We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

4,100
Open access books available

116,000
International authors and editors

120M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Chapter 3

MHC Genotyping in Human and Nonhuman Species by PCR-based Next-Generation Sequencing

Takashi Shiina, Shingo Suzuki and Jerzy K. Kulski

Abstract

The major histocompatibility complex (MHC) is a highly polymorphic genomic region that encodes the transplantation and immune regulatory molecules. It receives special attention for genetic investigation because of its important role in the regulation of innate and adaptive immune responses and its strong association with numerous infectious and/or autoimmune diseases. Recently, genotyping of the polymorphisms of MHC genes using targeted next-generation sequencing (NGS) technologies was developed for humans and some nonhuman species. Most species have numerous highly homologous MHC loci so the NGS technologies are likely to replace traditional genotyping methods in the near future for the investigation of human and animal MHC genes in evolutionary biology, ecology, population genetics, and disease and transplantation studies. In this chapter, we provide a short review of the use of targeted NGS for MHC genotyping in humans and nonhuman species, particularly for the class I and class II regions of the Crab-eating Macaque MHC (Mafa).

Keywords: HLA, MHC, polymorphism, genotyping, NGS

1. Introduction

The major histocompatibility complex (MHC) genomic region consists of a large group of evolutionary-related genes involved functionally with the innate and adaptive immune systems in jawed vertebrates [1]. In humans, the MHC is located on the short arm of chromosome 6, band p21.3, and the MHC class I and class II genomic regions encode the highly polymorphic gene complex classified as the human leukocyte antigen (HLA) complex [2, 3]. The HLA class I and class II molecules expressed by the MHC play important roles in restricted
cellular interactions and tissue histocompatibility due to cellular discrimination of “self” and “nonself” that require an essential knowledge of the effects of HLA allele matched and mismatched donors in transplantation medicine [4] and transfusion therapy [5]. While the HLA class I molecules are expressed by all nucleated cells to present processed peptides of intracellular origin to CD8+ cytotoxic T cells and serve as ligands for natural killer cells, the class II molecules are expressed by antigen-presenting cells such as B cells, dendritic cells, or macrophages to present exogenous peptides to CD4+ helper T cells of the immune system [6]. In addition, the classical HLA class I genes, HLA-A, HLA-B, and HLA-C, and the classical HLA class II genes, HLA-DR, HLA-DQ, and HLA-DP are distinguished by their extraordinary polymorphisms, whereas the nonclassical HLA class I genes, HLA-E, HLA-F, and HLA-G, are distinguished by their tissue-specific expression and limited polymorphism [2, 3, 7].

The highly polymorphic HLA genomic region is critically involved in the rejection and graft-versus-host disease (GVHD) of hematopoietic stem cell transplants [8, 9], the pathogenesis of numerous autoimmune diseases [10–13], and infectious diseases [14]. Apart from regulating immunity, the MHC genes may have a role in reproduction and social behavior, such as pregnancy maintenance, mate selection, and kin recognition [15]. The MHC genomic region also appears to influence drug adverse reactions [16, 17], CNS development and plasticity [18–22], neurological cell interactions [23, 24], synaptic function and behavior [25, 26], cerebral hemispheric specialization [27], and neurological and psychiatric disorders [28–32]. Hence, the MHC is one of the most biomedically important genomic regions that warrant special attention for genetic investigation.

In general, the study of the diversity and polymorphic variation of the MHC genomic region has been focused more on humans than any other species and animal population [1] largely because of the high cost and limited throughput of the first generation Sanger sequencing method [33, 34]. However, this is now changing because the next-generation sequencing (NGS) technologies are becoming the method of choice for lower-cost, high-throughput genotyping of MHC genes that are composed of highly homologous multiple loci such as those found in the macaque primate species [35]. Thus, the NGS technologies are expected to perform precise MHC genotyping in human and model animals that already have a collection of MHC allele references, and to facilitate MHC genotyping of wild animals that as yet have no MHC allele references. In addition, the NGS technologies are likely to replace traditional genotyping methods such as subcloning, Sanger sequencing, and previously developed PCR-based MHC typing methods (PCR-RFLP, PCR-SSP, and so on) in the near future. Recently, many articles concerning the development of NGS technologies for precise MHC genotyping and genotyping data of MHC genes using the new NGS technologies have been published on the investigations of human and nonhuman MHC polymorphisms in various fields of study such as medical science, evolutionary biology, ecology, and population genetics.

In this chapter, we provide a short review of the current HLA polymorphism information and the use of PCR-based NGS for MHC genotyping in human and nonhuman species, particularly for the Filipino crab-eating macaque MHC (Mafa) class I (Mafa-A, -B, -E, -F, and -I) and class II loci (Mafa-DPA1, -DPB1, -DQA1, -DQB1, and -DRB1).
2. HLA allele number

A total of 13,840 HLA allele sequences, 10,297 in the class I and 3543 in the class II gene regions, were released by the IMGT/HLA database [7] release 3.22 in October 2015 (Table 1).

<table>
<thead>
<tr>
<th>Category</th>
<th>Locus</th>
<th>Allele no.</th>
<th>Protein no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>HLA-A</td>
<td>3285</td>
<td>2313</td>
</tr>
<tr>
<td></td>
<td>HLA-B</td>
<td>4077</td>
<td>3011</td>
</tr>
<tr>
<td></td>
<td>HLA-C</td>
<td>2801</td>
<td>1985</td>
</tr>
<tr>
<td></td>
<td>HLA-E</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>HLA-F</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>HLA-G</td>
<td>51</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Pseudogene</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>10,297</td>
<td>7337</td>
</tr>
<tr>
<td>Class II</td>
<td>HLA-DRA</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>HLA-DRB1</td>
<td>1825</td>
<td>1335</td>
</tr>
<tr>
<td></td>
<td>HLA-DRB3</td>
<td>60</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>HLA-DRB4</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>HLA-DRB5</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>HLA-DQA1</td>
<td>54</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>HLA-DQB1</td>
<td>876</td>
<td>595</td>
</tr>
<tr>
<td></td>
<td>HLA-DPA1</td>
<td>42</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>HLA-DPB1</td>
<td>587</td>
<td>480</td>
</tr>
<tr>
<td></td>
<td>HLA-DMA</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>HLA-DMB</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>HLA-DOA</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>HLA-DOB</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Pseudogene</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>3543</td>
<td>2561</td>
</tr>
</tbody>
</table>

Table 1. Number and genomic distribution (loci) of HLA alleles

The IMGT/HLA database is a specialist database for HLA sequences. Ten years ago, the allelenumbers were only 2182, but since then, the numbers have increased by 1000 allele sequences each year. Of 10,297 HLA class I alleles, 3285, 4077, 2801, 18, 22, and 51 alleles were counted in HLA-A, HLA-B, HLA-C, HLA-E, HLA-F, and HLA-G genes, respectively (Table 1);
10,163 and 91 alleles were counted in the classical and nonclassical HLA class I genes, respectively. Of the 3,543 HLA class II alleles, 7, 18, 25, 99, 54, 87, 62, 7, 13, 12, and 13 alleles were counted in HLA-DRA, HLA-DRB1, HLA-DRB3/4/5, HLA-DQA1, HLA-DQB1, HLA-DPA1, HLA-DPB1, HLA-DMA, HLA-DMB, HLA-DOA, and HLA-DOB genes, respectively (Table 1), with 3,490 and 45 alleles in the classical and nonclassical HLA class II genes, respectively.

3. History of HLA genotyping methods

Many variations of the conventional HLA genotyping methods such as incorporating restriction fragment polymorphisms (RFLP) [36], single strand conformation polymorphism (SSCP) [37], sequence-specific oligonucleotides (SSOs) [38], sequence-specific primers (SSPs) [39], and sequence-based typing (SBT), like the Sanger method [33], have been used for the efficient and rapid HLA matching in transplantation therapy [40–43], research into HLA-related diseases [2, 3], population diversity studies [44–46], and in forensic and paternity testing [47]. The HLA genotyping methods mainly applied today are PCR-SSOP, such as the Luminex commercial methodology [48, 49], and SBT by the Sanger method employing capillary sequencing based on chain-termination reactions [33, 34]. However, both methods often detect more than one pair of unresolved HLA alleles because of chromosomal phase (cis/trans) ambiguity [50, 51]. To solve the phase ambiguity problem, new HLA genotyping technologies have been reported and commercialized that combine the PCR amplification of targeted HLA genomic regions with NGS platforms such as the ion PGM system (Life Technologies), GS Junior system (Roche), and the MiSeq system (Illumina) [52]. The PCR/NGS methods are expected to produce genotyping results that detect new and null alleles efficiently without phase ambiguity.

4. Summary of NGS-based human MHC genotyping methods

Table 2 shows a list of publications on NGS-based human MHC genotyping that includes information for PCR range, targeted HLA locus, NGS platform, and allele assignment method. The MHC genotyping methods in human are basically composed of three steps, PCR, NGS, and allele assignment. We summarize the important points in each of the three steps below. The more detailed information is described in our previous publication [52].

4.1. PCR step

4.1.1. Long- and short-range PCR

PCR methods produce amplicons of different sequence lengths depending on the primer design and the type of DNA polymerase used for the PCR. The amplicon sizes are usually classified into two size ranges: the short-range system where the amplicon size is <1 kb and the long-range system where the amplicon size is >1 kb as shown in Figure 1.
The short-range PCR system is a method based on PCR amplification of each exon that includes polymorphic exons 2 and 3 in HLA-A, HLA-B, and HLA-C and exon 2 in HLA-DR, HLA-DQ, and HLA-DP. One of the advantages of the short-range system is that it is the most suitable for application of physically fragmented DNA samples as templates such as those extracted from swabs because the PCR length is relatively short, ranging from 250 bp to 900 bp, per

<table>
<thead>
<tr>
<th>No.</th>
<th>PCR range</th>
<th>Sorting from PCR range</th>
<th>Targeted HLA locus</th>
<th>NGS platform</th>
<th>Allele assignment method</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>410–790 bp</td>
<td>Short-range system</td>
<td>A, B, C, DRB1/3/4/5, DQA1, DQB1, DPB1</td>
<td>454 GS FLX</td>
<td>Conexio Assign ATF</td>
<td>[66]</td>
</tr>
<tr>
<td>2</td>
<td>400–900 bp</td>
<td>Short-range system</td>
<td>A, B</td>
<td>454 GS FLX</td>
<td>GS-FLX amplicon variant analyzer</td>
<td>[67]</td>
</tr>
<tr>
<td>3</td>
<td>Unknown</td>
<td>Long-range system</td>
<td>A, B, C, DRB1, DQB1</td>
<td>454 GS FLX</td>
<td>Conexio Assign-NG</td>
<td>[51]</td>
</tr>
<tr>
<td>4</td>
<td>410–790 bp</td>
<td>Short-range system</td>
<td>A, B, C, DRB1/3/4/5, DQA1, DQB1, DPB1</td>
<td>454 GS FLX</td>
<td>Conexio Genomics ATF</td>
<td>[68]</td>
</tr>
<tr>
<td>5</td>
<td>381–537 bp</td>
<td>Short-range system</td>
<td>A, B</td>
<td>454 GS FLX</td>
<td>SSAHA2</td>
<td>[69]</td>
</tr>
<tr>
<td>6</td>
<td>4.6–11.2 kb</td>
<td>Long-range system</td>
<td>A, B, C, DRB1, DQA1, DQB1, DPB1, DPA1, DPB1</td>
<td>454 GS Junior, Ion PGM</td>
<td>SeaBass</td>
<td>[70]</td>
</tr>
<tr>
<td>7</td>
<td>2.7–4.1 kb</td>
<td>Long-range system</td>
<td>A, B, C, DRB1</td>
<td>HiSeq2000, MiSeq</td>
<td>Alignment with IMGT/HLA data</td>
<td>[71]</td>
</tr>
<tr>
<td>8</td>
<td>410–790 bp</td>
<td>Short-range system</td>
<td>A, B, C, DRB1/3/4/5, DQA1, DQB1, DPB1</td>
<td>454 GS FLX, (or GS Junior)</td>
<td>Conexio Assign ATF 454</td>
<td>[72]</td>
</tr>
<tr>
<td>9</td>
<td>3.4–13.6 kb</td>
<td>Long-range system</td>
<td>A, B, C, DRB1, DPB1, DQB1</td>
<td>MiSeq</td>
<td>BWA, Samtools, GATK, PerlScript</td>
<td>[73]</td>
</tr>
<tr>
<td>10</td>
<td>5.1–5.6 kb</td>
<td>Long-range system</td>
<td>DRB3, DRB4, DRB5</td>
<td>454 GS Junior</td>
<td>SeaBass</td>
<td>[74]</td>
</tr>
<tr>
<td>11</td>
<td>250–270 bp</td>
<td>Short-range system</td>
<td>A, B, C, DRB1, DPB1, DQB1</td>
<td>MiSeq</td>
<td>neXtype</td>
<td>[75]</td>
</tr>
<tr>
<td>12</td>
<td>250–270 bp</td>
<td>Short-range system</td>
<td>DRB1/3/4/5, DQA1, DQB1, DPA1, DPB1</td>
<td>MiSeq</td>
<td>Genetics Management System</td>
<td>[76]</td>
</tr>
<tr>
<td>13</td>
<td>Unknown</td>
<td>Long-range system</td>
<td>A, B, DRB1</td>
<td>PacBio</td>
<td>Bayes' theorem, NGSeqengine</td>
<td>[77]</td>
</tr>
<tr>
<td>14</td>
<td>4.0–7.2 kb</td>
<td>Long-range system</td>
<td>A, B, C, DRB1/3/4/5, DQB1, DPB1</td>
<td>Ion PGM</td>
<td>SeaBass</td>
<td>[54]</td>
</tr>
</tbody>
</table>

Table 2. Publication list of NGS-based MHC genotyping in human. Bold letter shows publications from the author’s group
amplicon. On the other hand, the short-range system is less effective for genotyping recombinant alleles that have been generated by recombination events of the HLA genes because it is difficult to avoid the phase ambiguities generated by recombinations. For example, in Figure 2, B*15:20 has an identical nucleotide sequence with B*15:01 in exon 2 and B*35:01 in exon 3, but B*35:43 has an identical nucleotide sequence with B*35:01 in exon 2 and B*15:01 in exon 3. When we genotype a DNA sample that has B*15:01 or B*15:20 and B*35:01 or B*35:43, ambiguous genotyping can result in assignments such as B*15:01/20 and B*35:01/43 that are difficult to assign correctly and definitively.

The long-range PCR system is a method based on PCR amplification of the entire HLA gene region including the promotor-enhancer region, 5′ untranslated region (UTR), all exons, all introns, and the 3′ UTR or partial gene regions that include polymorphic exons and adjacent introns (Figure 1). Primer sets for long-range systems have already been developed and published for HLA-A, HLA-B, HLA-C, HLA-DRB1/3/4/5, HLA-DQA1, HLA-DQB1, HLA-DPA1, and HLA-DPB1 (Table 2). The advantage of long-range PCR is that this system can easily solve phase ambiguity even if recombinant alleles such as those shown in Figure 2 are present in DNA samples. Also, the long-range PCR method is expected to detect new polymorphisms and variations throughout the entire HLA gene region. Therefore, the long-range

---

Figure 1. Outline of NGS-based human MHC typing.
system is an important and useful alternative to the short-range system for donor-recipient matching in bone marrow transplantation and HLA-related disease studies. In fact, one of the main themes of the upcoming 17th International HLA and Immunogenetics Workshop (IHIWS) in 2017 [53] is “NGS of full length HLA genes,” with the following objectives: (1) to complete the sequence of all HLA alleles of the reference cell lines from the 13th IHIWS and (2) to perform HLA genotyping of 10,000 quartet families of varied ancestry, utilizing at least one NGS method.

4.1.2. Development of multiplex PCR methods

Recently, we developed four kinds of multiplex PCR methods based on the long-range system for genotyping nine HLA loci (HLA-A, -B, -C, -DRB1/3/4/5, -DQB1, and -DPB1) [54] (Figure 3).
The multiplex PCR methods contributed greatly to simplifying, accelerating, and reducing costs and the number of reagents for the PCR step that is used to prepare samples and libraries for NGS in the NGS-based HLA genotyping method. The multiplex methods also conserved on the amounts of DNA samples needed to genotype a multiple number of HLA loci. Overall, the multiplex PCR method is a powerful tool for providing precise genotyping data without phase ambiguity, with a strong potential to replace the current routine genotyping methods to find polymorphisms. Commercialized PCR amplification reagents such as NEType (One-Lambda) that are based on multiplex PCR methods will be made available in the near future, whereas those based on the one-locus, one-tube PCR methods (left side of Figure 3) such as the TruSight HLA panel (Illumina) and NGSgo (GenDX) are already available in the market place.

4.2. NGS step

Although the 454 GS FLX was used often in the early stages of development of NGS-based HLA genotyping, the benchtop next-generation sequencers such as the GS Junior system, Ion Torrent PGM system, and the MiSeq system have been used more recently for the development and application of the HLA genotyping methods (Table 2). At the moment, complicated operations such as the preparation of NGS libraries are necessary for each of the different second generation sequencing platforms. However, the NGS companies are attempting to overcome these procedural bottlenecks by simplifying, automating, and speeding-up of the preparatory steps for NGS. For example, a new protocol using Ion Isothermal Amplification Chemistry that enables sequence reads of up to and beyond 500 bp, and Ion Hi-Q™ Sequencing Chemistry that reduces consensus insertion and deletion (indel) errors, including homopolymer errors, might lead to further simplification and cost reduction with higher data quality.

4.3. Allele assignment step

A variety of different allele assignment methods have been developed with some allele assignment software packages such as Assign (CONEXIO), OMIXON Target (OMIXON), and NGSengine (GenDX) commercially available, and others such as TypeStream (Life Technologies) still to be made commercially available in the near future. From our knowledge, Assign and NGSengine only support NGS data obtained from the one-locus, one-tube PCR method, whereas OMIXON Target and TypeStream also support NGS data obtained by the multiplex PCR methods. However, accuracy rates of the assignment methods are not 100% with genotyping errors caused by (1) missing HLA allele sequences, (2) generation of excessive allelic imbalance (ratio of sequence read numbers of allele 1 and allele 2), and (3) interference of HLA-DRB1 genotyping by participation of sequence reads originating from highly homologous HLA-DRB3/4/5 and other HLA-DRB pseudogenes. To avoid the errors raised in point 1, it is necessary to have a full and proper collection of all the HLA allele sequences to achieve precise HLA genotyping. In this regard, a much greater collection of high-quality full-length HLA allele sequences are expected to be obtained by way of international collaborations at the 17th IHIWS meeting in 2017 [53].
Recently, we developed a new next sequence allele assignment program (Sequence Alignment-Based Assigning Software; SeaBass) to solve the problems previously outlined in points 2 and 3 above. The program includes (1) output of sequence reads, (2) homology search using the Blat program [55] with the “match” variable set to 100% to detect identical exons within the known HLA alleles released from the IMGT-HLA database [7], (3) selection of allele candidates, (4) mapping of the sequence reads to the selected allele candidates as references with the “match” set at 100% using Reference Mapper (Roche), (5) calculation of coverages, and (6) confirmation of the mapping data and allele assignment (Figure 4).

The operations from Eqs. (2) to (5) are automatically processed. If a new polymorphism is included in the exon, we can detect its presence at the Blat search stage as shown in Figure 5, and if a new polymorphism is included in the intron, we can detect its presence during the calculation of the coverage and the final confirmation stages (Figure 6).

After the detection of the new polymorphisms, we further confirm them by traditional methods such as Sanger sequencing and subcloning. In addition, we validated the use of the SeaBass assignment methods for three next-generation sequencers, the GS Junior system, the Ion Torrent PGM system, and the MiSeq system. To evaluate the SeaBass program, we used a total of 2414 HLA sequences from all the classical HLA loci that have frequent HLA alleles in Caucasians, African-Europeans, and Japanese, and we obtained an overall accuracy rate of >99.8% and 100% for the Japanese subjects (Table 3).

The accuracy rate was not 100% for HLA-DRB1/3/4/5 and HLA-DPB1 of the non-Japanese subjects because the complete coding sequences have not been determined as yet for some of their HLA-DRB and HLA-DPB1 alleles. Nevertheless, the allele assignment method that we developed for SeaBass appears to be the most accurate and efficient way to detect new and null alleles by NGS.
Figure 5. Detailed information concerning selection of allele candidates using the SeaBass computer program. (A) “Extraction of allele candidates” by Blat search. We select allele candidates that are extracted in all of the exons. (B-1) New allele detection. In this example, one allele was called B*15:18:01, but the other allele was called B*44:03:01 excluding the exon 3. (B-2) Confirmation of the new allele by NGS. Mapping of the sequence reads with B*44:03:01 as a reference suggested six nucleotide differences with B*44:03:01 were detected in exon 3. We confirmed the polymorphisms by Sanger sequencing and deposited the sequence to DDBJ and IMGT-HLA database. Now the formal allele name is B*44:184 [94].
5. NGS-based MHC genotyping methods in nonhuman species

NGS technology provides the opportunity to genotype MHC sequences either by PCR targeted DNA sequencing or by PCR targeted RNA sequencing, that is, by DNA sequencing after converting the RNA samples to cDNA by reverse transcriptase. Usually, one or other of the
sequencing methods is chosen rather than using both methods on the same samples. In the following sections, we compare the use and limitations of targeted NGS sequencing using DNA or RNA samples for MHC genotyping of MHC class I and class II genes in nonhuman species such as the Filipino cynomolgus macaques.

5.1. Advantage and disadvantage of using DNA and RNA samples for NGS

Table 4 shows a summary of the advantages and disadvantages of using DNA and RNA samples for NGS-based MHC genotyping.

<table>
<thead>
<tr>
<th></th>
<th>DNA</th>
<th>RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty of sampling</td>
<td>Easy</td>
<td>Difficult</td>
</tr>
<tr>
<td>Extraction cost of nucleic acid</td>
<td>Cheap</td>
<td>Expensive</td>
</tr>
<tr>
<td>Preparation before PCR</td>
<td>No</td>
<td>RT reaction</td>
</tr>
<tr>
<td>Primer location</td>
<td>Both of exons and introns</td>
<td>Exons only</td>
</tr>
<tr>
<td>Required sequence read number</td>
<td>Few</td>
<td>Many</td>
</tr>
<tr>
<td>Exclusion of pseudogene</td>
<td>Difficult</td>
<td>Easy</td>
</tr>
<tr>
<td>Estimation of expression level</td>
<td>Impossible</td>
<td>Possible</td>
</tr>
</tbody>
</table>

Table 4. Advantages and disadvantages of DNA and RNA samples for NGS-based MHC genotyping
The advantages of using DNA samples instead of RNA samples are that (1) the sampling and extraction of the DNA nucleic acids are easier and cheaper than RNA samples, (2) PCR amplification can be performed directly without an additional reaction such as the reverse transcriptase (RT) reaction, (3) design of primers in the exon and intron regions, and (4) fewer read sequences are required for DNA than RNA samples if all alleles are amplified without allelic imbalance. Although many more read sequences are necessary for RNA samples than DNA samples to genotype all the MHC alleles that have different transcription levels, the advantages of using RNA samples for genotyping are that (1) they provide an opportunity to examine MHC gene expression, (2) transcription levels are possible to be estimated for each of MHC alleles from the read sequence depth [56], and (3) only transcribed MHC genes are detected without contamination of PCR products originating from pseudogenes if the primer locations cross over to at least two homologous exons. Thus, the use of RNA samples is thought to be more effective for precise MHC genotyping on duplicated MHC genes that have high similarities among the genes. However, DNA and RNA samples have their own unique advantages and disadvantages for informative NGS-based MHC genotyping and widen the choices for experimentation and data collection.

5.2. Methodology

Table 5 shows a publication list of the MHC genotyping by PCR-based NGS methods in different animal species, and it includes the MHC species name, target gene, PCR method, degree of allele data accumulation, and the allele assignment method.

<table>
<thead>
<tr>
<th>Species</th>
<th>MHC name</th>
<th>Animal model or nonmodel type</th>
<th>Template</th>
<th>Target gene</th>
<th>NGS platform</th>
<th>Degree of allele data accumulation</th>
<th>Allele assignment method</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammal Rhesus macaque</td>
<td>Mamu</td>
<td>Model</td>
<td>RNA</td>
<td>Class I and 454, II</td>
<td>Illumina</td>
<td>Relatively rich</td>
<td>De novo assembly</td>
<td>[78, 79]</td>
</tr>
<tr>
<td>Cynomolgus macaque</td>
<td>Mafa</td>
<td>Model</td>
<td>RNA</td>
<td>Class I and 454, II</td>
<td>Illumina, PacBio</td>
<td>Relatively rich</td>
<td>De novo assembly</td>
<td>[35, 78, 80, 81]</td>
</tr>
<tr>
<td>Pig-tailed macaque</td>
<td>Mane</td>
<td>Model</td>
<td>RNA</td>
<td>Class I and 454, II</td>
<td>Illumina</td>
<td>Relatively rich</td>
<td>De novo assembly</td>
<td>[78, 82]</td>
</tr>
<tr>
<td>Swine SLA</td>
<td>SLA</td>
<td>Model</td>
<td>RNA</td>
<td>Class I</td>
<td>454</td>
<td>Relatively rich</td>
<td>De novo assembly</td>
<td>[56]</td>
</tr>
<tr>
<td>Grey mouse lemur</td>
<td>Mimu</td>
<td>Nonmodel DNA</td>
<td>DRB and DQB</td>
<td>454</td>
<td>Poor</td>
<td>De novo assembly</td>
<td>[83]</td>
<td></td>
</tr>
</tbody>
</table>
As discussed previously, for humans, the HLA alleles obtained by next-generation sequencers are mainly assigned by mapping to known allele sequences that are used as the read references because a large number of HLA allele sequences already have been collected in the IMGT-HLA database [7] (Table 2). On the other hand, de novo assembly of read sequences and subcloning of PCR products identifies novel allele sequences. Of the nonhuman species, RNA samples tend to be used for MHC genotyping in experimental animals (model animals) such as macaque species and swine, whereas DNA samples are mainly used for MHC genotyping wild (nonmodel) animals because collecting RNA samples from them in their natural environment is more difficult than sampling captured or domesticated experimental animals (Table 5).

### 5.2.1. MHC genotyping RNA samples collected from Filipino cynomolgus macaques

MHC alleles in humans and experimental animals such as the macaque species and swine are mainly assigned by mapping methods because of the large amount of MHC allele information.
already available for them than for most other species. This allele information is collected and released by the IPD-MHC database [57]. When novel alleles are detected, de novo assembly of the read sequences and subcloning of PCR products identifies the sequences.

We identified homozygous and heterozygous cynomolgus macaques (Mafa) that have specific Mafa MHC haplotypes by genotyping the MHC of more than 5000 Filipino animals, and we found that they have a smaller number of different Mafa-class I and Mafa-class II alleles than the Indonesian and Vietnamese populations. In this section, we outline the MHC genotyping method using RNA samples and provide some results as an example of the method. Figure 7 shows a comparative genomic map of MHC regions between human and Filipino cynomolgus macaque.

![Figure 7. Comparative genomic map of the human (HLA) and the Filipino cynomolgus macaque (Mafa) Class I and Class II transcribed genes.](image)

The MHC class I genomic region has many more Mafa-class I genes than HLA-class I genes generated by gene duplication events, whereas the organization of Mafa-class II genes are well conserved between the two species. Also, there are many Mafa-class I pseudogenes located in the Mafa-class I region. Therefore, we performed MHC genotyping by amplicon sequencing with the Roche GS Junior system using RNA samples from the Filipino cynomolgus macaques to prevent contamination of PCR products originating from the pseudogenes (Figure 8).

The workflow that we used is composed mainly of five steps: (1) RNA extraction and cDNA synthesis, (2) multiplex PCR amplification, (3) pooling of the PCR products, (4) amplicon NGS sequencing, and (5) allele assignment. In step 1, we usually extracted total RNA from the peripheral white blood cell samples using the TRIzol reagent (Invitrogen/Life Technologies/Thermo Fisher Scientific, Carlsbad, CA) and synthesized cDNA by oligo d(T) primer using the ReverTraAce for the reverse transcriptase reaction (TOYOBO, Osaka, Japan) after treatment of the isolated RNA with DNase I (Invitrogen/Life Technologies/Thermo Fisher Scientific, Carlsbad, CA). In step 2, we designed a single Mafa-class I-specific primer set in exon 2 and exon 4 (PCR product size: 514 bp or 517 bp) that could amplify all known Mafa-class I alleles, whereas the Mafa-class II locus-specific primer sets included the polymorphic exon 2 in Mafa-DRB (420 bp), Mafa-DQA1 (435 bp), Mafa-DQB1 (396 bp), Mafa-DPA1 (407 bp), and Mafa-DPB1 (333, 336 or 339 bp) for massively parallel pyrosequencing (Figure 9).
In addition to these primer sets, we also designed 50 different types of fusion primers that contained the 454 titanium adaptor (A in forward and B in reverse primer), 10 bp MID (multiple identifier), and MHC-specific primers (Figure 8). Moreover, we constructed a multiplex PCR method using the primer sets by carefully optimizing primer composition and PCR conditions and by comparing the sequence read data obtained by NGS (Figure 10).

As a result of these primer designs, 51.5%, 13.6%, and 8.6–8.9% of all read sequence numbers were detected in Mafa-class I, Mafa-DRB, and the other Mafa-class II genes, respectively, and we confirmed that the genotypes obtained by the multiplex PCR method were consistent with
our previous uniplex PCR methods. Therefore, the multiplex PCR method greatly simplified the procedures required in preparing the DNA samples for NGS by reducing the time of preparation and the amount and cost of reagents. In the pooling step of the PCR products, we quantified the purified PCR products by the Picogreen assay (Invitrogen) with a Fluoroskan Ascent micro-plate fluorometer (Thermo Fisher Scientific, Waltham, MA), mixed each of the PCR products at equimolar concentrations and then diluted them according to the manufacturer’s recommendation. In the NGS amplicon sequencing step, we perform emulsion PCR (emPCR) and emulsion-breaking according to the manufacturer’s protocol (Roche, Basel, Switzerland). After the emulsion-breaking step, we enriched and counted the beads carrying the single-stranded DNA templates, and deposited them into a PicoTiterPlate to obtain the sequence reads.

A schematic workflow of the allele assignment process as a follow on from Figure 8 is shown in Figure 11.

After the sequencing run, image processing, signal correction, and base calling are performed by the GS Run Processor Ver. 3.0 (Roche) with full processing for shotgun or paired-end filter analysis. Quality-filter sequence reads that are passed by the assembler software (single sff file) are binned according to the MID labels into each separate sequence sff file using the sff file software (Roche). These files are further quality trimmed to remove poor sequence at the end of the reads with quality values (QVs) of less than 20. After separation of the trimmed and MID-labeled sequence reads in each of forward and reverse side read sequences, we independently detect the Mafa-class I and Mafa-class II allele candidates from both sides of the forward and reverse reads by using the BLAT program to match the trimmed and MID labeled sequence reads at 99% and 100% identity while setting the minimum overlap length at 200 and the alignment identity score parameter at 10 against all the known Mafa-class I and Mafa-class II allele sequences released in the IMGT/MHC-NHP database [58]. After the extraction of common allele candidates from both sequencing sides, we finally assign the “real alleles” by confirming nucleotide sequences of the allele candidates using the GS Reference Mapper Ver.
3.0. To discover novel Mafa-class I sequences, we perform the de novo assembly set to detect >85% matches using the trimmed and MID-binned sequences after converting the outputs to ace files for the Sequencher Ver. 5.01 DNA sequence assembly software (Gene Code Co., Ann Arbor, MI). We then use the defined consensus sequence obtained from the de novo assembly as a reference sequence to identify and map the correct allele sequences. Using this process, we genotyped a set of 400 unrelated animals by the Sanger sequencing method and high resolution pyrosequencing and identified 190 different alleles, 28 at Mafa-A, 54 at Mafa-B, 12 at Mafa-I, 11 at Mafa-E, 7 at Mafa-F, 34 at Mafa-DRB, 13 at Mafa-DQA1, 13 at Mafa-DQB1, 9 at Mafa-DPA1, and 9 at Mafa-DPB1 alleles [35, 59].

On the basis of our large-scale project to genotype the MHC of 5000 Filipino cynomolgus macaques by NGS, we so far have detected 15 different types of Mafa haplotypes (HT1~HT15) in 45 homozygous animals. These Mafa homozygous animals provided the basis to efficiently estimate other Mafa haplotypes. For example, we estimated a variety of Mafa-A, Mafa-B/I, Mafa-E, and Mafa-class II (Mafa-DRB, Mafa-DQA1, Mafa-DQB1, Mafa-DPA1, and Mafa-DPB1) haplotypes by comparing the homozygous animals with heterozygous animals that

![Figure 11. A schematic workflow of the allele assignment process using the SeaBass software.](image-url)
carry the identical Mafa-class I and Mafa-class II alleles in the homozygous animals. In addition, we estimated the Mafa haplotypes and haplotype frequencies by the PHASE 2.1.1 program [60] using the allele data obtained by amplicon sequencing. From these procedures, we estimated a total of 84 Mafa-class I and 18 Mafa-class II haplotypes. Of the 15 different Mafa HT haplotypes, the haplotype frequencies of HT1, HT2, HT4, and HT8 were the highest. Of them, HT1 and HT8 have entirely different Mafa alleles, whereas HT2 and HT4 are thought to be recombinants of HT1 and HT8 (Figure 12).

![Gene composition of representative Mafa MHC haplotypes HT1 and HT8 and their recombinants HT2 and HT4.](image)

Figure 12. Gene composition of representative Mafa MHC haplotypes HT1 and HT8 and their recombinants HT2 and HT4.

Namely, the Mafa-A allele in HT2 is identical to that in HT8, whereas HT2 also has alleles at other loci that are identical with those in HT1. Similarly, HT4 has alleles in Mafa-class I loci that are identical with those in HT8, and alleles in the Mafa-class II loci that are identical with those in HT1. Therefore, Mafa homozygous animals with known haplotypes such as H1 and H2 are important for biomedical research, such as the transplantation outcomes of induced pluripotent stem (iPS) cells (Figure 13) because such studies are undertaken on animals with a defined genetic background and relatively well-characterized MHC haplotypes that might regulate the adaptive immune system in different ways and efficiencies.

5.2.2. MHC genotyping using DNA samples of wild animals

At this time in the development of MHC genotyping by NGS, it is difficult to apply the RNA-sequencing mapping method to accurately genotype the MHC of wild animals using known allele sequences as references. This is because the present allele information is relatively poor for most of them (Table 5). Therefore, MHC genotyping of wild animals or poorly studied species by NGS is based on de novo assembly of DNA sequences. In this case, the definition of “real alleles” and “artifact alleles” is important because NGS errors such as monostretch sequences are frequently observed in the assembled consensus sequences. Some of the allele assignment approaches based on de novo assembly that have been published include the allele validation threshold (AVT) method [61], clustering method [62–64], and the relative sequencing depth modeling methods [65]. These methods suppose that the contigs that have a sequence depth greater than the threshold level are the “real alleles,” and they are determined by
statistical calculation of the threshold using the sequence depth values of all contigs obtained in \textit{de novo} assembly. Therefore, the detection of exact or “real” alleles depends largely on the setting of the threshold level and the quality of the sequence reads \cite{65}. To enable the correct setting of the threshold level, it is important to use primers that can amplify all alleles of the target locus or loci without allelic imbalance. Furthermore, additional considerations such as repeating independent NGS experiments at least three times and detecting identical allele sequences in at least two animals are necessary to distinguish between real and artifactual alleles.

6. Conclusion

Genotyping the polymorphisms of MHC genes using targeted NGS technologies has been developed for humans and some nonhuman species to replace the use of other more cumbersome and less accurate procedures. We found that targeted NGS of DNA or RNA samples is feasible, productive, and generates high-quality MHC allele information from a large number of samples not easily achievable by other genotyping methods. We used second-generation sequencing protocols to target the DNA region and RNA subsets of interest in our NGS studies. It is likely that the longer sequence reads produced by third-generation platforms such as the...

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure13.png}
\caption{Application of Mafa homozygous and heterozygous animals for nonclinical trials of induced pluripotent stem (iPS) cells.}
\end{figure}
Pacific Biosciences single-molecule real-time sequencing or the Oxford nanopore sequencing platform will enable and improve the task of MHC sequence phasing and haplotyping, although this has yet to be demonstrated and proved to be advantageous and more economical. Continued allele data collection for different species, improvements to the reagents, protocols, and data analysis tools also are likely to simplify procedures and lower the costs of generating sequencing data in future. Most species have numerous highly polymorphic MHC loci; hence, the many benefits of using NGS technologies are likely, in the near future, to replace many of the traditional genotyping methods for the investigation of human and animal MHC genes and their role in evolutionary biology, ecology, population genetics, disease, and transplantation.

Author details

Takashi Shiina1*, Shingo Suzuki1 and Jerzy K. Kulski1,2

*Address all correspondence to: tshiina@is.icc.u-tokai.ac.jp

1 Division of Basic Medical Science and Molecular Medicine, Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa, Japan
2 Centre for Forensic Science, The University of Western Australia, Nedlands, WA, Australia

References


[28] O’Keefe GM, Nguyen VT, Benveniste EN. Regulation and function of class II major histocompatibility complex, CD40, and B7 expression in macrophages and microglia:


Kulski JK, Suzuki S, Ozaki Y, Mitsunaga S, Inoko H, Shiina T. In phase HLA genotyping by next generation sequencing—a comparison between two massively paral-


[55] Blat program. Available from: http://genome.ucsc.edu/


