ProDP_3	6	02:01	01:01	VYQG	EEYA	AAE	ILEEK	V	DEAV	2	3	G
ProDP_3	1	02:02	01:01	VYQG	EEYA	AAE	ILEEK	V	DEAV	2	3	G
ProDP_3	5	02:01	11:01	VYQL	QEYA	AAE	LLEER	M	DEAV	2	3	G
ProDP_3	2	02:01	13:01	VYQL	EEYA	AAE	ILEEE	I	DEAV	2	3	G
ProDP_3	3	02:01	17:01	VHQL	EEFV	DED	ILEEE	M	DEAV	2	2	A
ProDP_3	1	02:01	86:01	VHQL	EEFV	DED	ILEEE	M	GGPM	2	2	A
ProDP_4	4	02:01	09:01	VHQL	EEFV	DED	ILEEE	V	DEAV	2	1	G
ProDP_4	8	02:01	10:01	VHQL	EEFV	DEE	ILEEE	V	DEAV	2	1	G
ProDP_4	1	02*	10:01	VHQL	EEFV	DEE	ILEEE	V	DEAV	2	1	G
ProDP_4	7	02:01	14:01	VHQL	EEFV	DED	LLEEK	V	DEAV	2	2	G
ProDP_4	2	02:01	35:01	VHQL	EEFV	DED	ILEEK	V	DEAV	2	2	G
ProDP_4	1	02:01	45:01	VHQL	EEFV	DEE	LLEEK	V	DEAV	2	2	G
ProDP_5	5	02:02	01:01	VYQG	EEYA	AAE	ILEEK	V	DEAV	3	3	G
ProDP_5	1	02:02	14:01*	VHQL	EEFV	DED	LLEEK	V	DEAV	3	2	G
ProDP_5	10	02:02	05:01	LFQG	EELV	EAE	ILEEK	M	DEAV	3	3	G
ProDP_5	1	02:01	05:01	LFQG	EELV	EAE	ILEEK	M	DEAV	3	3	G
ProDP_5	1	02:02	22:01	LFQG	EELV	EAE	ILEEE	M	DEAV	3	3	G
ProDP_12	2	02:02	135:01	LFQG	EELV	EAE	ILEEK	M	DEAV	3	3	G
ProDP_5	1	02:02	02:02*	LFQG	EELV	EAE	ILEEE	M	GGPM	3	3	A
ProDP_11	5	04:01	107:01	VYQL	EEYA	AAE	ILEEE	I	DEAV	4	3	G

Table 2. DP Promoter sequences compared with the correlated DPB1 hypervariable regions.
Distinct correlations can be seen between the promoter sequences and hypervariable regions 1,5 and 6. The imputed TCE group and expression SNP are also shown.

	DPA1	Prom	DPB1	rs9277534	TCE
DPA1		0.860	0.687	0.945	0.841
Prom	0.860		0.780	0.645	0.608
DPB1	0.687	0.780		1	1
rs9277534	0.945	0.645	1		0.459
TCE	0.841	0.608	1	0.459	

Table 3. Linkage disequilibrium constants
The Wn constant, a symmetric measure of linkage disequilibrium between multi-allelic loci, is used to compare the strength of the linkage disequilibrium between loci and expression markers. Wn constants fall between 0 and 1, where 1 indicates the maximum observed linkage between loci, and the table cells are shaded to visually show the strength of the linkage, where red represents a strong linkage, and blue/purple are weaker linkage. DPA1, Promoter sequences, DPB1, the rs9277534 expression SNP, and imputed TCE group are compared.

Linkage Disequilibrium was estimated using the Wn statistic.[27] This statistic is a number between 0 and 1 which represents a symmetric measure of the correlation between two multi-allelic polymorphic loci. To estimate the linkage between the promoter with DPA1 and DPB1, Wn was calculated on the 2-field typings of DPB1, DPA1 and the promoter sequence, as well as the imputed expression SNP and TCE groups.

3. Results

3.1 Promoter SNP haplotypes

In the panel of 100 samples, 12 distinct promoter SNP haplotypes, named ProDP-1 to -12, were inferred within the region between the start codons of DPA1 and DPB1 loci. The promoter SNP haplotype can be distinguished by 16 characteristic SNPs (Table 1). Frequency of transversions and transitions by nucleotide types was 43.75% for C/T (7), 37.5% for A/G (6), 12.5% for C/G (2) and 6.25% for T/G (1). ProDP-11 was distinctively diverging from the rest of haplotypes, as it included additional 60 polymorphic positions following a 12-base deletion. The polymorphic sites in ProDP-11 can be found in Supplementary Table 1. The most commonly observed promoter SNP haplotype was ProDP-1 (95 of 200), followed by ProDP-2 (24 of 200), and ProDP-4 (23 of 200). ProDP-3 and ProDP-5 were both identified in 18 and 19 haplotypes, respectively. ProDP-8 and ProDP-10 were only observed once in the panel of 100 samples.

Through multiple sequence alignments, the patterns of SNP association suggest that there were three distinct sections within the intergenic region between two DP loci. The first segment consisted of two polymorphisms, rs2051548 and rs2856830, that are located at (-243) and (-388), respectively upstream from exon 1 and within the transcription regulatory region of DPA1. The connective region (CR) between the DP genes included 10 polymorphic sites as described in table 1, from rs9380340 to rs987870. The last four substitutions, namely rs2071349, rs2071350, 33,043,565, rs140559351, located within 300

bp upstream of DPB1 start codon, therefore within the regulatory region of this gene, and made up the last segment of the intergenic region.

Within the CR between two the DP loci, four distinct lineages were observed based on the sequence homology and haplotype-linked SNPs (Figure 2). The first pattern was defined as CR1 consisting of multiple promoter SNP haplotype ProDP-1, ProDP-2, ProDP-7, ProDP-8, ProDP-9 and ProDP-10. ProDP-6 and ProDP-10 haplotypes were classified as CR1A and CR1B, respectively, as they were from one clade as the aforementioned haplotypes in CR1. The second pattern, CR2, consisting of ProDP-3 and ProDP-4 was delineated based on the presence of 4 characteristic SNPs including rs9469344, rs9469345, rs60349783, and rs987870. Likewise, the third pattern (CR3) was determined on the presence of another set of 4 SNPs, rs9380340, rs9394130, rs9296073, and rs9296074, that were found in promoter haplotype ProDP-5 and ProDP-12. Lastly, ProDP-11 sequence formed its own CR clade due to the diverging polymorphism content compared to the remaining 11 CR sequences.

3.2 Analysis of promoter SNP haplotypes with the hypervariable region in DPB1 and the connective region (CR) in the intergenic section

DPA1~Promoter~DPB1 haplotypes were examined in respect of DPB1 amino acid motifs, specifically 6 hypervariable regions HV1-HV6[6, 28] (also named A-F)[13, 29] and CR in the intergenic region between DP genes. There was a significant association between all 12 promoter SNP haplotypes, CR lineages and amino acid positions 84-87 in the HV6 at near complete linkage as shown in Table 2, with only a few exceptions. The exceptions are described in the result section 3.5.

The patterns of association between promoter haplotype contents and amino motifs in DPB1 further highlighted there were potential multiple lineages within ProDP-1, which was the most commonly observed haplotype and linked to numerous DPB1 alleles. The first lineage of ProDP-1 sequence formed a haplotype with a distinct group of DPB1 alleles (DPB1*02:01, *04:01, *04:02, *23:01, *33:01, *80:01, *105:01, *126:01, *128:01, *138:01) which were all correlated with LFQG in HV1, M in HV5 and GGPM in HV6. This lineage was found in 55 of 200 haplotypes and in striking association with DPA1*01:03, 01:04 and 03:01 alleles. Interestingly, DPA1*01:03, *01:04 and *03:01 alleles share multiple similarities in the amino acid composition. For instance, DPA1*01:04 and DPA1*03:01 only differs from DPA1*01:03 at one and five codons, respectively, in contrast to DPA1*02 and DPA1*04 alleles with more than 15 different amino acids in comparison to DPA1*01:03 allele. The second pattern in ProDP-1 (33 of 200) was associated with only DPA1*01:03 and the following DPB1*03:01, *05:01, *06:01, *11:01, *20:01, *21:01, *36:01, *104:01, *124:01, *130:01 alleles. This lineage was in near complete agreement with amino acid motif VYQL in HV1 (with one exception of LFQG found in one sample positive for DPB1*05:01, described in 3.5) and 100% with DEAV in HV6 of DPB1 alleles. The third lineage in ProDP-1

(7 of 200) was observed with DPB1*15:01, *18:01 and amino acid motif VYQG in HP1, M in HV5 and VGPM in HV6. This lineage was also observed in association with DPA1*01:03, *01:04 and *03:01 alleles.

ProDP-7 and ProDP-10 showed identical pattern of association as the first lineage in ProDP-1 as both promoter SNP-linked haplotypes were linked to DPA1*01:03 ~ DPB1*04 ~ LFQG in HV1 ~ M in HV5 ~ GGPM in HV6. Likewise, ProDP-6 shared a common DEAV in HV6 as the second lineage in ProDP-1. However, the ProDP-1 and ProDP-6 haplotypes diverged from HV1 through to HV5. The linkage between ProDP-10 and ProDP-6 to the associated amino acid motifs GGPM and DEAV in HV6, respectively, was consistent with haplotype content in the connective region and phylogenetic tree clades, therefore defined as CR1A and CR1B.

ProDP-2 diverged into two subgroups based on the association with GGPM or DEAV amino acid motifs in HV6. Interestingly, both lineages were in 100% correlation with DPA1*01:03 ~ LFQG in HV1 ~ ILEEE in HV4 (with one exception of ILEEK found in one sample positive for DPB1*04:01) ~ M in HV5 haplotype. The DPB1 alleles in association with first ProDP-2 lineage (21 of 200 haplotypes) were DPB1*02:01, *02:02, *04:01, *47:01, *81:01, *416:01. The DPA1*01:03 ~ HV1/4/5/6 of DPB1 haplotype in the first lineage of ProDP-2 was also observed with ProDP-8 (1 of 200) and ProDP-9 (3 of 200). On the other hand, DPB1*16:01 and DPB1*652:01 alleles were associated with the second lineage of ProDP-2 and DEAV motif in HV6. DPB1*16:01 and DPB1*652:01 alleles share a high degree of exon sequence homology with one single difference in codon 205 of exon 4, which encodes for the transmembrane region of the DPB1 chain, whereas there are multiple differences in their intronic sequences.

ProDP-3 and ProDP-4 SNP haplotypes have many common features within the connective region (CR) between two DP genes and therefore clustered into two CR2 pattern. Two promoter haplotypes only differed at one SNP rs2071350, which defined ProDP-4. Interestingly, ProDP-4 (23 of 200) was observed in 100% agreement with VHQL in HV1, EEFV in HV2, V in HV5 and DEAV in HV6 of DPB1*09:01, *10:01, *14:01, *35:01 and *45:01. The pattern of association between promoter haplotype ProDP-3 and the hypervariable region in DPB1 allele was not as distinctive as ProDP-4. ProDP-3 SNP haplotype (18 of 200) was linked to DPB1*01:01, *11:01, *13:01, *17:01, *86:01 allele and DEAV motif in HV6 (with one exception of GGPM in a sample positive for DPB1*86:01).

Promoter haplotypes ProDP-5 and -12 were grouped in CR3 and distinguished by a single SNP (rs140559351). ProDP-5 was associated with DPB1*01:01, *02:02, *05:01, *14:01, *22:01 and ProDP-12 was only observed with DPB*135:01. Both ProDP-5 and ProDP-12 were associated with DEAV in HV6 (one exception of GGPM found once in sample positive

for DPB1*02:02). ProDP-5 sequences (19 of 200) were observed predominantly with DPB1*05:01 and ProDP-12 sequences (2 of 200) were with DPB1*135:01. DPB1*05:01 and DPB1*135:01 alleles share high degree of exon sequence homology with one single difference in codon 205 of exon 4, similar pattern as DPB1*16:01 and DPB1*652:01 alleles.

ProDP-11 was distinctive different from other promoter haplotypes. ProDP-11 was exclusively in association with DPA1*04:01 and DPB1*107:01, found in 5 of 200 haplotypes. Interestingly, DPB1*107:01 and DPB1*13:01 alleles were identical in exon 2 and downwards. The two alleles only differed in two codons in exon 1, yet associated to different DPA1 allele and promoter haplotypes. DPB1*13:01 was linked to ProDP-3 and DPA1*02:01 alleles; whereas DPB1*107:01 was associated with ProDP-11 and DPA1*04:01.

3.3 Phylogenetic tree

The phylogenetic relationship among consensus sequence of the 12 promoter haplotype sequence is depicted in Figure 2. The tree showed that promoter haplotype sequences with the same CR pattern clustered together and supported the view on sequence homology in the intergenic region. The majority of the sequences derived from Pro-DP1, ProDP-6, ProDP-7, ProDP-10 were on one branch. This grouping pattern indicated that ProDP-1, ProDP-6, ProDP-7 and ProDP-10 could be diverged from a common ancestral gene. ProDP-2, ProDP-8 and ProDP-9 sequences clustered in a different clade as the ProDP-1/6/7/10 branch, due to the additional SNPs in the 3′ end of the CR. ProDP-2, ProDP-8 and ProDP-9 all shared one SNP rs2071349 in the regulatory region of DPB1 gene. ProDP-3/4 and ProDP-5/12 haplotypes clustered into 2 separate branches consistent with previously classification as CR2 and CR3 in the haplotype analysis. ProDP-11 linked-SNPs formed a distinct branch and distanced from all other branches.

3.4 Promoter SNP haplotype analysis with TCE group and expression marker

The association between promoter SNP haplotype, TCE group and expression marker is also shown in Table 2. There was no apparent pattern between promoter SNP-linked haplotype, TCE group and expression marker as one promoter haplotype was observed with DPB1 alleles from different TCE groups and both variants of rs9277534 SNP. However, when the promoter haplotypes were clustered in respect of DPB1 amino acid motifs in the HV, individual lineage within the promoter haplotype exhibited an agreement with the expression model, and less so with the TCE model. For example, the first lineage in ProDP-1 was observed in strong association with poor immunogenetic and lowly expressed DPB1 alleles as they were all classified in TCE group 3 and linked to rs9277534A marker. On the other hand, the second lineage in ProDP-1 was observed with DPB1 alleles from both TCE group 2 and 3, while almost all samples in this lineage were linked to rs9277534G marker with only one exception of DPB1*124:01 allele (rs9277534A). DPB1*124:01 and

DPB1*03:01 alleles are identical in exon 2, yet different from intron 2 and downwards as evidence of evolutionary recombination between exon 2 and exon 3.

Out of all promoter haplotypes, the association between ProDP-4, TCE and rs9277534A marker was the most noticeable. All DPB1 alleles correlated with ProDP-4 were classified as highly expressed allele due to their linkage with rs9277534G polymorphism in the expression model. In regards to the TCE model, ProDP-4 haplotype was found in both TCE group 1 and TCE group 2 alleles with high immunogenetic functionality. This group of DPB1 alleles also had complete agreement in the amino acid motifs of HV1/2/5/6 and LD with DPA1*02:01 as mentioned previously.

The Linkage Disequilibrium constants are shown in Table 3. The Wn statistic, which represents the strength of linkage disequilibrium between two multi-allelic loci, is shown between DPA1, promoter SNP-linked haplotype, DPB1, the rs9277534 SNP, and TCE group. The colour of the table cells depicts the strength of the linkage, where red represents a strong linkage (maximum 1), and blue represents a weak linkage (minimum 0). DPB1 demonstrates a strong linkage to the rs9277534 expression SNP and TCE group due to the use of DPB1 as a predictor of these loci, but these markers are not strongly linked to DPA1. The promoter sequences were found to be strongly linked to DPA1, and less so to DPB1. Neither TCE group or expression SNP were correlated with the promoter sequences.

3.5 Exceptions in DP haplotype analysis and hypervariable amino acid motifs

There are a few exceptions in the DP haplotype analysis in respect of the hypervariable amino acid motifs in DPB1 alleles. ProDP-1 was only observed with DPB1*05:01:01:New ~ DPA1*01:03 in one sample. The DPB1*05:01:01:New sequence in this sample had 17 differences in the intron 1 compared to the intronic sequence of other DPB1*05:01:01 sequences in the selected DNA panel, which were all in LD with ProDP-5 and DPA1*02 allele. Another exception was observed in a different sample, which was positive for DPA1*01:03 ~ ProDP-2 ~ DPB1*04:01. DPA1*01:03 and DPB1*04:01 haplotype was observed with ProDP-1 30 times, but only once with ProDP-2. The DPB1*04:01:01:New allele in this sample was also different from other DPB1*04:01:01 alleles in the 5′ UTR at gDNA position -299.

Similar observations were found in ProDP-3 and ProDP-5 haplotypes. DPB1*86:01 allele shared similarity in HV1 to HV5 with other DPB1 alleles associated with ProDP-3 haplotypes. However, DPB1*86:01 segregated with others at GGPM motif in HV6 as the exception. DPB1*02:02:01 was seen in association with ProDP-2 as well as ProDP-5. DPB1*02:02:01:01 allele was linked to DPA1*01:03:01:01 and ProDP-2, whereas DPB1*02:02:01:02 was with DPA1*02:02:02:01 and ProDP-5. The two DPB1*02:02:01 alleles were different at 22 sites in the intron 1 sequence and therefore diverged into two different DPA1 ~ promoter ~

DPB1 haplotypes. Likewise, DPB1*14:01:01 allele was linked to DPA1*02:01:01 ~ ProDP-4 in 7 samples, however, only once was observed with DPA1*02:02:02 ~ ProDP-5. These were the intronic subtypes of DPB1*14:01 alleles, where intron 1 region of these two DPB1*14:01:01:New subtypes were distinctive different at 10 polymorphic sites.

4. Discussion

Polymorphisms in the entire 21.5 kb region containing the DP genes were evaluated in this study. Specifically, the haplotype content within the intergenic region between DPA1 and DPB1 genes were closely delineated. We have identified 76 SNPs and 12-base deletion in approximately 2500 bp region between the two start codons in 100 samples and in which 12 promoter SNP haplotypes were inferred, named ProDP-1 to ProDP-12. The organization of polymorphism in the DP promoter region was also elucidated using multiple sequence alignments and classified into three segments. The first segment, located 388 bp upstream of DPA1 exon 1 including the 5′ UTR and transcription regulatory regions of this gene. The connection region (CR) between two genes spanned for 1800 bp and consisted of 10 characteristic SNPs that diverged into three distinct lineages of the CR (without taking into account polymorphisms in ProDP-11 haplotype). The last segments located 300 bp upstream of DPB1 start codon and contained specific regulatory sequences, the X, Y and W′ box of the DPB1 gene.

SNP-linked haplotype analysis revealed that DPA1 has stronger linkage disequilibrium (LD) with promoter SNP haplotypes than DPB1. The association between promoter SNP haplotypes and DPA1 locus also highlights the similarities between DPA1*01:03, 01:04 and 03:01 alleles as they shared the CR and promoter lineages, which were distinctive different from DPA1*02 and DPA1*04 allele groups. In addition, sequence analysis of the whole region gave insights on the association of hypervariable regions (HV) in exon 2 of DPB1, promoter haplotype and correlated DPB1 allele groups. DPB1 antigen group was not as well defined as other HLA loci due to the scarcity of distinguishing anti-HLA antibodies and lower cell surface expression making serologically typing more challenging. Based on the extended haplotypes, we could potentially cluster DPB1 alleles using their association with DPA1 and promoter sequences. For example, our data suggested that DPB1*16:01 and DPB1*652:01 alleles were clustered as one based on the sequence homology and functionality of the protein, which was however not reflected in the current nomenclature.

Through haplotype content analysis, the relationship between DPA1~promoter~DPB1 and amino acid motifs was strictly related to HV1 (position 8 to 11) and HV6 (position 84 to 87) with minor exceptions. Of 12 SNP-linked haplotypes, DPB1 alleles in LD with ProDP-4 stood out due to their complete agreement in VHQL motif in HV1, EEFV motif in HV2, V

motif in HV5, DEAV motif in HV6 as well as high expression variant rs9277534G, and highly immunogenic TCE group. The differences in HV3 and HV4 of each DPB1 allele associated with ProDP-4 represented the immunogenicity of different β-chains in this group. This observation was consistent with early observations by Cesbron *et al*[28] in which the authors indicated mismatches in the third and fourth HV of different DPB1 alleles can influence the risk of graft-versus-host disease in bone marrow transplantation.

In this study, we have unravelled multiple unique LD of different intronic subtypes of the same DPB1 alleles. For example, DPB1*02:02:01:01 allele was seen in association with DPA1*01:03:01:01 and ProDP-2, where DPB1*02:02:01:02 was with DPA1*02:02:02:01 and ProDP-5. The two DPB1*02:02:01 alleles were different at 22 sites in the intron 1 sequence and therefore diverged into two different DPA1 ~ promoter ~ DPB1 haplotypes. Likewise, there were also subtypes of DPB1*14:01:01 alleles. DPB1*14:01:01:01/02 was commonly associated with ProDP-4 ~ DPA1*02:01:01 as observed in 7 samples, whereas the other variant of DPB1*14:01:01:New was associated with ProDP-5 ~ DPA1*02:02:02. The intron 1 region of these two DPB1*14:01:01 alleles are distinctive different at 10 polymorphic sites. Those DPB1 alleles shared identical sequence at exon 2 and 3, therefore based on conventional matching at Antigen Recognition Site (ARS) these alleles would be considered as DPB matched, yet they were on different extended haplotypes. Furthermore, they would have the same TCE group based on exon 2, and the same expression SNP based on exon 3, so we would never see the difference based on conventional matching. This finding highlights the importance of full-gene or haplotype analysis, since SNP haplotype diversity has important clinical implication as previously reported by Petersdorf *et al* (2013).[30]

The LD between DPA1 ~ Promoter ~ DPB1 was strong in our cohort. However, we need to confirm these findings in other populations, which requires the collaborative effort by the International HLA & Immunogenetics Workshop. It is important to expand the study panel and investigate the same assigned allele in different populations as it was evident in our study that intronic variations of one DPB1 allele could segregate to different DPA1, promoter and extended haplotypes. In addition, we also observed certain haplotype was restricted to the ethnicity. For example, ProDP-11 SNP haplotype was exclusively in association with DPA1*04:01 ~ DPB1*107:01. This promoter haplotype was reported as DP-PRO4 by Varney *et al*[10] and Hap3 by Liu *et al*[17]. Both studies denoted that this haplotype was observed in individuals with Southern Asian origin, which was also consistent with the samples in our study cohort.

Overall, we report for the first time the clustering of polymorphisms in the DP region, especially in the intergenic region between DPA1 and DPB1 genes. The polymorphic clusters of the DPA and DPB alleles, the CR have been profiled. The linkage between DPA1

~ promoter ~ DPB1 and amino acid motifs in HV1/HV6 is highly conserved and more significant than previously described.[6] This new view on the full region haplotype shows the relation of genes and alleles; and provides a basis for future functionality related nomenclature. Lastly, the novel clustering of the DP extended haplotype warrants for future investigations of DP haplotype matching in the outcome of HSCT and validate the permissive view of DP matching.

Acknowledgements

We appreciate the contribution of Mats Bengtsson for DNA material of DPB1*86:01 allele. We are grateful to all the colleagues at the Department of Clinical Immunology, PathWest and the Department of Transplantation Immunology, Maastricht University Medical Center for their technical assistance and troubleshooting.

References

1. Dausset J. Leuco-agglutinins IV: leuco-agglutinins and blood transfusion. Vox Sang. 1954;4:190-8.
2. Van Rood JJ, Van Leeuwen A. LEUKOCYTE GROUPING. A METHOD AND ITS APPLICATION. The Journal of clinical investigation. 1963;42:1382-90.
3. Payne R, Tripp M, Weigle J, Bodmer W, Bodmer J. A NEW LEUKOCYTE ISOANTIGEN SYSTEM IN MAN. Cold Spring Harbor symposia on quantitative biology. 1964;29:285-95.
4. Leffler EM, Gao Z, Pfeifer S, Segurel L, Auton A, Venn O, *et al*. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science (New York, NY). 2013;339:1578-82.
5. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. Nucleic acids research. 2015;43:D423-D31.
6. Hollenbach JA, Madbouly A, Gragert L, Vierra-Green C, Flesch S, Spellman S, *et al.* A combined DPA1~DPB1 amino acid epitope is the primary unit of selection on the HLA-DP heterodimer. Immunogenetics. 2012;64:559-69.
7. Servenius B, Gustafsson K, Widmark E, Emmoth E, Andersson G, Larhammar D, *et al*. Molecular map of the human HLA-SB (HLA-DP) region and sequence of an SB alpha (DP alpha) pseudogene. The EMBO journal. 1984;3:3209-14.
8. Nezu N, Ryu K, Koide Y, Yoshida TO. Regulation of HLA class II molecule expressions by IFN-gamma. The signal transduction mechanism in glioblastoma cell lines. Journal of immunology (Baltimore, Md : 1950). 1990;145:3126-35.
9. Sugawara M, Ponath PD, Yang Z, Strominger JL. Interferon-gamma response region in the promoter of the class II MHC gene, DPA. Hum Immunol. 1992;35:157-64.
10. Varney MD, Gavrilidis A, Tait BD. Polymorphism in the regulatory regions of the HLA-DPB1 gene. Hum Immunol. 1999;60:955-61.
11. Liu X, Xu Y, Shen Y, Zhang H, Fu Y, Liu Z, *et al*. HLA-DPA1 promoter haplotypes are differently distributed in southern Chinese ethnic groups. Tissue antigens. 2005;65:172-7.
12. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409:928-33.
13. Zino E, Frumento G, Marktel S, Sormani MP, Ficara F, Di Terlizzi S, *et al*. A T-cell epitope encoded by a subset of HLA-DPB1 alleles determines nonpermissive mismatches for hematologic stem cell transplantation. Blood. 2004;103:1417-24.
14. Thus KA, de Hoop TA, de Weger RA, Bierings MB, Boelens JJ, Spierings E. Predicted Indirectly ReCognizable HLA Epitopes Class I Promote Antileukemia Responses after Cord Blood Transplantation: Indications for a Potential Novel Donor Selection Tool. Biology of Blood and Marrow Transplantation. 2016;22:170-3.
15. Petersdorf EW, Malkki M, O'hUigin C, Carrington M, Gooley T, Haagenson MD, *et al*. High HLA-DP Expression and Graft-versus-Host Disease. New England Journal of Medicine. 2015;373:599-609.

16. Balgansuren G, Regen L, Sprague M, Shelton N, Petersdorf E, Hansen JA. Identification of the rs9277534 HLA-DP expression marker by next generation sequencing for the selection of unrelated donors for hematopoietic cell transplantation. Hum Immunol. 2019;80:828-33.
17. Liu X, Liu Z, Lin B, Liu Y, Chen Z, He W, *et al*. Catalog of 162 single nucleotide polymorphisms (SNPs) in a 4.7-kb region of the HLA-DP loci in southern Chinese ethnic groups. Journal of human genetics. 2004;49:73-9.
18. Truong L, Matern B, D'Orsogna L, Martinez P, Tilanus MGJ, De Santis D. A novel multiplexed 11 locus HLA full gene amplification assay using next generation sequencing. HLA. 2019.
19. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, *et al*. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue antigens. 2012;80:305-16.
20. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994;22:4673-80.
21. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987.
22. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research. 2019;47:W256-w9.
23. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update--a software pipeline for large-scale multilocus population genomics. Tissue antigens. 2007;69 Suppl 1:192-7.
24. Do CB, Batzoglou S. What is the expectation maximization algorithm? Nat Biotechnol. 2008;26:897-9.
25. Crivello P, Zito L, Sizzano F, Zino E, Maiers M, Mulder A, *et al.* The impact of amino acid variability on alloreactivity defines a functional distance predictive of permissive HLA-DPB1 mismatches in hematopoietic stem cell transplantation. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2015;21:233-41.
26. Schone B, Bergmann S, Lang K, Wagner I, Schmidt AH, Petersdorf EW, *et al.* Predicting an HLA-DPB1 expression marker based on standard DPB1 genotyping: Linkage analysis of over 32,000 samples. Hum Immunol. 2018;79:20-7.
27. Thomson G, Single RM. Conditional Asymmetric Linkage Disequilibrium (ALD): Extending the Biallelic r2 Measure. Genetics. 2014;198:321-31.
28. Cesbron A, Moreau P, Milpied N, Harousseau JL, Muller JY, Bignon JD. Crucial role of the third and fourth hypervariable regions of HLA-DPB1 allelic sequences in the mixed lymphocyte reaction. Hum Immunol. 1992;33:202-7.
29. Urlacher A, Dormoy A, Tongio MM. DP epitope mapping by using T-cell clones. Hum Immunol. 1992;35:100-8.
30. Petersdorf EW, Malkki M, Horowitz MM, Spellman SR, Haagenson MD, Wang T. Mapping MHC haplotype effects in unrelated donor hematopoietic cell transplantation. Blood. 2013;121:1896-905.

CHAPTER 9

9

Specific amino acid patterns define split specificities of HLA-B15 antigens enabling conversion from DNA based typing to serological equivalents

B. Duygu, B.M. Matern, L. Wieten, C.E.M. Voorter, M.G.J. Tilanus

Transplantation Immunology, Tissue Typing Laboratory, Maastricht University Medical Center, Maastricht, The Netherland

Manuscript Submitted (HLA).

Abstract

Abstract

HLA-B typing by serological approaches has defined the B15 serological group and its subgroups (or splits) B62, B63, B75, B76, B77, as well as the serological group B70 with its splits B71 and B72. The scarcity of sera with specific anti-HLA antibodies makes the serological typing method difficult to discriminate between the wide variety of HLA antigens, especially between the B15 antigen subgroups. Advancements in DNA based technologies have led to a switch from serological typing to high resolution DNA typing methods. DNA sequencing techniques assign B15 specificity to all alleles in the HLA-B*15 allele group, without distinction of the serological split equivalents. However, the presence of antibodies in the patient defined as split B15 antigens urges the identification of HLA-B*15 allele subtypes of the donor, since the presence of donor specific antibodies is an important contra-indication for organ transplantation. Although the HLA dictionary comprises information regarding the serological subtypes of HLA alleles, there are currently 394 B15 protein alleles out of 516 in the IPD-IMGT/HLA database (3.38.0) without any assigned serological subtype. In this regard, we aimed to identify specific amino acid patterns characteristic of each B15 serological split, in order to facilitate the assignment of B*15 serological equivalents to alleles after high resolution molecular typing. As a result, serological specificities of 372/394 not yet assigned alleles could be predicted based on amino acid motifs. Furthermore, two new serological types were defined and added, B62-Bw4 and B71-Bw4.

Introduction

Serological typing has been used for a long time to determine HLA typing of patients and donors. This method is based on complement-dependent cytotoxicity (CDC) test or microlymphocytotoxicity assay, which measures the reactivity of a panel of sera containing well-characterized anti-HLA antibodies [1]. This technique has also been performed in the International Histocompatibility Workshops, and resulted in identification of different serological specificities of HLA genes including HLA-B15. In the HLA-B15 antigen group the serological splits B62, B63, B75, B76, B77 and the B70 group (B71 and B72 split antigens) have been defined [2-6]. However, the scarcity of sera with specific anti-HLA antibodies, in particular antibodies against serological splits or infrequent HLA antigens, makes it highly challenging to discriminate the high variety of serologically defined HLA antigens. Furthermore, the need for living cells to perform the CDC method creates additional challenges. The advancements in DNA based technologies have led to a switch from serological to high resolution DNA typing approaches to obtain a refined HLA typing of patients and donors [7]. A major advantage of the high resolution full length sequencing is the recognition of allele polymorphisms enabling distinction of epitopes. On the other hand, identification of serological specificities remains important for transplantation, especially for determination of donor specific antibodies (DSA) present in the patient's serum. When an antibody against a B15 split antigen in the serum of the recipient is detected, the serological subtype of the B15 antigen present on the donor cells must be identified in order to determine whether this anti-B15 antibody is donor specific. Because DSA can mediate and promote acute and chronic graft rejection, the presence of DSA is an important contraindication for solid organ transplantation [8, 9]. In our current practice, high resolution typing of HLA-B*15 is always performed for kidney patients and living donors. The serological subtype is determined using the HLA data dictionary [10] from the IPD-IMGT/HLA database [11]. However, the number of alleles with assigned serological split types is limited in the database, and serological typing by CDC is not always possible to identify serological type due to aforementioned limitations. Without knowing the B15 serological equivalent, the risk of rejection for patients having anti-B15-split HLA antibodies is present and therefore all B*15 typed donors are considered contra-indicated for transplantation for this patient.

The expert assignment given by the HLA data dictionary (available as searchable form in the IPD-IMGT/HLA database) has been generated with the data obtained from different sources, including WHO Nomenclature Committee [12] the International Cell Exchange (UCLA program), National Marrow Donor Program, and recent publications and individual laboratories [10]. There are 728 different B*15 alleles in the database (version 3.38.0), 37 of them are null or questionable alleles, and 516 are B15 antigens (based on 2 field allele assignment). Of these 516 antigens, 394 are without any serological assignment

in the IPD-IMGT/HLA database (version 3.38.0). Neural network (NN) analysis increased the number of alleles assigned to serological split specificities [13] but these results have already been included in the 2008 edition of the HLA data dictionary, as seen in the IPD-IMGT/HLA database website. In this study, we aim to provide a reliable, fast and straightforward method to predict serological specificities of HLA-B*15 alleles based on amino acid sequence patterns. For this purpose, we identified specific amino acid patterns for each B*15 serological subtype to predict the serological equivalents of B*15 alleles. In addition, we identified two new HLA-B*15 alleles by full length allele specific Sanger sequencing [14]; HLA-B*15:03:01:03 and HLA-B*15:16:01:03 and for both alleles we predicted the serological specificity using our new approach and confirmed this by HLA class I serological typing.

Materials and Methods

Dataset

HLA-B15 antigens: HLA-B*15 amino acid sequences included in the analysis have been obtained from the IPD-IMGT/HLA Database (release 3.38.0), and each B*15 allele that differs in the second field was considered as a potentially different antigen. Null and questionable alleles were excluded. In this way, we identified 516 different HLA-B15 antigens in the IPD-IMGT/HLA Database (version 3.38.0) (Figure 1a).

B15 antigens used for analysis (B*15 alleles in the dictionary); The amino acid patterns of the different serological B15 split antigens were analysed, using only alleles with defined serological assignments in the HLA dictionary. The serological information from the expert assignment was used, and when this assignment is not conclusive, then the sources WHO and NN assignment were used. The original 2008 report of HLA data dictionary included 123 HLA-B*15 alleles, but the previously B*95:30 is modified to B*15:27:02, resulting in 122 antigen-based HLA-B*15 alleles in total (Figure 1a). Out of the 122, 105 have been assigned to a certain serological split antigen and these alleles have been used in this study to analyse the amino acid motif for each serological subtype (Figures 1b and c and table 1). For 6 alleles, the expert assigned serological type is discrepant since the sources WHO assignment and Neural network (NN) gave different serological assignments (Figure 1c and table 2). The remaining 11 alleles have undefined expert assigned type (Figure 1c and table 3).

B*15 alleles used for prediction (B*15 alleles **not** in the dictionary); 394 B*15 alleles in the IPD-IMGT/HLA were not included in the HLA data dictionary (Figure 1d). We used this set of alleles to predict the serological assignments by using the analysed amino acid patterns (Figure 1d and table 1).

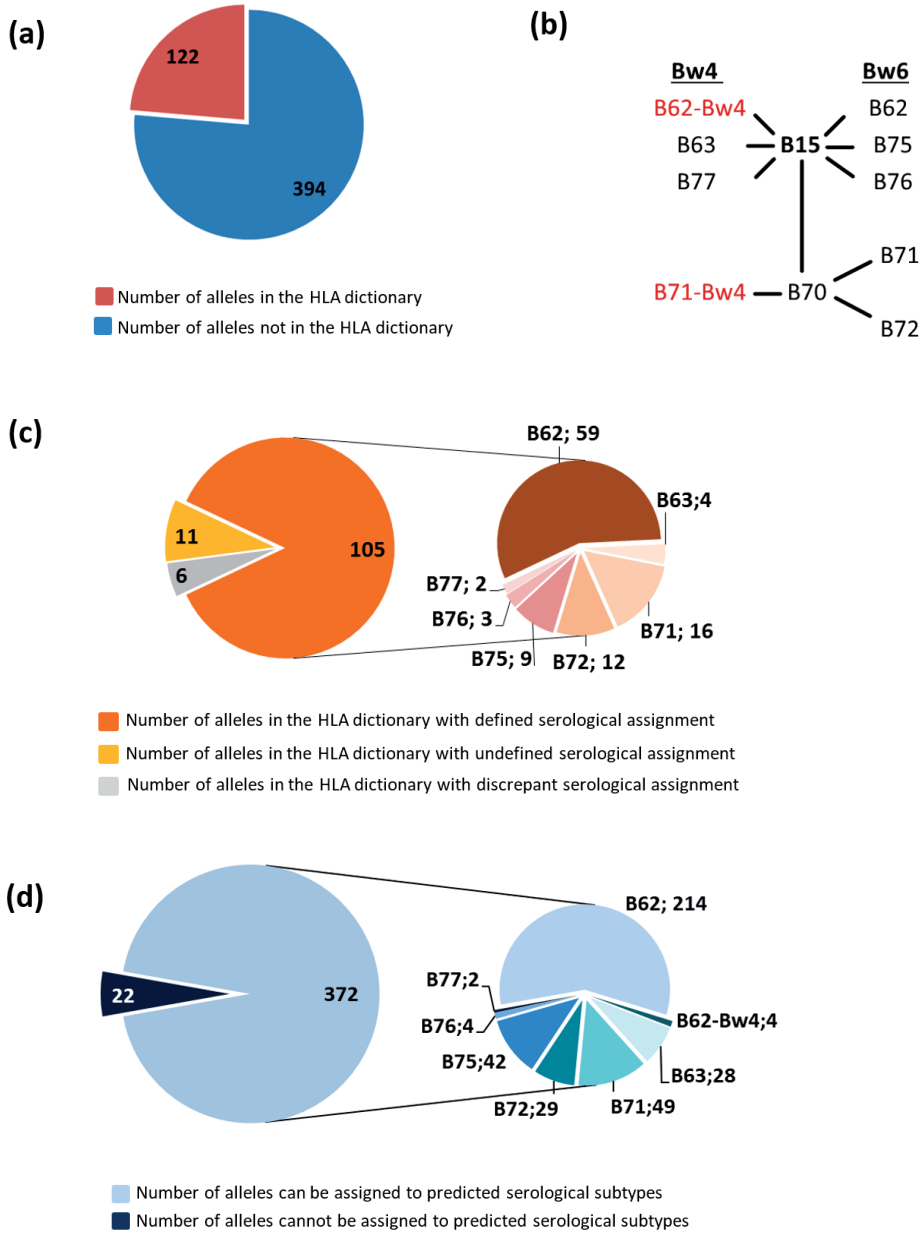


Figure 1. Overview of the HLA-B15 antigens and B15 serological specificities reported in the IPD-IMGT/HLA database. (a) Number of HLA-B*15 alleles in IPD-IMGT/HLA database (release 3.38.0) present or absent in HLA dictionary. There are in total 516 HLA-B*15 alleles (based on 2 field allele assignment). (b) Serological subtypes of B15 antigen, clustered according to the Bw6 and Bw4 epitopes. The new serological types B62 with Bw4 and B71 with Bw4 epitope are added in red. (c) Distribution of the 122 B15 alleles present in the dictionary according to serological assignment (for details see supplemental table 1). (d) Distribution of the 386 B15 alleles absent from the dictionary according to predicted serological assignment (for details see supplemental table 2).

Sequence Based Typing (SBT)

For ultrahigh resolution typing, full-length allele-specific sequencing was performed by group-specific amplification and sequencing according to our previously published protocol [14]. In brief, allele group specific amplification was performed with primers in 5′ and 3′ untranslated regions followed by Sanger sequencing using generic sequencing primers in both forward and reverse direction by means of cycle sequencing. The 3730 DNA-analyzer was used for electrophoresis whereas analysis was performed with SeqPilot (JSI, Germany) and Lasergene (DNASTAR, Madison, Wisconsin) software, as previously described [14].

	Amino acid positions						
	exon 2						exon 3
Serological types	24	45-46	63	65-67	70	77/80-83	166-167
B62	A	MA	E	QIS/QIF/QIC	N	S,NLRG	EW
B62-Bw4	A	MA	E	QIS	N	**N,IALR**	EW
B63	A	MA	E	**RNM**	**S**	**N,IALR**	EW
B71	**S**	**EE**	**N**	QIC/QIF/QIS	N	S,NLRG	EW
B71-Bw4	**S**	**EE**	**N**	QIC	N	**N,IALR**	EW
B72	**S**	**EE**	**E**	QIS	N	S,NLRG	EW
B75	A	MA	**N**	Q I S / Q I Y / QIC	N	S,NLRG	EW
B76	A	MA	E	QIS	N	S,NLRG	**DG/ES**
B77	A	MA	**N**	QIS	N	**N,IALR**	EW

Table 1. Overview of characteristic amino acid motifs based on the alignment of B15 serological subtypes. Amino acid motifs specific for each serological subtype are bold and grey shaded. Numbers represent the amino acid positions according to alignments of full length amino acid sequences.

	Amino acid positions							The information in the HLA dictionary			
	exon 2						exon 3	**Expert**	**WHO**	**NN**	**Based on**
Alleles	**24**	**45-46**	**63**	**65-67**	**70**	**77/80-83**	**166-167**	**assigned**	**Assigned**	**assigned**	**aa pattern**
B*15:08	A	MA	**N**	QIF	N	S, NLRG	EW	B75/62	B75(15)	B15	B75
B*15:15	A	MA	**N**	QIS	N	S, NLRG	EW	B75/62	B62(15)	B15	B75
B*15:23	**S**	**EE**	**N**	QIC	N	**N, IALR**	EW	B70/B5/ Blank	-	Not assigned	B71 Bw4
B*15:43	A	MA	E	QIS	N	**D, TLLR**	EW	B15	-	B62	B62 Bw4
B*15:87	A	MA	E	QIS	N	**S, IALR**	EW	B15		B15	B62 Bw4
B*15:115	**S**	**EE**	**N**	QIC	N	**S, TALR**	EW	B70	-	B70	B71 Bw4

Table 2. Overview of the six B*15 alleles with discrepant serological assignment in the dictionary, with on the left side the characteristic amino acid motifs and on the right side the information from the dictionary and the serological assignment prediction based on the amino acid pattern. Amino acid motifs specific for each serological subtype are bold and grey shaded. Numbers indicate the amino acid positions, WHO refers to typings assigned by World Health Organization, NN typings were assigned by Neural Network, aa refers to an amino acid pattern.

	Amino acid positions							The information in the HLA dictionary			
	exon 2						exon 3	**Expert**	**WHO**	**NN**	**Based on**
Alleles	**24**	**45-46**	**63**	**65-67**	**70**	**77/80-83**	**166-167**	**assigned**	**Assigned**	**assigned**	**aa pattern**
B*15:36	A	MA	E	QIS	N	N, TALR	EW	Undefined	-	B77	B62 Bw4
B*15:46	A	**KE**	E	QIS	N	S, NLRG	EW	Undefined	B72(70)	B15 B62	
B*15:52	S	EE	N	QIC	N	S, NLRG	EW	Undefined	B15	B71	B71
B*15:53	T	**KE**	E	QIS	N	S, NLRG	EW	Undefined	-	Not assigned	
B*15:62	S	EE	E	QIS	N	S, NLRG	EW	Undefined	-	B70 B72	B72
B*15:68	S	EE	E	QIS	N	S, NLRG	EW	Undefined	B35	B70 B72	B72
B*15:76	A	MA	N	QIY	**Q**	S, NLRG	EW	Undefined	-	B15	
B*15:86	A	MA	E	QIS	N	S, NLRG	EW	Undefined	-	B15 B62	B62
B*15:91	**S**	**EE**	E	QIS	N	S, NLRG	**ES**	Undefined	-	B70 B72 B12	
B*15:101	A	MA	E	QIY	**Q**	S, NLRG	ES	Undefined	-	B15 B46	
B*15:106	A	**KE**	E	QIS	N	S, NLRG	ES	Undefined	-	B15 B62	

Table 3. Overview of B*15 alleles with undefined serological assignment in the dictionary, with on the left side the characteristic amino acid motifs and on the right side the information from the dictionary and the serological assignment prediction based on the amino acid pattern. Unique amino acid motifs that prevent prediction of serological assignment are bold and grey shaded. Numbers indicate the amino acid positions, WHO means World Health Organization, NN means Neural Network, aa means amino acid.

	Amino acid positions						
				exon 2			exon 3
Alleles	**24**	**45-46**	**63**	**65-67**	**70**	**77 80-83**	**166-167**
B*15:143	A	**KE**	N	QIS	N	S, NLRG	EW
B*15:183	**T**	MA	E	QIS	N	S, NLRG	EW
B*15:202	A	**TA**	E	QIS	N	S, NLRG	EW
B*15:212	**T**	**KE**	E	QIS	N	S, NLRG	EW
B*15:239	A	**TE**	E	QIS	N	S, NLRG	EW
B*15:251	**S**	**MA**	E	QIS	N	S, NLRG	EW
B*15:259	**A**	**EE**	E	QIS	N	S, NLRG	EW
B*15:308	A	**TE**	N	QIS	N	S, NLRG	EW
B*15:336	**T**	MA	E	QIS	N	S, NLRG	EW
B*15:345	**T**	MA	N	QIS	N	S, NLRG	EW
B*15:376	S	**TE**	E	QIS	N	S, NLRG	EW
B*15:392	A	MA	**G**	QIS	N	S, NLRG	EW
B*15:429	**S**	**MA**	E	QIS	N	S, NLRG	EW
B*15:430	A	MA	E	QIS	N	**S, KLRG**	EW
B*15:434	S	**GE**	N	QIC	N	S, NLRG	EW
B*15:436	S	**GE**	N	QIC	N	S, NLRG	EW
B*15:504	**A**	**EE**	N	QIC	N	S, NLRG	EW
B*15:511	**T**	EE	E	QIS	N	S, NLRG	EW
B*15:525	A	**KE**	E	QIS	N	S, NLRG	EW
B*15:545	**S**	**MA**	E	QIC	N	S, NLRG	EW
B*15:553	**A**	**EE**	N	QIF	N	S, NLRG	EW
B*15:556	**S**	**MA**	E	QIS	N	S, NLRG	EW

Table 4. Overview of the 22 B*15 alleles with unique amino acid combinations preventing serological assignment prediction. Unique amino acid motifs that prevent prediction of serological assignment are bold and grey shaded. Numbers indicate the amino acid positions.

HLA class I Serological Typing

Heparinized blood was collected and lymphocytes were isolated by centrifugation on Ficoll-Hypaque. After counting the cells using pocH-100i (Sysmex) and adjusting to $4x10^6$ cells/ml, the serological typing was performed on this lymphocyte suspension using the standard NIH complement dependent cytotoxicity (CDC) assay and a local set of sera. In short, 1 ul cell suspension was added per well of typing plate, containing 1 µl of specific typing serum, and incubated 30 min at 20°C. Complement activation was initiated by addition of 5 µl rabbit complement (CEDARLANE®) and incubation at 20°C for 60 min. After incubation with complement, FluoroQuench™ (Acridine orange (AO)/ ethidium bromide (EB) (One Lambda)) was used for staining, reading the plates by fluorescence microscopy after 10 min incubation at room temperature in the dark. Plates were scored based on percentage of dead and live cells and evaluated for serological typing assignment.

Results

Identification of specific amino acid motifs for each B15 serological subtype

HLA-B15 represents one of the largest broad antigen groups with different serological subtypes, containing B62, B63, B75, B76 and B77 specificities and is also associated with B71 and B72 (belonging to the B70 broad antigen) (Figure 1b). Based on the 105 HLA-B*15 alleles with defined serological assignment in the HLA dictionary we were able to identify 7 characteristic amino acid motifs for each serological type. These characteristic amino acid motifs were located at positions 24, 45- 46, 63, 65-67, 70 and 166-167 (table 1). In addition, amino acids at position 77, 80, 81, 82 and 83 have already been described to characterize the Bw4 and Bw6 motifs [15]. Bw4 specificity has been defined by the presence of N, D or S at residue 77 and IALR, TLLR or TALR amino acid motifs at residues 80-83 [16]. Since these motifs facilitate distinction between serological subtypes, they were included in the analysis. B62 is the most common subtype and therefore used as a reference type. The difference in amino acid pattern compared to B62 are highlighted/colored in the table and these highlighted motifs are used to determine each serological subtype. The amino acid pattern of 'RNM and S' at locations 65-67 and 70 together with Bw4 motif enabled prediction of B63 split antigen. Only the location 65-67 was already sufficient to specify B63 serological subtypes since 'RNM' is highly conserved among B63 subgroup. Other serological subtypes all have one of the following combinations: QIC, QIF, QIS or QIY. For the B70 broad antigen the amino acids S and EE at locations 24 and 45-46 are specific, whereas amino acid 'N' at location 63 specifies B71 and 'E' indicates B72 split antigens. B75 and B77 can be distinguished from B62 by the presence of an N at position 63. B77 carries, in addition to this N, the Bw4 motif, making this antigen in fact a B75-Bw4 type. Lastly, B76 antigen can be discriminated from B62 by the absence of the amino acids EW at locations 166 and 167. In the three known B76 antigens (B*15:12, 15:14 and 15:19), two different

motifs are present, namely DG or ES. Since these amino acid patterns are quite distinct, it seems more logical that the B76 is defined as missing the EW motif, which fits with the serological finding that the B76 bearing cells react with less sera than the B62.

Furthermore, based on the 6 discrepant types (table 2), we were able to identify two additional variants: B62-Bw4 which carries both a B62 and Bw4 pattern (B*15:43 and B*15:87) and B71-Bw4, which bears both B71 and Bw4 pattern (B*15:23 and B*15:115). These variants are included in the list of serological types (Figure 1b and table 1). The remaining 2 discrepant HLA-B*15 alleles (B*15:08 and B*15:15) could be assigned to the serological subtype B75 according to their amino acid pattern ('N' at location 63 with Bw6 motif) (tables 1 and 2).

The amino acid patterns of the 11 HLA-B15 alleles that were unassigned in the dictionary are shown in table 3. Based on this pattern 5 of them could be assigned to one of the 9 defined serological types. The remaining 6 alleles revealed unique amino acid patterns at designated locations, indicated in grey in table 3, and therefore the serological type could not be reliably predicted for these alleles.

Verification of B15 serological splits of 2 new B*15 alleles

Two new HLA alleles; HLA-B*15:03:01:03 and HLA-B*15:16:01:03 were identified during routine high resolution DNA typing of a kidney patient and a donor by full length allele specific sequencing method (Supplementary Figure 1). The full-length sequences were confirmed by sequencing 2 different polymerase chain reaction products, from both individuals. The HLA-B*15:03:01:03 was most similar to HLA-B*15:03:01:02 with one nucleotide difference at position 1054 (G>T) in intron 3 while HLA-B*15:16:01:03 resembled HLA-B*15:16:01:01 with a single nucleotide change at position 119 (G>C) in intron 1 (Supl. Figure 1). In both cases the new allele showed no amino acid differences with the most similar allele, since the single nucleotide changes were detected in the introns. The genomic sequences of these new alleles have been submitted to the EMBL Nucleotide Sequence Database (accession numbers LT618821 and LT898179 respectively) using a new allele submission tool called saddlebags [17] and to the IPD-IMGT/HLA database. The names HLA-B*15:03:01:03 and HLA-B*15:16:01:03 have been officially assigned by the World Health Organization (WHO) Nomenclature Committee [12]. Serological typing of these two samples was performed and confirmed the presence of B70 in the B*15:03:01:03 sample and B63 in the B*15:16:01:03 individual, as was already assigned by the experts for B*15:03 and B*15:16 respectively, as well as predicted by the amino acid composition. The B*15:03:01:03 is most probably B72 subtype, but this could not be confirmed by CDC method due to lack of B71 and B72 specific sera in our laboratory.

Prediction of serological specificities of B*15 alleles without defined serological types
Amino acid motifs were applied to the B*15 alleles that were not included in the HLA dictionary and were without assignment of serological subtype. 372 alleles out of 394 could be predicted to a serological subtype by using the amino acid pattern shown in table 1. The alleles and the prediction are indicated in supplementary table 1. The remaining 22 unassigned alleles revealed unique amino acid combinations at the determined amino acid motif positions that were not present in the serological assigned alleles (Table 4). In 20 of these cases, it concerns either a change of amino acid 24, not being A or S, a change of amino acid 45, not being E or M, or a change of the combination of these two motifs, not being A-MA or S-EE. The influence of these amino acid changes compared to known HLA-B15 subtypes on the serological reaction is unclear and could not be determined by serological typing due to lack of viable cells with these B15 alleles.

Discussion

In this study, we provide a straightforward approach to predict serological splits of HLA-B*15 alleles based on amino acid polymorphisms. HLA-B*15 represents the largest broad antigen comprising 9 different serological splits. Currently, 516 HLA-B15 alleles are found in the IPD-IMGT/HLA database (release 3.38.0 [11]), while no information is available regarding serological subtype of 394 alleles. Advancements in molecular techniques have led to a switch from serological typing to DNA typing of HLA alleles. Although DNA typing enables easy and specific allelic distinctions, it does not provide information about the corresponding serological type of HLA antigens by itself. Therefore, despite the rapid increase in the number of identified HLA alleles, information about their serological subtype remains limited. In addition, the scarcity of sera, especially with anti-HLA antibodies against split antigens, limits serological methods to determine serological splits. For instance, due to the unavailability of a B72 specific antiserum in our laboratory, the serological type of the new HLA-B*15 allele, HLA-B*15:03:01:03, could be only assigned as B70 by CDC serotyping. However, based on amino acid motifs identified in this study, we could now assign this new allele to the B72 serological subtype. Thus, our new approach facilitates the determination of serological subtype of HLA-B*15 alleles based on their DNA sequence.

In 2003, the neural network (NN) analysis has been developed by a machine learning model with the polypeptide sequences of HLA-A, HLA-B and HLA-DRB1 alleles alongside well-defined serological subtypes in order to predict the split assignments of 393 alleles [13]. Subsequently, this computational model was able to predict serological assignments for most alleles (95% HLA-A, 85% HLA-B, 96% HLA-DRB1). The information from this analysis has been included in the HLA Dictionary in 2008 [10]. However, this

method is not generally available and demands expert skills to be executed. Therefore, our new approach based on the distinct and specific amino acid patterns of each B15 serological split provides an easy and practical solution. After evaluation of the serological subtypes of 105 alleles and determining the defining amino acid motifs, it was possible to predict not only the serological types of 6 alleles with conflicting information about their serological group in the HLA dictionary, but also 5 out of 11 HLA-B*15 alleles which were undefined according to expert assignment and 372 out of 394 B*15 alleles that were not yet assigned. The remaining 22 alleles carry different amino acids than the identified motifs, warranting further analysis for definite serological assignments.

The number of antigens assigned to each serological split greatly varies. In total, B62 is the largest subtype since it is the serological equivalent of 274 different B*15 alleles (both in dictionary and predicted by us). The second subtype is B71 with 66 alleles assigned to this split. Furthermore, different B*15 alleles have been assigned 43 to B72, 53 to B75, 32 to B63, 7 to B76 and 4 to B77 serological subtype. The new types B62-Bw4 and B71-Bw4 were assigned for 7 and 2 B*15 alleles. From the total of 516 B*15 alleles we were able to assign 488 to a specific serological split, whereas 28 remain undetermined, because their amino acid patterns do not match with the identified ones. Since there is no empirical evidence regarding their reactivity against any serological group, it is currently not possible to reliably predict their serological subtype.

The serological subtype of HLA alleles is of particular interest in the setting of organ and stem cell transplantation, especially when the patients' sera contain antibodies against a split antigen. Current techniques for antibody profiling of patient sera, such as Luminex bead assays allow the identification of anti-HLA antibodies at the split level [18]. Since the presence of donor specific antibodies (DSA) in the patient serum has been associated with graft failure, the identification of donor HLA type at the split level remains to be crucial for successful transplantation outcome [8]. For this reason, Eurotransplant, a non-profit organ allocation organization of 8 European member countries, recommends transplantation centers to report HLA typing for both organ donors and patients at the serological split antigen level in order to obtain optimal organ allocation (ETRL Newsletter issue 9).

HLA typing of donor and patients for solid organ transplantation is at the moment generally performed at low resolution. In this paper, we used high resolution full length sequencing results to determine the crucial positions needed for serological subtyping of B15. With this analysis we can now determine which method will obtain sufficient information to perform the serological subtyping of this antigen. Since the NGS methods with full length sequencing has become rather cost effective, these methods are ideal for patients and living donors, because there are less time constraints. For deceased donors this is not possible yet, because even the fastest NGS method will still take more than 24

hrs. Therefore, we have also investigated whether the crucial amino acid positions can be determined with the faster typing method of real time PCR using the LinkSeq technique. We have therefore checked the primer recognition sites, recognizing one or more of the crucial amino acids determining the B15 subtypes. All primer combinations that recognized a B15 subtype were used for the primary analysis to determine which combinations were crucial to make the difference between the different serological subtypes. These crucial primer combinations are indicated in supplementary table 3, together with the reactions of the different B15 subtypes. From this table it is clear that each B15 subtype can be recognized by a unique pattern of positive and negative reactions. This positive/negative reaction pattern fits perfectly with the specific amino acid patterns identified in this study. Therefore, accurate prediction of B15 subtypes from a potential deceased donor is definitely possible, even during night shifts, revealing the possibility of exclusion based on the presence of antibodies against a certain B15 subtype for a potential recipient.

The presence of donor specific antibodies is also important in the setting of stem cell transplantation (SCT) as it can influence donor engraftment and transplant outcome [19] and is especially important in the haploidentical transplantation setting, where one HLA haplotype is completely mismatched with the patient. For SCT both patient and donor high resolution HLA typing is performed in most centers, enabling conversion of HLA-B*15 alleles to serological split equivalents with our approach, to determine whether the antibodies present in the patient are indeed donor specific.

In conclusion, we provide a straightforward practical approach to predict serological subtypes of HLA-B*15 alleles. This approach is useful for patients waiting for a stem cell - or solid organ transplant, that have antibodies against B15 subtypes, by predicting the donors B15 subtype and therewith circumventing donor specific antibodies that have an impact on graft survival and acute and chronic rejection.

Acknowledgement

The authors wish to thank Christel Meertens for her contribution in HLA typing and submitting the genomic sequence of two new alleles and Lisette Groeneveld for helping to collect samples for Linkseq analysis. Diana van Bakel for her contribution to prepare the manuscript for submission.

References:

1. Terasaki, P.I. and J.D. McClelland, *Microdroplet Assay of Human Serum Cytotoxins.* Nature, 1964. **204**: p. 998-1000.
2. Laundy, G.J., C.C. Entwistle, and K. Hassenkamp, *Bu--a new antigen at the HLA-B locus.* Tissue Antigens, 1978. **11**(2): p. 121-8.
3. Hildebrand, W.H., *et al.*, *HLA-B15: a widespread and diverse family of HLA-B alleles.* Tissue Antigens, 1994. **43**(4): p. 209-18.
4. Lin, L., *et al*., *Further molecular diversity in the HLA-B15 group.* Tissue Antigens, 1996. **47**(4): p. 265-74.
5. Steiner, N., *et al*., *HLA-B alleles associated with the B15 serologically defined antigens.* Hum Immunol, 1997. **56**(1-2): p. 84-93.
6. Elsner, H.A., *et al*., *Identification of the novel allele HLA-B*1546 which belongs to the serological B72 type: implications for bone marrow transplantation.* Tissue Antigens, 2000. **55**(1): p. 83-5.
7. Erlich, H.A., G. Opelz, and J. Hansen, *HLA DNA typing and transplantation.* Immunity, 2001. **14**(4): p. 347-56.
8. Michielsen, L.A., *et al*., *A paired kidney analysis on the impact of pre-transplant anti-HLA antibodies on graft survival.* Nephrol Dial Transplant, 2019. **34**(6): p. 1056-1063.
9. DeVos, J.M., *et al*., *De novo donor specific antibodies and patient outcomes in renal transplantation.* Clin Transpl, 2011: p. 351-8.
10. Holdsworth, R., *et al*., *The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens.* Tissue Antigens, 2009. **73**(2): p. 95-170.
11. Robinson, J., *et al*., *The IPD and IMGT/HLA database: allele variant databases.* Nucleic Acids Res, 2015. **43**(Database issue): p. D423-31.
12. Marsh, S.G.E., *et al*., *Nomenclature for factors of the HLA system, 2010.* Tissue Antigens, 2010. **75**(4): p. 291-455.
13. Maiers, M., *et al*., *Use of a neural network to assign serologic specificities to HLA-A, -B and -DRB1 allelic products.* Tissue Antigens, 2003. **62**(1): p. 21-47.
14. Voorter, C.E., F. Palusci, and M.G. Tilanus, *Sequence-based typing of HLA: an improved group-specific full-length gene sequencing approach.* Methods Mol Biol, 2014. **1109**: p. 101-14.
15. Muller, C.A., *et al*., *Genetic and serological heterogeneity of the supertypic HLA-B locus specificities Bw4 and Bw6.* Immunogenetics, 1989. **30**(3): p. 200-7.
16. Lutz, C.T., *Human leukocyte antigen Bw4 and Bw6 epitopes recognized by antibodies and natural killer cells.* Curr Opin Organ Transplant, 2014. **19**(4): p. 436-41.
17. Matern, B.M., *et al*., *Saddlebags: A software interface for submitting full-length HLA allele sequences to the EMBL-ENA nucleotide database.* HLA, 2018. **91**(1): p. 29-35.
18. Picascia, A., T. Infante, and C. Napoli, *Luminex and antibody detection in kidney transplantation.* Clin Exp Nephrol, 2012. **16**(3): p. 373-81.

19. Ciurea, S.O., *et al.*, *The European Society for Blood and Marrow Transplantation (EBMT) Consensus Guidelines for the Detection and Treatment of Donor-specific Anti-HLA Antibodies (DSA) in Haploidentical Hematopoietic Cell Transplantation.* Bone Marrow Transplant, 2018. **53**(5): p. 521-534.

CHAPTER 10

10

General Discussion

Discussion

This thesis has touched upon many scientific and analytical themes through a bioinformatic lens. It explores how we can identify, store, and catalog HLA polymorphism (**Chapters 2,3,5**), and new ways to think about the sequence patterns and role the polymorphism plays in human biology (**Chapters 4, 7, 9**). It touches on the basic nature of DNA, and how polymorphism affects the peptide sequence of the resulting protein (**Chapters 7,9**), and how these protein-coding genes are arranged within the MHC (**Chapters 6,7,8**). It discusses how non coding polymorphism can affect the expression of HLA on the cell surface, and it explores how we can use this knowledge of polymorphism and apply it to a transplantation immunology setting (**Chapters 4, 5, 9**), ultimately improving patient outcomes. Bioinformatics has played a critical role in the development of these scientific questions, as well as the analytical approaches used to answer them. This thesis shows an urgent need to further develop bioinformatics science, and create new tools and approaches for molecular analysis of DNA sequences, and how to manage and interpret the growing pool of available data.

DNA Sequencing & Molecular Analysis

The techniques of DNA sequencing are regularly evolving. Like in biological evolution, new sequence platforms (*e.g.* MinION) are selected and improved, and platforms that are not fit for survival (*e.g.* Roche 454) tend to disappear. The succession of developments in new technologies bring with them cost reductions, as well as improvements in the throughput and quality of the generated data. The four main costs in sequencing are equipment, reagents, labor, and data analysis. The early 2000s saw the most rapid cost reductions due to developments in technology, but costs continue to decrease in recent years. The transition from Sanger to NGS sequencing saw some reductions in reagent costs, but the labor cost reduction brought by the high throughput capabilities is most apparent. Those costs are further reduced by the introduction of nanopore sequencing, which also has very low upfront equipment costs, but the platform provides new challenges in data analysis. The cost reductions have slowed in recent years,[1] likely because all of these platforms still require comparatively expensive PCR amplification. It is clear that the most effective cost reductions are provided by major platform shifts, especially those that implement new sequencing technologies or eliminate expensive steps like PCR. The concept of a $1000 human genome has been realized, and is even being pushed further. In a recent Nanopore Community Meeting in New York, Clive Brown, the CTO of Oxford Nanopore sequencing, claimed that the current nanopore technology allows $725 genomes, and with minor improvements the cost could push as low as $145.[2] The current goal is a sub-$100 human genome, but there's no reason to assume this is an absolute minimum cost.

As costs decrease, the opportunities for analysis of genomes are vastly improved. As DNA sequencing and typing become more routine, the financial burdens of performing DNA-based analysis in a clinical setting becomes less restrictive, and thus the amount of available data is increased, extending our understanding of the existing polymorphism. Full genome sequencing can circumvent the need for GWAS studies by increasing the pool of observable polymorphism, and by directly identifying the relationship between disease and genomes. The availability of full genome sequence allows researchers to compare the differences in genomics between populations and species. Understanding the distribution of polymorphism and how it has evolved provides context to our evolutionary history, and indicates areas for exploration in the future.

Aside from cost reductions, new platforms bring with them improvements to research capabilities. Traditional shotgun-based short read sequencers generate short reads, perhaps a few hundred base pairs in length, which can identify unphased polymorphism within a small region. MinION sequencing is capable of sequencing long pieces of continuous DNA, tens or hundreds of thousands of base pairs in length.[3] These individual reads can span across complicated genomic features such as homopolymers and STRs, and allows the phasing of polymorphism even across multiple genes. It can clarify the organization of genes, and help to identify historic recombinations and patterns of evolution.

DNA sequencing technologies give the ability to directly observe sequence polymorphism. It allows the identification of SNPs, and provides techniques to infer the function and features of SNP polymorphism. We can explore whether polymorphism affects splicing behavior or expression levels of mRNA transcripts. Sequencing allows us to ask questions about what polymorphism is common or rare, or is present in higher frequencies in distinct populations. We can identify polymorphism that belongs to a gene, or polymorphism that lies in intergenic non-coding regions. And, most pertinently, we can explore how this polymorphism relates to the HLA genes and immunology.

Molecular Analysis & HLA

Patterns in graft rejection were identified by Dr. Rose Payne and Dr. Jon van Rood in the late 1950s,[4,5] marking the beginning of serology-based HLA typing. Ever since, the level of understanding and the resolution of identified HLA types has rapidly increased. Typing by serology has been, and continues to be, gradually replaced by DNA-based typing methods. It has become clear that increased resolution in HLA typing provides more understanding of the nature of HLA genes and how to match them for transplantation compatibility.

The polymorphism that exists in HLA genes has been studied extensively, but the alleles represented in IPD-IMGT/HLA only represents a small portion of the total HLA

polymorphism.[6] Identifying the full extent, and understanding the polymorphism in the context of its function has only been touched upon. We have some idea about the function of an HLA molecule and how it interacts with the immune system and human disease, but there are continued lessons to learn, and we will never reach a complete understanding of all the complexities in this system. It is clear that high resolution sequence based typing is currently the most effective way to study these genes, and to match for SCT.[7]

The full-length sequencing of HLA alleles regularly reveals novel allele sequences, and for many laboratories the identification of novel polymorphism is a routine process. It seems that every study related to HLA polymorphism, especially with a novel population or less-studied loci (**Chapter 6**), identifies new polymorphism that has never been represented before. In order to accurately compare HLA alleles and determine compatibility, it is necessary to store and curate the data in an effective way. IPD-IMGT/HLA is the only HLA database with WHO-assigned HLA nomenclature, and it is the best available repository of allele sequence, which is made clear by its widespread adoption in clinical diagnostics and integration with commercial software packages. Allele sequences, with corresponding alignments and relevant metadata, are easily available to anyone interested in HLA. Bioinformaticians can download allele sequences and analyze them for more specific scientific questions, which has proven to be very useful for analysis in all projects in this thesis. The WHO nomenclature committee has come up with effective, albeit imperfect, nomenclature to represent polymorphism in a biologically meaningful way.[8]

Having the wealth of HLA polymorphism is valuable to individuals and the community, but there are reasons laboratories may not submit novel polymorphism. First, filling the database with all available data may present challenges in patient / donor allele matching; with so many available allele sequences, it can be nearly impossible to identify a perfect HLA match. Secondly, laboratories are often unwilling to commit the labor necessary to create high-quality consensus sequences, gather relevant metadata and specific protocol details, and submit novel sequences. The first difficulty will be resolved in future studies, which identify what polymorphism is the most immunogenic, and which polymorphism is the most important consideration for transplantation. Analyzing sequences in the context of the most important polymorphism eases the burden of having so much available data. The second difficulty can be resolved by easing the process of allele submission. Saddlebags (**Chapter 2**) represents an attempt to ease the process of annotating and submitting allele sequences. Adopting the software to support batch allele submission to IPD-IMGT/HLA remains a major goal, in order to improve the amount of available high-quality allele sequences.

MHC and HLA Gene Organization

The MHC is a very complicated region of the human genome.[9,10] It is dense with genes, including HLA and genes related to inflammation and immune function, as well as some genes with seemingly unrelated function, such as olfactory receptors. These genes are arranged in common patterns, but early maps of the MHC are based on homozygous cell lines, and likely do not represent the diversity in MHC gene organization. Matching of MHC haplotypes in addition to HLA genotype is important in Stem Cell Transplantation,[11] with benefits that are likely related to the interactions between the HLA and non-HLA genes within the region. Even if the HLA genes were absent from the region, the MHC is a fascinating collection of genes and provides countless opportunities for interesting studies.

Diversity of HLA is apparent when aligning and comparing polymorphism in allele sequences, and much can be learned by studying the individual polymorphisms and alleles in HLA sequences. Meiotic recombinations, however, create a different kind of diversity in this region, causing changes in haplotype arrangements.[12-14] Recombinations can take place in regions between genes, resulting in gene reorganizations which might not be detected by normal HLA genotyping,[15] but recombinations can also happen in regions within genes (**Chapter 3**), resulting in hybrid allele sequences. Recombinations are a major driving force behind much of the diversity in HLA sequences, and due to homology between the HLA loci and the variability in haplotype patterns, it can raise questions about what polymorphism belongs to what gene locus.

Defining a gene locus depends on a few factors. A gene is a region of DNA that encodes a protein. It contains expressed exons, which are translated into proteins, and it often contains introns, which are spliced from the mRNA sequence and do not affect the amino acid sequence. The gene is surrounded by 5′ and 3′ UTR sequences, which have functions which are probably related to protein expression. But there still remains the question of how much sequence is sufficient to define a gene. Certainly protein-encoding exons that define the sequence and structure of the molecule should be included, and it is important to know intron sequences to determine the splicing behavior, but how much UTR sequence is sufficient? The longest 5′ UTR sequence in IPD-IMGT/HLA is 17,561bp (HLA-DQA1), and the longest 3′ UTR sequence is 11,200bp (HLA-DPA1). These are extreme examples, but illustrate the difficulty in defining the region that contains an entire gene. I would posit that any sequence that can be demonstrated to affect the expression or behavior of the resulting protein, or any sequence that is in a mRNA transcript in the cytoplasm, should be included in the gene definition.

The 5′ UTR contains the promoter sequence for most genes, including HLA.[16] The promoter sequence, transcription start site, and regulatory regions *e.g.* activators and repressors,

exist within the 5′ UTR sequence. The length of the complete regulatory region can vary depending on a gene, but it may be possible to estimate an upper bound based on gene pairs with bidirectional[17] promoters. A convenient example is the DPA1~DPB1 region, where two head-to-head genes are separated by 2.5kb of intergenic 5′ sequence (**Chapter 8**). This suggests that 2.5kb could be argued as an acceptable 5′ UTR region to represent a gene.

The 3′ UTR is perhaps less clear, although there is significant evidence for functional sequence in this region. The 3′ UTR contains polyadenylation sites,[18] which can affect expression regulation by altering the poly-A tails on mRNA transcripts. In HLA-DPB1, the rs9277534 SNP (4989bp downstream of HLA-DPB1), has been shown to be correlated with DPB1 expression levels,[19] and subsequent transplantation outcomes. It has been suggested that this SNP may directly affect expression by interacting with microRNA sequences,[20] but it is not completely clear if this is the functional polymorphism, or represents a SNP with linkage disequilibrium to the functional differences. Deletions in the 3′ UTR of HLA-G have been correlated with expression levels of soluble HLA-G, which may have effects on the success of a pregnancy,[21] or with susceptibility to parasite infection.[22] It could be hypothesized that the 5′ UTR region is primarily responsible for mRNA transcript regulation, and the 3′ UTR sequence is primarily responsible for post-transcriptional and translational regulation, as evidenced by polymorphism related to expression found in both the 5′[23] and 3′[24] UTRs of HLA-C. Furthermore, polymorphism within the 3′ UTR is significant in identifying the genomic context of a gene,[25,26] and thus downstream polymorphism is an important indicator of haplotypes (**Chapters 6 and 7**). All of this functional sequence should be represented when considering a complete gene.

It is clear that there are ambiguities in how we represent genes, and the overlapping nature of coding sequence within the MHC creates many questions. The arrangement of the genes relative to each other is not a completely solved problem. I believe that the ultimate solution for storing and presenting MHC reference sequences is based on complete and continuous MHC sequences. If we have unambiguous reference sequence across the entire region, we can more completely compare how an individual's MHC relates to the rest of the population. This is a lofty goal, that is not feasible with current laboratory and computational techniques. But techniques such as Region Specific Extraction[27] will improve, and cooperative efforts like Dr. Paul Norman's 2021 Amsterdam workshop component "Creating Fully Representative MHC Reference Haplotypes" are valuable steps to achieving this goal.

It has become clear that the MHC is a dynamic region. It is not cold and static, and it does not have a constant or consistent arrangement. Although haplotype patterns can be conserved over long periods of human evolution,[28] genes can be duplicated, recombined,

and translocated, resulting in completely novel arrangements. The HLA gene loci as we understand have been assigned are based on sequence homology and similarities in biological function. While this is an established and effective way to represent HLA polymorphism, it is useful to think of the HLA genes as flexible genes that lie in a flexible MHC region. The extent of how flexibility of the region will be explored in the future.

Immunogenetics

Matching of HLA has been clearly demonstrated to be effective for Stem Cell Transplantation,[29] with better outcomes with higher resolution typing.[7] Better results are obtained still by matching of phased haplotypes in addition to matched high resolution gene sequences.[11] It is likely that by matching the HLA loci within the MHC, we are also matching sequences that are in linkage disequilibrium with the typed regions, which indicates more matched sequence. The importance of haplotypes and linkage disequilibrium may be most apparent in the rise of haplo-transplants.[30-32] While these transplants do cause a donor T cell response against the recipient's mismatched HLA, there seems to be some apparent protective effect provided by the single matched HLA haplotype. The mechanism of the protection is poorly understood, but it could be hypothesized that the interactions between the linked loci in the matched haplotype have an effect of silencing the genes on the mismatched haplotype. In general, for the best possible matching and for the identification of permissive mismatches, it is important to identify what polymorphism is the most clinically relevant. It is important to consider which polymorphism is correlated with disease, and what polymorphism is the most likely to generate an immune response if mismatched.

Exon 2 (class II HLA) and exons 2 and 3 (class I HLA) are the regions that encode the antigen presentation domain, and are therefore the most important polymorphism to identify, but it is important to type polymorphism outside this region as well. The region of exon 4 of HLA-A contains a C homopolymer region, and polymorphism here can result in null alleles (**Chapter 3**). The rs9277534 expression SNP of HLA-DPB1 provides a clear indicator of polymorphism that is related to the function of a gene. While the function of this polymorphism is still being explored,[20] laboratories are already performing typing methods based on this polymorphism.[33]

Considering the flexible nature of HLA genes, and the fact that alleles at one loci can share functional epitopes (*e.g.* Bw4) with alleles at different loci, it can be more tangible to compare alleles at a more functional level. This is the concept behind epitope matching; Comparing epitope sequences between HLA molecules provides a more tangible way to compare HLA alleles. The studies in this thesis explore the differences in amino acid sequences of HLA alleles (**Chapters 7, 8, and 9**), but only hint that these epitopes may be the target of antibodies. Determining which epitopes are the actual target of

antibodies is critical to understanding the differences in immunogenicity of HLA alleles and considerations for transplantation matching. This illustrates a shift in thinking about HLA matching, from matching based on HLA alleles to matching based on differences in epitopes or the SNPs that define them.

Epitope matching is not, of course, a new idea.[34] In the HLA-DRB genes, polymorphism in the exon sequences has direct correlation with allele groups. Alleles with similar exon sequences are clustered into allele groups, which is mostly correlated with serological subtypes. This is an example of clustering allele sequences based on epitopes, and these clusters can often be used to match patient and donor HLA. In the case of HLA-DP, allele clustering is not based on serological information, because we haven't identified antibodies that correspond to distinct HLA-DP alleles. The lack of serological typing means that we cannot form HLA-DPB1 allele groups based on antigen differences. The nomenclature indicates allele groups based on some sequence homology, but difficulties in identifying general patterns means that HLA-DPB1 nomenclature does not indicate protein function, and cannot be used for matching patient and donor alleles in a conventional sense. However, there have been efforts to cluster and match HLA-DPB1 alleles based on functional polymorphism. This thesis has discussed that polymorphism within the 5′ UTR can indicate haplotypes, which are correlated with HLA-DPB1 hypervariable epitope regions. (**Chapter 8**) This helps identify allele clusters, which will be explored to find their role in matching for transplantation. This may ultimately indicate a method to match HLA-DP based on functional polymorphism.

Collaboration

It is increasingly clear that science is not done in isolation, and collaboration is the key for successful projects. Tackling complex projects cannot be done by an individual, and the benefit of the expertise of others plays a major role in the success of a project. Every one of the projects in this thesis required collaboration, often international collaboration, with other individuals in the field.

In order for the efforts of international researchers to be effective, it is necessary for some coordination and consensus. The International HLA and Immunogenetics Workshop (IHIWS) represents a culmination of collaboration in this field. The first international workshop was hosted in Durham, North Carolina in 1964, with the aim of coming to a consensus on the identification of serologically relevant antigens. The workshop will continue with the 18th workshop in Amsterdam, in 2021. The workshop is a recurring theme in this thesis. The sequencing of full-length HLA alleles (**Chapter 3**) was the result of a workshop component at the 17th workshop, where Saddlebags (**Chapter 2**) was used for the submission of allele sequences to EMBL ENA database. Sequencing of the

HLA-DPA1~Promoter~HLA-DPB1 haplotypes (**Chapter 8**) will be continued at the 18th workshop as well.

Bioinformatics is by definition a collaborative effort, as it is often described as an interface between biological and computational sciences. It is the bridge between biological scientific questions and the use of algorithms and computational methods to solve them. When creating in-silico analysis tools and software, there is a benefit in designing software to be reusable and flexible. NMDP Bioinformatics / CIBMTR has a major focus on creating an environment of usable and sharable services for the community.[35] There is extensive work being performed on how to transmit and communicate high-quality genotyping data,[36-38], and tools for annotating HLA alleles or share allele and haplotype frequencies.[39] These tools are not kept for internal analysis, but are hosted online, and members of the community are encouraged to use and share them, or possibly even improve them. This spirit of collaboration is most apparent at the Data Standards Hackathon (DASH) events. At a DASH hackathon, members of the community, whether focused on bioinformatics analysis or not, come together to create or discuss software tools and data standards. This spirit of collaboration has encouraged me to keep all of the software that I write as open source,[40] with the hope that it is useful or reusable in a future study.

Future Perspectives

The studies presented in this thesis represent a single point in a continuous timeline. All of these studies are performed in the context of both past and future studies; although are carried out based on the current knowledge, formed by answers that have been obtained in the past, they are also carried out with an eye to the future. This research was performed with the intention that these discoveries will play a role in the questions that are asked in the future.

The progression of sequencing technologies and their technological improvements are constantly evolving. It is convenient to attach names like "next generation," "third generation," "high resolution," "ultra-depth," or "full-length" to a technology or group of technologies, but these phrases are only meaningful when considered in the context of technological progression. The current generation of technologies are just that, the current generation. We will continue to see paradigm shifts, improvements in the methodologies, accuracies, cost, and throughput of sequencing techniques that will further increase the available data, and the importance of analysis.

Sequence based typing of full-length HLA alleles is commonly performed by amplicon sequencing, and that trend is reflected in the projects in this thesis. PCR primers were

designed to cover a specific region that represents a gene. The sequencing of HLA-DRA in **Chapters 6 and 7** includes a larger region than the gene as represented in IPD-IMGT/HLA, and the polymorphism identified in this surrounding region was found to be important in defining distinctions in haplotype patterns. Likewise, **Chapter 8** showed that the intergenic promoter sequence is important in defining haplotype patterns. Moving forward, we will need to expand our represented sequence. Not just by increasing the length of available sequences, but by putting in context of the entire MHC and the human genome.

Attempts to sequence the entire MHC are already in development,[27] and have the potential to generate clear sequences of the entire MHC. Probe-based extraction methods carry the promise of full-length MHC sequences, but it is necessary for the techniques to be as robust as possible. Although region-specific extraction is demonstrated to be possible using PacBio SMRT sequencing, early attempts to use MinION to sequence DNA extracted from an MHC region using this technique were not successful, likely due to poor interactions with nanopores and biotinylated nucleotides. This indicates that haplotype sequencing analysis will continue to require adaptations in the laboratory techniques and sequencing platforms, as well as in bioinformatics interpretation and analysis. But the results will help us to identify and define the flexible haplotypes in the MHC.

Just as the move from full-length gene sequencing to haplotype analysis reveals additional value, considering MHC polymorphism in the context of the whole human genome provides additional value. The MHC does not act alone, and although it is the most complex region, it is important to consider how polymorphism in the MHC relates to the whole human genome. The rise in full genome sequencing means that it is much more likely that an individual may have full-genome sequence data available. Tools such as ALPHLARD[41] work towards the analysis of HLA sequences from data obtained from whole genome sequencing. These attempts hope to unearth HLA polymorphism which can be directly related to genome sequence.

Future studies will continue to identify which genome-wide polymorphism is the most clinically significant. More tests and assays for the typing of the most relevant polymorphism will be developed and used in clinical applications. Sequencing of cDNA and mRNA transcripts will help to clarify how transcription and alternative splicing affect the resulting HLA molecule. Epitope-based matching techniques will increase in the future, and will help to elucidate the behavior of HLA and how it interacts with immune cells. These studies will help to increase our understanding of the expression and biological behavior of HLA.

As scientific projects generate increasing amounts of data, it becomes even more important to share data and ideas. International collaboration will continue to play a critical role in these scientific pursuits. The 2021 International Workshop in Amsterdam and data-standards meetings like DASH will continue to foster a sense of community, where researchers come together and share ideas. It will continue to be important to share the analysis tools we have available. Sharing of software tools on open source websites, and hosting of valuable services will continue to be important for improving the research capabilities of the general community.

The nature of human biology remains a huge mystery, and the function of all of the moving parts within a human body continue to present fascinating areas for research. Humans will continue to have questions about how our genome works, and how it encodes the mind-boggling array of functionality that might seem so simple and trivial in day-to-day life. Every time we seek an answer to one of these questions, we are shedding light on an area of understanding where there was darkness before. This certainly contributes to the greater understanding of how humans work, and hopefully leads to improvements in medical treatments or quality of life. I'm happy that I could contribute this piece of the puzzle, and I know that we'll continue to ask questions and find clever ways to answer those questions in the future.

References

1. KA W. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcostsdata. Accessed Sept 12, 2019.
2. Technologies ON. Clive Brown CTO update | NCM 2019. 2019; https://www.youtube.com/watch?v=fFceCr4O284. Accessed Dec 17, 2019.
3. Payne A, Holmes N, Rakyan V, Loose M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv.* 2018:312256.
4. Payne R, Rolfs MR. Fetomaternal leukocyte incompatibility. *The Journal of clinical investigation.* 1958;37(12):1756-1763.
5. van Rood JJ, van Leeuwen A, Eernisse JG. Leucocyte antibodies in sera of pregnant women. *Vox Sang.* 1959;4:427-444.
6. Robinson J, Guethlein LA, Cereb N, *et al*. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genet.* 2017;13(6):e1006862.
7. Mayor NP, Hayhurst JD, Turner TR, *et al*. Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. *Biology of Blood and Marrow Transplantation.* 2019;25(3):443-450.
8. Marsh SGE, Albert ED, Bodmer WF, *et al*. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens.* 2010;75(4):291-455.
9. Horton R, Wilming L, Rand V, *et al*. Gene map of the extended human MHC. *Nature Reviews Genetics.* 2004;5(12):889-899.
10. Leffler EM, Gao Z, Pfeifer S, *et al*. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science.* 2013;339(6127):1578.
11. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Med.* 2007;4(1):e8.
12. Cullen M, Noble J, Erlich H, *et al*. Characterization of recombination in the HLA class II region. *Am J Hum Genet.* 1997;60(2):397-407.
13. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. High-Resolution Patterns of Meiotic Recombination across the Human Major Histocompatibility Complex. *The American Journal of Human Genetics.* 2002;71(4):759-776.
14. Lam TH, Shen M, Chia JM, Chan SH, Ren EC. Population-specific recombination sites within the human MHC region. *Heredity (Edinb).* 2013;111(2):131-138.
15. Traherne JA, Horton R, Roberts AN, *et al*. Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet.* 2006;2(1):e9.
16. Varney MD, Gavrilidis A, Tait BD. Polymorphism in the regulatory regions of the HLA-DPB1 gene. *Human Immunology.* 1999;60(10):955-961.
17. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. An abundance of bidirectional promoters in the human genome. *Genome Res.* 2004;14(1):62-66.

18. de Klerk E, Venema A, Anvar SY, *et al*. Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. *Nucleic Acids Res.* 2012;40(18):9089-9101.
19. Petersdorf EW, Malkki M, O'hUigin C, *et al*. High HLA-DP Expression and Graft-versus-Host Disease. *New England Journal of Medicine.* 2015;373(7):599-609.
20. Shieh M, Chitnis N, Clark P, Johnson FB, Kamoun M, Monos D. Computational assessment of miRNA binding to low and high expression HLA-DPB1 allelic sequences. *Hum Immunol.* 2019;80(1):53-61.
21. Craenmehr MHC, Haasnoot GW, Drabbels JJM, *et al*. Soluble HLA-G levels in seminal plasma are associated with HLA-G 3'UTR genotypes and haplotypes. *HLA.* 2019;94(4):339-346.
22. Sonon P, Gomes RG, Brelaz-de-Castro MCA, *et al*. Human leukocyte antigen-G 3' untranslated region polymorphism +3142G/C (rs1063320) and haplotypes are associated with manifestations of the American Tegumentary Leishmaniasis in a Northeastern Brazilian population. *Human Immunology.* 2019;80(11):908-916.
23. Thomas R, Apps R, Qi Y, *et al*. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet.* 2009;41(12):1290-1294.
24. Kulkarni S, Savan R, Qi Y, *et al*. Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature.* 2011;472(7344):495-498.
25. Hongming F, Tilanus M, Eggermond Mv, Giphart M. Reduced complexity of RFLP for HLA-DR typing by the use of a DRβ3'cDNA probe. *Tissue Antigens.* 1986;28(3):129-135.
26. Bontrop RE, Broos LAM, Pham K, Bakas RM, Otting N, Jonker M. The chimpanzee major histocompatibility complex class II DR subregion contains an unexpectedly high number of beta-chain genes. *Immunogenetics.* 1990;32(4):272-280.
27. Dapprich J, Ferriola D, Mackiewicz K, *et al*. The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC genomics.* 2016;17:486-486.
28. Degli-Esposti MA, Leaver AL, Christiansen FT, Witt CS, Abraham LJ, Dawkins RL. Ancestral haplotypes: conserved population MHC haplotypes. *Human Immunology.* 1992;34(4):242-252.
29. Dehn J, Spellman S, Hurley CK, *et al*. Selection of unrelated donors and cord blood units for hematopoietic cell transplantation: guidelines from the NMDP/CIBMTR. *Blood.* 2019;134(12):924-934.
30. Gao L, Zhang C, Gao L, *et al*. Favorable outcome of haploidentical hematopoietic stem cell transplantation in Philadelphia chromosome-positive acute lymphoblastic leukemia: a multicenter study in Southwest China. *J Hematol Oncol.* 2015;8:90.
31. Wang Z, Zheng X, Yan H, Li D, Wang H. Good outcome of haploidentical hematopoietic SCT as a salvage therapy in children and adolescents with acquired severe aplastic anemia. *Bone Marrow Transplant.* 2014;49(12):1481-1485.
32. Fabricius WA, Ramanathan M. Review on Haploidentical Hematopoietic Cell Transplantation in Patients with Hematologic Malignancies. *Adv Hematol.* 2016;2016:5726132.
33. Balgansuren G, Regen L, Sprague M, Shelton N, Petersdorf E, Hansen JA. Identification of the rs9277534 HLA-DP expression marker by next generation sequencing for the selection of unrelated donors for hematopoietic cell transplantation. *Human Immunology.* 2019;80(10):828-833.

34. Marsh SGE, Bodmer JG. HLA-DR and -DQ epitopes and monoclonal antibody specificity. *Immunology Today.* 1989;10(9):305-312.
35. Research C-CfIBaMT. CIBMTR Bioinformatics Github. 2019; https://github.com/nmdp-bioinformatics.
36. Milius RP, Heuer M, Valiga D, *et al*. Histoimmunogenetics Markup Language 1.0: Reporting next generation sequencing-based HLA and KIR genotyping. *Human immunology.* 2015;76(12):963-974.
37. Mack SJ, Milius RP, Gifford BD, *et al*. Minimum information for reporting next generation sequence genotyping (MIRING): Guidelines for reporting HLA and KIR genotyping via next generation sequencing. *Human immunology.* 2015;76(12):954-962.
38. Milius RP, Mack SJ, Hollenbach JA, *et al*. Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string. *Tissue Antigens.* 2013;82(2):106-112.
39. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human Immunology.* 2013;74(10):1313-1320.
40. Matern B. Nanopore Prospector Github. https://github.com/transplantation-immunology-maastricht/nanopore-prospector. Accessed Sept 1, 2019.
41. Hayashi S, Yamaguchi R, Mizuno S, *et al*. ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data. *BMC Genomics.* 2018;19(1):790.

Final Summary

Bioinformatics is taking a more prominent role in analysis of HLA and its role in scientific advancement. Every improvement in the tools for molecular analysis, and the techniques that employ them brings about new methods of approaching scientific queries. Improvements in bioinformatic methods enable us to create new scientific queries, which creates a self-perpetuating cycle. This thesis has discussed my use of bioinformatics to make scientific contributions to our understanding of immunogenetics, HLA and the MHC, and how these ideas apply to clinical diagnostics. HLA diagnostics has been evolving, from serological typing, to exon-based molecular genotyping, to identification of full-length gene sequence, to elucidating entire MHC haplotypes. It is critical that our ways of thinking match the evolving technology, in order to fully realize our potential in understanding HLA and immunogenetics.

This thesis describes a two-fold approach to scientific developments. In the first section (**Rules and Tools of HLA Analysis**) I have described techniques that enable molecular analysis, and how we can use these techniques to answer scientific questions. **Chapter 2** describes efforts in collecting and curating sequences in standard databases, which provides more high-quality data to be used in sequencing analysis and diagnostics. Collecting and submitting allele sequences is a difficult and time-consuming project, but the availability of full-length reference sequence is critical to identifying HLA polymorphism, and efforts to simplify this process are of a great value to the community. Community-based collaboration is clearly critical for successful scientific endeavors, which is reflected in **Chapter 3**. Several laboratories contributed samples and sequencing data to a collaborative workshop component, and resulted in a valuable expansion of the IPD-IMGT/HLA database. As our understanding of the MHC advances, the content of IPD-IMGT/HLA will advance as well, and the community will still need to have reference data to understand HLA polymorphism and what it represents.

Targeting and collecting high quality HLA sequences commonly begins with PCR, and one successful technique is described in **Chapter 5.** This assay was analyzed critically, and proven that the resulting amplification produces product that is accurate as well as reliable. We also showed that this product is suitable for use on applicable sequencing platforms, and subsequent use in diagnostics and high-throughput typing. **Chapter 4** described our contribution in bringing a new analysis technique to clinical diagnostics. In a diagnostic setting, it is valuable to have a tool that can quickly and inexpensively type HLA, but there is no room for ambiguities and errors. Applying a new technique (*e.g.* MinION) to a clinical setting brings challenges, and these challenges were met by a focused effort to identify analysis techniques that overcome them. Our techniques were validated, demonstrating that this sequencing platform can be used as a diagnostic tool in a clinical setting.

In the second section (**What's in a Haplotype?**) I have described our contributions to our understanding of MHC haplotypes and immunogenetics. The importance of full-gene polymorphism was discussed in **Chapter 6**, where we indicated how extended polymorphism of HLA-DRA indicates haplotype patterns. Limiting analysis to exons can give an incomplete picture of an HLA gene, and indicates that this gene plays a very limited role. Expanding analysis to intron and UTR sequence gives important insights into the evolutionary history of HLA-DRA, and how it reshapes our understanding of DR~DQ haplotypes. **Chapter 7** expanded on the importance of UTR polymorphism by demonstrating its role in dividing DRB1*13 allele groups and showing that it represents multiple haplotypes. Allele groups are often intended to correlate to serological subtypes, and demonstrating that alleles in this group do not possess characteristic epitopes suggests that another look should be taken at the definition of the DRB1*13 group.

Defining allele groups based on functional distinctions is especially challenging in the HLA-DP region, as we have not yet reached an understanding of what polymorphism is relevant in transplantation. This was explored in **Chapter 8**, where DPA1~Promoter~DPB1 haplotypes were defined and clustered by the intergenic promoter sequences. Correlating the promoter patterns with hypervariable regions demonstrated that clustering the DP haplotypes by this method may be a more effective way to think about this unique HLA region, and may suggest a more functional nomenclature. Allele groups and their correlations with nomenclature and immunogenetics were also a major focus in **Chapter 9**, where we predicted serological specificity of HLA alleles based on amino acid patterns. Efforts to identify serological typing can help to identify permissive mismatches in transplantation, and our efforts to correlate sequence to serology improve the understanding of their relationship, hopefully contributing to general understanding of the nature of HLA.

Together, these projects have created a story about the role of bioinformatics in HLA and immunogenetics. Bioinformatics' multi-faceted and wide reaching approaches makes it perfectly suitable to apply to these fields. I have applied these approaches to identify patterns in HLA sequences, and to observe how these sequences contribute to MHC haplotypes and identifying immunogenic epitopes. Bioinformatics has given me the tools and methodology that I need to approach scientific questions using effective techniques, and it provides new ways to think about problems and develop scientific vision for the future. Bioinformatics techniques will continue to develop, and I am excited to be a part of bringing it into the future.

Valorisation

Rules and Tools of HLA Analysis

Typing and matching of HLA alleles is clearly beneficial in Stem Cell Transplantations, and matching of HLA reduces the effects of Graft-vs-Host disease. Sequencing and matching the full length of HLA at high resolution has also been correlated with improved outcomes, and matching of phased HLA haplotypes improves outcomes even further. High resolution HLA matching is also a strong consideration for Solid Organ Transplantations. The presence of anti-HLA antibodies is the main contraindication for SOT, and high resolution sequencing defines the epitopes that are recognized by the antibodies. Advancements in the platforms and techniques used in HLA sequencing improve the speed and cost-effectiveness of HLA typing, and allow the characterization of full-length HLA polymorphism.

A PCR approach that reliably amplifies 11 HLA loci in four reactions to help in library preparation for sequencing is described in **Chapter 5**. PCR is one of the greatest costs involved in sequence-based typing of HLA alleles, and reduction of these costs enables more HLA laboratories to easily and accurately type these alleles. The availability of a standard primer set allows a more reliable sequencing assay and more consistent analysis and comparison of sequencing data. **Chapter 4** describes the validation of an HLA typing approach which uses nanopore sequencing. MinION sequencing is portable, requires less up-front costs, and requires only minimal laboratory equipment. It generates full-length single-molecule reads, which allows phasing of relatively distant polymorphism and reduces the inherent difficulties of phasing cis/trans polymorphism in heterozygous sequencing by short-read technology. The smaller form factor and relatively short time required for sequencing makes it an attractive target as an on-call typing device. The benefits of MinION are however balanced by challenges in implementation. Basecalling models can struggle with regions of low sequence variation, especially homopolymer sequences, and bioinformatics approaches are necessary to correctly interpret the data. Interpretations of sequencing data, especially from novel platforms, must be validated for accuracy and reliability. MinION sequencing, combined with a validated analysis technique, enables a wider variety of laboratories to sequence and type HLA, to the benefit of the HLA and transplantation communities.

The HLA genes are hyperpolymorphic, which is apparent in the number of unique allele sequences in IPD-IMGT/HLA. The sequence data in this repository is freely available, and is often used in commercial software packages for HLA analysis. The availability of a standard HLA database with official names from the WHO nomenclature committee is of great value to the community. It allows standardization and unambiguous typing and comparison of HLA alleles which can be communicated between any HLA laboratory. Many of the alleles

have only partial sequences available; just 27.3% of the available HLA alleles have full-length (5′ UTR to 3′ UTR) sequences available (release 3.39.0), a significant improvement over the <8% reported by Dr. Steven J. Mack in 2015. This improvement is thanks to local and international efforts to fill the gaps in available full-length sequences. **Chapter 3** describes the results of an international collaboration at the 17th HLA workshop where 34 HLA alleles were extended with complete full-length sequences. Matching of full-length HLA sequences allows matching of a greater amount of polymorphism, compared with matching only the antigen presentation domains, and the availability of full-length sequences allows more specific studies that compare polymorphism between groups.

Sequencing of an individual's HLA genes, especially individuals from under-represented populations, regularly produces novel allele sequences. The submission and naming of these sequences in IPD-IMGT/HLA provides a continuous increase in the known HLA polymorphism, to the benefit of HLA researchers and transplantation clinicians. However, the submission process can be cumbersome, and highly-curated databases often have higher requirements for submission. Gathering the necessary metadata and documentation requires some human effort, **Chapter 2** describes an effort to ease that process. Saddlebags is a freely available tool designed to simplify the process of submission to EMBL/ENA, an important step in submission to IPD-IMGT/HLA, allowing laboratories to more easily participate in submission of novel HLA alleles. Saddlebags has been used by laboratories around the world for submission of HLA class I sequences, and development is continuing to support HLA class II and bulk sequence submission.

What's in a Haplotype?

The HLA genes do not exist in isolation, they are part of a complex and variable MHC region. The second part of this thesis is entitled "What's in a Haplotype?" which reflects a major theme of this thesis. Outside of the HLA field, a haplotype may represent only two linked SNPs, but for HLA researchers a haplotype represents polymorphism in multiple genes, and possibly all polymorphism across an entire chromosome. Regardless, haplotypes are a critical concept in HLA studies. Determining haplotype patterns is an important step in identifying patterns in linkage disequilibrium between SNPs within a gene, or polymorphism at completely different loci. Haplotype studies allow researchers to identify polymorphism that is conserved through evolution, or polymorphism that is commonly inherited together. We can find the relationship between polymorphism of alleles at two loci that encode a protein heterodimer, and clarify how it affects the behavior of the resulting molecule. Haplotypes help us to find new patterns in the organization of genes, and sheds light on the nature of the MHC.

In addition to applications in answering research questions, haplotypes have an important role in transplantation. Matching of phased HLA haplotypes in addition to the unphased

genotypes provides further benefits in stem cell transplantations, perhaps due to implicit matching of unsequenced polymorphism. Haplotypes provide an advantageous effect in the context of haploidentical transplants, that seems to overcome the effects of mismatched HLA alleles. Sequencing haplotypes may help to clarify the linkage disequilibrium patterns and poorly understood mechanisms that provide these beneficial effects. It is clear that identifying patterns in haplotypes increases our understanding of the mechanisms within the MHC, to the benefit of both scientific and clinical applications.

This thesis has expanded our understanding of HLA haplotypes, especially in the class II region. In **Chapter 6**, we explored the role of HLA-DRA polymorphism in DR~DQ haplotypes. Previous literature has described DRA as monomorphic, with a consistent locus within well-defined haplotype patterns. The exon sequences were found to have minimal polymorphism compared to other HLA genes, but we described 20 novel SNPs in the introns and UTR sequences. Haplotype analysis revealed that patterns of polymorphism are correlated with specific HLA-DRB and DQB1 alleles, suggesting that although the DR alpha subunit is evolutionarily conserved, the non-coding polymorphism of HLA-DRA suggests distinct evolutionary lineages and plays an important role in defining DR-DQ haplotypes.

Although previous studies have categorized haplotypes into one of just a few patterns, **Chapter 7** expands our understanding of HLA-DRB1*13 haplotypes and explores the theory of a flexible MHC. We have suggested that the MHC is a flexible and dynamic system which is subject to continued evolution, and that existing haplotypes may not always fall within the definitions of known patterns. This model is presented, not as a conclusive and final definition, but as an idea that can be expanded in further studies by others in the community. As more individuals are sequenced, and more research projects to determine haplotype patterns are carried out by researchers worldwide, the community will further understand how the HLA and non-HLA genes fit together, and how evolutionary pressures affect differentiation between individuals and ethnic groups.

Our understanding of haplotypes was further extended in our studies of the HLA-DP region (**Chapter 8**). DPA1 and DPB1 have an interesting head-to-head orientation with a shared overlapping promoter region. Unlike other HLA loci, HLA-DP nomenclature is not based on allele groups defined by serology. Sequencing the entire region identified common promoter patterns, and haplotype analysis indicates that sequence clusters based on these patterns form strong correlations with the hypervariable regions in DPB1. This suggests a relationship between the promoter region, which likely affects HLA-DP expression levels, and the hypervariable regions in the antigen presentation domain, which affect the HLA-DP immunogenicity. The allele clusters defined by promoter sequences were defined with the goal that future studies and collaborations, such as the

International HLA & Immunogenetics Workshop, can expand on the patterns and clarify their clinical consequences.

The relationship between polymorphism and immunogenicity was further explored in **Chapter 9**. The use of serological HLA typing in a clinical setting is generally decreasing, and the serological typing for many allele sequences is unknown. Serological subtypes of specific HLA-B alleles are not known, and can be difficult to assign due to scarcity of available sera. The serotyping is critical in determining if patient donor-specific antibodies (DSAs) are specific to the transplanted tissue, and models have been proposed to predict serology based on sequence polymorphism. We have proposed one technique for using patterns in specific amino acid polymorphism, compared with alleles with known serotypes, to predict the potential serological subtype of an unknown HLA-B*15 sequence. This method is proposed as an alternative model to existing models that use machine learning-based serology prediction, and it's accuracy and efficacy are free to explore by the community.

The HLA Community

For many of the projects in this thesis, specific software tools were developed for analysis, and software that we created for analysis of HLA sequences is provided as open-source software whenever possible. This includes the code for Saddlebags, as well as Nanopore Prospector, the collection of code and scripts that has provided some capability to analyze MinION reads and HLA allele sequences. The code is available on Github, a widely-used repository for open-source software, and is provided under the GNU GPL 3.0 license, which means that it can be freely downloaded and modified and repurposed for different applications. Providing open-source software has remained a high priority during these studies, since it increases the clarity of how the analysis was performed, and benefits the community by helping other researchers to formulate techniques for analyzing sequencing data or HLA alleles.

The International HLA & Immunogenetics Workshop is a worldwide gathering of researchers and clinicians who work to standardize methodologies, definitions, nomenclatures, and concepts and collaborate on community-focused well-defined projects related to HLA and immunogenetics. The workshop occurs once every 2-5 years, and workshop projects have been a recurring theme in several chapters in this thesis. **Chapter 3** is the direct result of a 17th workshop project where labs collaborated to sequence and submit (**Chapter 2**) full-length HLA allele sequences. This project will be expanded and continued at the 18th workshop. The 18th workshop also features a project focused on DPA1-promoter-DPB1 haplotypes, which will expand on the results identified in our HLA-DP project (**Chapter 8**). We explored the ideas of polymorphic epitopes in **Chapters 7 & 9**, which are related to the planned projects of identifying immunogenetic epitopes and an update of the

HLA dictionary. The 18th workshop will also feature projects focused on bioinformatics, including analysis of recombinations in inherited haplotypes, population genetics, and a community-focused DASH data standards hackathon, all of which relate to projects in this thesis.

All studies in this thesis have been performed with a goal of improving our understanding of HLA for the benefit of patients, clinicians, and researchers in the HLA community. We have put priority on sharing of our results and data whenever feasible, and on active participation in the collaborative congresses and hackathons. This thesis has been focused on the creation and use of software tools which clarify our knowledge of the MHC and which can be applied to many HLA research questions. The projects represented by this thesis are a snapshot in a continuing timeline; it expands on the discoveries of earlier HLA researchers, and the results have the goal of extending the capabilities of future researchers to continue to advance the field of HLA and immunogenetics.

List of Publications

Matern BM, Olieslagers TI, Voorter CEM, Groeneweg M, Tilanus MGJ: Insights into the polymorphism in HLA-DRA and its evolutionary relationship with HLA haplotypes. HLA 2020 Feb;95(2):117-127. doi: 10.1111/tan.13730.

Truong L, Matern BM, D'Orsogna L, Martinez P, Tilanus MGJ, De Santis D: A novel multiplexed 11 locus HLA full gene amplification assay using next generation sequencing. HLA 2020 Feb;95(2):104-116. doi: 10.1111/tan.13729.

Voorter CEM, Matern BM, Tran TH, Fink A, Vidan-Jeras B, Montanic S et al.: Full-length extension of HLA allele sequences by HLA allele-specific hemizygous Sanger sequencing (SSBT). Human Immunology 2018 Nov;79(11):763-772. doi: 10.1016/j.humimm.2018.08.004.

Matern BM, Groeneweg M, Voorter CEM, Tilanus MGJ: Saddlebags: A software interface for submitting full-length HLA allele sequences to the EMBL-ENA nucleotide database. HLA 2018 Jan;91(1):29-35. doi: 10.1111/tan.13179.

Duygu B, Matern BM, Groeneweg M, Voorter CEM, Tilanus MGJ: Polymorphism at residue 156 of the new HLA-A*02:683 allele suggests immunological relevance. HLA 2017 Aug;90(2):107-109. doi: 10.1111/tan.13059.

Matern BM, Olieslagers TI, Groeneweg M, Tilanus MGJ: Division of HLA-DRB1*13 haplotypes by extended HLA-DRA 3'UTR polymorphism refines HLA-DRB1*13~HLA-DRB3~HLA-DQB1 haplotypes and gives clues to HLA-DR13 immunogenicity. In Preparation

Duygu B, Matern BM, Wieten L, Voorter CEM, Tilanus MGJ: Specific amino acid patterns define split specificities of HLA-B15 antigens enabling conversion from DNA based typing to serological equivalents. Submitted (HLA)

Truong L, Matern BM, Groeneweg M, D'Orsogna L, Martinez P, Tilanus MGJ, De Santis D: Polymorphism clustering of the 21.5kb DPA-Promoter-DPB region reveals novel extended full length haplotypes. Submitted (HLA)

Matern BM, Olieslagers TI, Groeneweg M, Duygu B, Wieten L, Tilanus MGJ, Voorter CEM: Long-read nanopore sequencing validated for HLA typing in routine diagnostics. Submitted (Journal of Molecular Diagnostics)

Curriculum Vitae

Ben was born April 28, 1986 in Minneapolis, Minnesota. Ben is the third of six children of Mark and Martha Matern. He gained an early interest in math, science, and computers during his early education, and graduated from Spring Lake Park High School in 2004. Ben completed an undergraduate degree in Applied Mathematics and Computer Science, with a concentration in Bioinformatics, from the University of Wisconsin: Stout in 2009. This interest in bioinformatics was explored further at the University of Minnesota, where he completed a Master's degree in Biomedical Informatics and Computational Biology in 2015. During his Master's studies he collaborated with Martin Maiers and Bob Milius of the National Marrow Donor Program bioinformatics group on a MIRING validator, a project related to data standards and transmission of high quality HLA genotyping data. He began participating in the Data Standards Hackathons (DASH), and formed a deep interest in HLA and DNA sequencing. It was through NMDP that he made contact with Mathijs Groeneweg, Lotte Wieten, and Marcel Tilanus of the Transplantation Immunology Laboratory of Maastricht University Medical Center, and decided to pursue a PhD while working on MinION sequencing full length HLA and exploring the nature of the identified polymorphism. Ben plans to continue and expand his research at the University Medical Center Utrecht, and will continue to participate and contribute to the International Histocompatibility Workshops and the HLA community.

Acknowledgements

Well, it seems that I'm nearing the end of my PhD training here in Maastricht. The experience of moving across the world to do science has been as rewarding as it has been challenging. And certainly life-changing. I can't imagine where I'd be if I hadn't decided to take that plunge, but when I see where I've ended up, it seems like it was certainly the right decision. This experience has been so valuable for me, and I am grateful to everyone who has supported me along the way.

When choosing which propositions to include in my thesis, I wanted to be sure to include one about all the support and collaboration that has been a theme throughout this PhD, which I think is summarized well as "We are greater than the sum of our parts." Although it may not be the most useful prompt for scientific discussion, it certainly rings true. The expertise, hospitality, and collaboration of others makes the whole experience more valuable. I was fortunate to gain many new friends and colleagues during the last four years, and I learned so much from the more experienced advisors and colleagues.

I start with **Marcel**. You draw upon your years of experience to help provide scientific vision and leadership in all of these projects. I'm grateful to have the direct and firm leadership, and for how you have helped me to define goals and scientific vision. You've shown me focus and drive and determination, as well as flexibility, which is inspiring. Sincere thanks for all of the opportunities, and for helping me meet the right people. I was not your first PhD student, and I hope you can continue your advisory roles into the future.

Next up is my co-promoter **Mathijs.** You've been stuck in a server closet with me for 4 years, and we've both made it out alive. Since the beginning, when you took me in your van to find furniture for my apartment, you've been accommodating and friendly. Thanks for all of the enlightening discussion, it seems like you can come up with a good answer to any random question. You've shown me how to translate my bioinformatic ramblings into something that makes sense to laboratory people. Thanks for all the nerdy movies, and I'm sure I'll be running into you at all the conferences.

Beyond my promoters, I've also been fortunate to work with some top-notch colleagues. Thanks **Lotte** for your excellent leadership. Your input at the scientific discussions was very valuable. You always were clear with expectations and boundaries, while considering what I need and providing me with all of the valuable opportunities over the last years. I'm so thankful to have worked with you. I'm also thankful for the experience of working with **Christien.** You were always prepared to do valuable detailed review of data, papers, and abstracts. The frank feedback and suggestions were extremely valuable to making my writing the highest quality I can make.

Thanks especially to the members of my thesis assessment committee, **Drs**. **Savelkoul**, **Marsh**, **Evelo**, **zur Hausen**, and **Spierings**. It takes a great deal of time to assess a thesis, and I appreciate the effort. I could not have done it without you, so thanks.

I have made many memories with my two paranymphs, **Timo** and **Burcu**. You two have both been a constant joy and encouragement during this journey. I daresay all three of us have benefited greatly from our unique perspectives, and I'm glad we've even found time for the occasional socializing as well. Our adventures in EFI will be fond memories for a long time. **Timo**, thanks for reminding me when it's lunchtime, and for just having that personality, you make the lab a joy to be in. And **Burcu**, thanks for always being approachable, and for giving a sense of sanity, and for showing me that this whole PhD thing can actually be done.

The Maastricht lab is fortunate to have a whole team of people who are a joy to work with. **Robbert**, you asked me about guitar in our first conversation. And I think you asked me about beer pretty soon after that. Thanks for being a friend, and welcoming me when I was in a strange place far from home. **Tom**, nobody knows board games like you do, thanks for the discussions as well as your help on my projects. **Laura**, thanks for being patient as I was testing the waters of an advisory role. Even when I didn't define tasks perfectly, you were willing to step up and always got things done. I still refer to your thesis sometimes, and I'm taking it with me when I move. Nice job. **Dorien**, **Lize**, **Esther**, **Simone**, and **Jeroen,** whether going to an escape room or Phantasialand or just eating fries. Thanks for the friendship, you made me feel welcome in Maastricht. **Christel,** I always loved working with you. You always explain things well and are just a joy to be around, thanks. We'll be in touch :) Thanks to everyone for the help and hospitality, **Stefan, Filiz, Fausto, Dominique, Jacqueline, Annette, Veerle, Lisette, Coline, Ilse, Sophie, Marjolijn, Carmen, Maud, Levi, and Bert.** Oh and **Thuur**, as far as I'm concerned you're still one of the Original Gangsters of MinION. **Sandra**, **Audrey**, **Brigitte**, and **Diana,** thanks for all the assistance with organizational items. You were always so helpful in sorting out the details, and with practicing my Dutch.

One of my best PhD role models is **Niken**. Seeing your success has been an inspiration, and I've been using your thesis as a model for mine, so I hope you did it right ;) Thanks to **Femke**. You started your PhD pretty soon after I did. It's clear that you're well on your way. You got this! **Denise,** Nobody can bring a smile to the room like you can. Even just "Hi Ben!" brightens up the day. Your dedication to the OVIDE project and to helping people is inspiring.

This PhD journey could not have taken place without the support and advice of a lot of people. **Martin,** basically the best boss anyone could hope for. Even when I suggested leaving your group to pursue this crazy Dutch adventure, you gave me the support I needed. You've given great advice, and checked in on my progress from time to time. **Bob,** your support during my Masters was crucial, and I'm inspired by how you're making everyone realize how cool data standards are. Sincere thanks to you and everyone else I've worked with at BeTheMatch, especially **Mike W, Mike H, Jane**, **Caleb**, **Wei**, **Julia**, **Hu**, **Jason, Joel, Abeer,** and **Loren**.

Thanks especially to **Linh.** You and I have been helping each other figure out this research thing. We met as two world travelers, displaced to do some science, and I'm glad to have your friendship. Your expertise in analysis was critical in our projects, and even when I can't keep up with you I think we're still learning from each other. **Dianne,** You have been a wonderful host, as well as a great motivator and scientific advisor. Thanks for all the opportunities and collaboration, and hopefully I'll get to visit again sometime soon. The Perth team, thanks for the warm welcome and friendship. You made my home away from home away from home a joy to visit, and helped keep me busy with scientific discussions and socialization when I was there.

Mike C, I'm working on collecting my titles, and I'm looking forward to being part of your celebration in May. Thanks for always being there. The rest of my friends back home, especially **Dave**, **Mike H**, **Kevin**, **Matt**. I miss you guys, time to meet up, either in Minnesota or Amsterdam.

Fenna, I'm so pleased with this cover design. Thanks for your patience with the preparation of the thesis, it's so valuable to me.

Yvonne, You have made the last year so much better just by being you. You listen and understand when I'm complaining, and we share the joy when things are great. I have loved our random trips to Antwerp or Amsterdam or Minnesota, and I'm so glad we can share interests (gotta catch em all) and make me comfortable to be myself. Thanks for always being there. The love and support and patience we have can't be measured. I'm excited to continue this adventure in Utrecht, and I look forward to seeing where the next years take us.

Mom and Dad, I don't even know where to start. Thanks for everything. You've both always encouraged me to do what makes me happy, and given me the space and flexibility to figure out what that is. Even when I stumble, you've just asked what I need and how you can help. Every time I hop on the plane and say goodbye for a few months I feel crushed, but even when you're across the ocean you're providing encouragement and inspiration.

Mom, when you donated your kidney, it scared everyone in our family. But with your strength you made it through, and thanks to you, my uncle **Adam** is still posting insightful memes on facebook. I always quote that as part of my inspiration for pursuing this field. **Dad,** your dedication to the family and all of the hard work you have done for everyone is legendary. I can't possibly have had a better Dad. And for better or worse, I have to thank you for all my personality traits, I think I inherited the chromosome for thinking before speaking, kindness, and helpfulness from you. Love you both.

And why do I have so many siblings? **Greg,** you're my older brother. My role model, and my original inspiration for becoming quite the geek. So thanks for that I guess. **Stacey,** you really bring a different flavor of humor to our family. Thanks for bringing the Canadian joy, it's authentic and smells like maple syrup. **Carrie,** my big sister. We grew up together, and even attended school together. Seeing your successes and joyous family makes me so proud to be your brother. **Jason,** when you came into the picture, us brothers weren't so sure. But when we saw that you're kind, funny, tall, and good at games, you became one of us very quickly. **Dominic,** I think of you as just a slightly younger version of myself, I think our personalities overlap to the point of confusing others. Thanks for the lan parties and fun when I'm at home. **Bridget,** seeing you turn into the responsible and caring adult you are inspires joy. It was just yesterday that you were a child, and I'm so excited to see where life brings you. **Peter,** the baby. You've seen the rest of us become adults, and some people, I suppose, might say you're becoming one yourself. You've taken up the torch as the best gamer in our family, a huge responsibility. Keep up the good work bro, I'm proud of you.

And all of the niblings, **Miles**, **Odin**, **Erin**, **Will**, **Ike**, and **Logan**. One of the biggest drawbacks to being in Europe is missing you guys growing up every week. Seeing how your personalities are forming is a joy that can't be put into words. I like playing the part of the scientist uncle, because I hope to show you some of the things that are possible. Whatever you do, we'll all be excited to watch you grow up. PS I'm still collecting every piece of art you've been giving me, I absolutely love it. I wish I could mention everyone by name, especially my grandparents and aunts and uncles and cousins, but this thesis is long enough already. Thanks to my whole family, I love you all.

It's been a real challenge, but it's been a real joy. Thanks again to everyone, and I look forward to seeing where life will take us next.