Published in final edited form as:

Immunogenetics. 2014 January ; 66(1): 15–24. doi:10.1007/s00251-013-0744-3.

Full-length novel MHC class I allele discovery by next-generation sequencing: two platforms are better than one

Dawn M. Dudley^{1,2}, Julie A. Karl², Hannah M. Creager², Patrick S. Bohn², Roger W. Wiseman², and David H. O'Connor^{1,2,*}

¹Department of Pathology and Laboratory Medicine, University of Wisconsin-Madison, Madison, WI 53705

²Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, WI 53715

Abstract

Deep sequencing has revolutionized major histocompatibility complex (MHC) class I analysis of nonhuman primates by enabling high-throughput, economical, and comprehensive genotyping. Full-length MHC class I cDNA sequences, which are required to generate reagents such as MHC:peptide tetramers, cannot be directly obtained by short read deep sequencing. We combined data from two next-generation sequencing platforms to discover novel full-length MHC class I mRNA/cDNA transcripts in Chinese rhesus macaques. We first genotyped macaques by Roche/454 pyrosequencing using a 530 bp amplicon spanning the densely polymorphic exons 2 through 4 of the MHC class I loci that encode the peptide-binding region. We then mapped short paired-end 250 bp Illumina sequence reads spanning the full-length transcript to each 530 bp amplicon at high stringency and used paired-end information to reconstruct full-length allele sequences. We characterized 65 full-length sequences from 6 Chinese rhesus macaques. Overall, approximately 70% of the alleles distinguished in these 6 animals contained new sequence information, including 29 novel transcripts. The flexibility of this approach should make full-length MHC class I allele genotyping accessible for any nonhuman primate population of interest. We are currently optimizing this method for full-length characterization of other highly polymorphic, duplicated loci such as the MHC class II DRB and killer immunoglobulin-like receptors. We anticipate that this method will facilitate rapid expansion and near completion of sequence libraries of polymorphic loci, such as MHC class I, within a few years.

Keywords

Major histocompatibility complex class I; Chinese rhesus macaques; Roche/454 pyrosequencing; Illumina MiSeq sequencing; MHC class I genotyping

Introduction

MHC class I proteins present intracellular peptides to CD8⁺ T-cells and therefore play a crucial role in cellular immune responses mounted against infectious diseases. To accommodate the large variety of peptides present in possible infectious pathogens, the MHC class I binding domains are highly polymorphic and differ between individuals. In nonhuman primates, the gene segments encoding the MHC class I classical A and B proteins

*Corresponding author: David H. O'Connor, 585 Science Drive, Madison, WI 53711, Phone: (608) 265-3389, Fax: (608) 265-8084, dhocconno@wisc.edu.

Conflict of interest: The authors declare that they have no conflict of interest.

are also highly duplicated (Otting et al. 2005). As a result, and unlike humans, nonhuman primates express more than 2 copies of each MHC class I gene. In fact, one sequenced rhesus macaque genomic MHC class I haplotype carries 22 different loci (Daza-Vamenta et al. 2004). This complexity creates challenges when trying to characterize MHC class I transcripts since each individual may express an unpredictable number of highly similar alleles.

We previously established MHC class I genotyping methods for nonhuman primates using Roche/454 pyrosequencing with cDNA amplicons spanning portions of exons 2 through 4 that encode the peptide-binding domain. Sequencing this highly polymorphic region of the MHC class I transcript allows us to distinguish >75% of the alleles within an animal (Budde et al. 2011; Fernandez et al. 2011; Wiseman et al. 2009). While this region is sufficiently diverse to distinguish between essentially all allele lineages, subtle differences between alleles of the same lineage cannot always be resolved. Rhesus macaques have more than two hundred MHC class I lineages; however, there are often several allelic variants (typically 1-10, but up to 60 or more) within each lineage that may be indistinguishable by exon 2-4 sequencing alone (de Groot et al. 2012).

Allelic variants within a specific MHC class I lineage may have important consequences for certain disease or treatment outcomes. For example, *HLA-B*57:01* is associated with differential HIV disease progression and hypersensitivity reactions to abacavir treatment for HIV, while *HLA-B*57:02* and *-B*57:03* are not (Kloverpris et al. 2012; Mallal et al. 2002; Migueles et al. 2000; Stocchi et al. 2012). Specific MHC class I and II allelic variants have also been implicated in drug hypersensitivity to the anticonvulsant drug carbamazepine and the uric acid reducer allopurinol, as well as susceptibility to autoimmune diseases such as type I diabetes (Profaizer and Eckels 2012; Erlich et al. 2013). Most studies correlating MHC class I with disease or treatment in either humans or nonhuman primates do not distinguish between allelic variants of a lineage. This is due primarily to cost and time associated with traditional Sanger-based cloning and sequencing techniques used for full-length transcript characterization, prohibiting the use of these techniques for large studies. Therefore, the impact of allelic variants on disease and treatment remains deeply understudied. Techniques like the one presented here to sequence the full-length MHC class I coding region using next-generation sequencing are essential to begin to delineate how allelic variants affect a broad spectrum of diseases and treatments.

We previously utilized next-generation Roche/454 sequencing of three overlapping amplicons spanning the ~1200 bp full-length MHC class I mRNA transcripts in Mauritian cynomolgus macaques (Budde et al. 2011). However, this technique was not ideal, as the 3' amplicon often produced inconsistent sequencing results and it was often difficult to obtain robust PCR amplification of this amplicon in the first place. Insertion/deletion artifacts inherent to Roche/454 sequencing also confounded rapid assembly of the overlapping amplicons. It should be noted that Roche/454 sequence read lengths are not yet long enough to span the full-length MHC class I transcript with a single amplicon. To attempt to overcome these issues, we PCR amplified full-length MHC class I mRNA transcripts, created randomly fragmented amplicon libraries, and sequenced the resulting products on the Illumina MiSeq. This generated randomly distributed overlapping 250 bp paired-end sequences. However, *de novo* assembly of these short highly similar reads resulted in chimeric assemblies of different alleles and inaccurate transcript reconstruction. An alternative assembly approach for these MiSeq sequences is to use reference-guided assembly methods. The challenge of this approach is to identify an appropriate set of reference sequences that can be used for all animals.

Given the limitations of using Roche/454 pyrosequencing or Illumina sequencing alone, we devised a hybrid sequencing approach that uses data from both platforms to generate full-length MHC class I sequences at a scale that will allow rapid characterization of sequences present at polymorphic loci, such as MHC class I. We first used the Roche/454 platform to sequence a 530 bp amplicon representing exons 2-4 of each MHC class I transcript present in an animal. The resulting sequences were previously validated against those obtained via traditional cDNA cloning and Sanger-based sequencing to demonstrate that they provide accurate allele sequences when used for MHC class I genotyping (Wiseman et al. 2009; Fernandez et al. 2011; Karl et al. 2013; Wiseman et al. 2013). The 530 bp amplicon sequences were used as scaffolds, or reference sequences, for a multi-round reference-guided assembly of paired-end MiSeq reads to generate full-length MHC class I transcripts. Because our scaffold assembly process begins in the most highly diverse region of the MHC class I gene, we are less likely to assemble incorrect sequences to this region (those with an incorrect paired end) providing a much more accurate assembly than is possible by the alternative approach of unguided *de novo* assembly.

Here we present this new MHC class I allele discovery approach in the context of novel full-length allele discovery in six Chinese rhesus macaques. Due to the limited supply of well-characterized Indian rhesus macaques, scientists have begun using Chinese rhesus macaques to study infectious disease such as HIV/SIV, influenza, and malaria, in addition to xenotransplantation and stem cell therapies (Southwick and Siddiqui 1994; Chen et al. 2011; Chen et al. 2009; Choi et al. 2011; Hutnick et al. 2012; Li et al. 2012; Ling et al. 2013; Mumbauer et al. 2013; Wei et al. 2011). However, characterization of MHC class I sequences in the Chinese rhesus macaque population remains incomplete. We have reported that Chinese and Indian rhesus macaques share many MHC class I lineage-based haplotypes and that 85% of Chinese rhesus macaques express at least one haplotype shared between these two populations (Karl et al. 2013). However, these putative ancestral haplotypes often vary at the allelic level between the two populations. As is the case in human studies, how often these variants lead to different outcomes in macaques is as yet unexplored.

Using the described genotyping approach, we identified 65 full-length classical *Mamu-A* and *Mamu-B* sequences in 6 Chinese rhesus macaques. Almost half (46%) of these transcripts were novel, and another quarter (26%) of the alleles extended previously known alleles by >30 bp, exemplifying the power of this technique to rapidly expand our knowledge of MHC class I alleles in understudied nonhuman primate populations. Most of the novel alleles discovered are variants of known lineages that would have likely been ambiguous with previously described sequences by genotyping techniques that do not evaluate full-length transcripts. Because our sequences are full-length, they can be used to derive reagents such as MHC class I tetramers and anti-MHC class I antibodies to study the immunology associated with the many disease models that use nonhuman primates. In addition, this technique can also be adapted to other model organisms or humans. Lastly, this approach could be highly valuable for characterizing other duplicated and/or polymorphic loci that are currently very difficult to study such as killer immunoglobulin-like receptors, T-cell receptors, and immunoglobulin heavy chain variable regions.

Materials and Methods

Animals

Peripheral blood mononuclear cells (PBMCs) were obtained from six Chinese rhesus macaques from Battelle Biomedical Research Center. These animals were previously genotyped by sequencing MHC class I exons 2-4 and were assigned MHC class I haplotypes in our laboratory (Karl et al. 2013). The animals used in this study were selected because they were predicted to be rich in novel allelic variants based on this haplotyping (Table 1).

cDNA synthesis and PCR amplification

A schematic representation of the methods used for MHC class I genotyping is shown in Fig. 1. RNA was isolated from frozen PBMCs using the Roche MagNA Pure instrument and RNA High Performance kit (Roche, Indianapolis, IN, USA) following manufacturer's protocols (Fig. 1a and 1b). cDNA was synthesized using Oligo (dt) primers and Superscript III reverse transcriptase (Life Technologies, Grand Island, NY, USA) with equivalent concentrations of RNA for each sample (Fig. 1c). The entire ~1.2kb MHC class I coding region was amplified using Phusion high fidelity DNA polymerase (New England Biolabs, Ipswich, MA, USA) and the following primers shown in Online Resource 1 and as previously published: TiFL-MHC1-F_MIDXX (5'UTR), TiFL-MHC1-FL_MIDXX (5' Leader), TiFL-MHC3-Ra_MIDXX, TiFL-MHC3-Rb_MIDXX, and TiFL-MHC3-Rb2_MIDXX (Budde et al. 2010). MIDXX refers to Roche multiplex identifiers (MIDs), which vary for each sample, but is one of Roche MID tags 01-96 (XX) (i.e. MID01, MID02 etc). The three reverse primers were mixed together and used in conjunction with either the 5'UTR or 5'Leader forward primer in two separate reactions. Three reverse primers are used to PCR amplify different MHC class I transcripts that vary subtly under the reverse primer regions. Similarly, MHC class I transcripts that mismatch under the 5'UTR primer are typically amplified by the 5'Leader forward primer, ensuring that we maximize the number of transcripts amplified in each animal. A 530 bp region of MHC class I exons 2-4 was amplified using the previously described primers SBT568F and SBT568R (Fernandez et al. 2011)(Karl et al. 2013)(Wiseman et al. 2013). The cycling conditions for all amplicons were 98°C for 3 minutes, 25 cycles of 98°C for 5 seconds, 60°C for 10 seconds, and 72°C for 20 seconds followed by a 5 minute final elongation at 72°C. All primers contained MID tags and adaptors utilized for Roche/454 pyrosequencing (Roche, Indianapolis, IN, USA), including those amplicons used for Illumina sequencing. All PCR products were confirmed on a FlashGel (Lonza Group Ltd, Basel, Switzerland).

Roche/454 pyrosequencing

The 530 bp PCR amplicon of MHC class I exons 2-4 was purified twice using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA) at a ratio of 1:1 (DNA: bead). Purified products were pooled together at equimolar ratios based on known amplicon size verified by gel electrophoresis concentration and as quantitated using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Life Technologies, Grand Island, NY, USA). Pools were subjected to emulsion PCR, breaking and enriching, and sequencing on a Roche/454 GS Junior using Titanium technology (Fig. 1d) according to the protocol provided by the manufacturer (Roche, Indianapolis, IN, USA).

Illumina MiSeq library generation and sequencing

PCR amplicons of full-length MHC class I cDNAs starting in the 5' UTR and at the 5' leader peptide region of the gene were pooled together for each animal. These pooled ~1.2kb amplicons were purified using Agilent AMPure XP (Beckman Coulter, Brea, CA, USA) beads at a DNA:bead ratio of 1:1 and then fragmented using Nextera XT tagmentation (Illumina, San Diego, CA, USA) (Fig. 1e). This process also adds Illumina IL7 and IL5 indices to each sample that allow deconvolution of pooled samples after sequencing. Libraries of fragmented products yielded average fragment lengths ranging from 606 bp to 663 bp as determined by bioanalysis with a high sensitivity chip (Agilent Technologies, Santa Clara, CA, USA). Libraries for each sample were Ampure XP bead purified (5:3 DNA:bead ratio), quantitated as described above, and normalized to 2 nM. Samples were pooled together, denatured according to the Illumina MiSeq sample preparation protocol using NaOH, and run on an Illumina MiSeq using a 500 cycle MiSeq Reagent Kit v2 (Illumina, San Diego, CA, USA) (Fig. 1f).

Data analysis: Roche/454 pyrosequencing

All data analysis was performed with Geneious Pro software (v6.1.2) (Biomatters Limited, Auckland, New Zealand) using a semi-automated pipeline. A schematic representing analysis of Roche/454 pyrosequencing and Illumina MiSeq data is shown in Fig. 2. FASTQ data files separated by MID tags from three Roche/454 runs of the 530 bp product were consolidated and processed in Geneious. Briefly, reads were trimmed at both ends with an error probability limit of 0.01. Reads representing each animal (those with a single MID) were *de novo* assembled at a minimum overlap identity of 100% with 0% maximum mismatches per read. Contigs with fewer than five sequences representing a variant were merged and not treated as independent variants due to low sequence coverage. Therefore, only variants represented by five or more sequences were analyzed. A consensus sequence representing each contig was extracted and grouped into a sequence list. These sequences represent either the forward or reverse reads from each allele assembled together for each sample. Primer sequences were removed from these consensus sequences and they were trimmed to remove low quality regions from the 5' and 3' ends using an error probability limit of 0.0001, 0 maximum low quality bases and 0 maximum ambiguities. These sequences were then *de novo* assembled again at reduced stringency allowing gaps, up to 1 mismatches/read and a minimum overlap identity of 99%. The consensus sequences from this second *de novo* assembly were trimmed again at an error probability limit of 0.0001. These trimmed sequences represented the forward and reverse reads of each allele assembled together to generate a 530 bp product. These consensus sequences were exported and used as a scaffold to which MiSeq reads were assembled in Geneious (Fig. 2a).

Assembly of MiSeq sequences to extend the 530 bp region of MHC class I

FASTQ files separated by index were extracted from the Illumina MiSeq and imported into Geneious for all sequence trimming and assembling. The parameters used for each step in the Geneious program are as described. First, forward and reverse reads for each sample were paired together. The paired reads were trimmed for quality using the modified –Mott algorithm as implemented in Geneious 6.1 (www.geneious.com/assets/documentation/geneious/GeneiousManual.pdf) and an error probability limit of 0.01. Paired reads with both sequences longer than 150 bp were extracted and primer sequences were trimmed. Sequences longer than 100 bp in length were extracted and used in a reference-guided assembly. All MiSeq sequences were assembled against the scaffolds (consensus sequences) built from the 454 pyrosequencing data representing all MHC class I alleles from a single animal (Fig. 2b). Gaps were not allowed, maximum mismatches per read was set to 0%, minimum overlap identity was set to 100% and best matches were set to “map to none” to prevent the same sequence read from mapping to more than one 530 bp scaffold. Therefore, any sequences mapping perfectly to two or more scaffolds were excluded from downstream analysis. Each assembly was examined to ensure reads were mapping with 100% stringency to each other and that false mappings were not included. A consensus sequence was then generated for each allelic variant. These consensus sequences were trimmed to an error probability limit of 0.0001 on both 5' and 3' ends with 0 maximum low quality bases and 0 maximum ambiguities. This set of consensus sequences for each animal was then used as a reference database to perform reference-based assembly of the paired-end MiSeq sequences again to extend out each consensus sequence (Fig. 2c). The assembled consensus sequences were trimmed using the same parameters described above and used as a reference database against which the original MiSeq reads were assembled again in the same way. This process was repeated a total of approximately seven iterations per animal, or until alleles were no longer extended (Fig. 2d). Each iteration was completed within a few minutes and the overall analysis pipeline to generate full-length allele sequences can be completed within a day for 6 animals on a server equipped with an Intel Xeon X5650@2.67 Ghz (24 core) CPU with 64GB of RAM. The stringent quality trimming applied to all MiSeq reads before

assembly and to all consensus sequences during each iteration ensured that only high quality sequence throughout each read was used in the analysis. Though overall depth of coverage decreased near the ends of the alignments, the quality of the sequence reads in those regions was identical to those located in the middle of the alignment.

Characterization of MHC class I transcripts

The final consensus sequences generated with the MiSeq assemblies (Fig. 2d) were compared against a local curated database containing all known rhesus macaque MHC class I alleles using the Basic Local Alignment Search Tool (BLAST). From this BLAST search, novel alleles, extensions to known alleles, and known alleles were determined for each animal. Number of full-length versus partial-length transcripts was also recorded. Novel alleles were submitted to GenBank (accession #s KF297354-KF297369 and see Online Resource 2) and the Immuno Polymorphism Database (IPD) for official nomenclature designation (Robinson et al. 2000; Robinson et al. 2013).

Results and Discussion

Roche/454 pyrosequencing of 530 bp of MHC class I amplicons

The MHC class I transcripts discovered by this method are ultimately dependent upon the allele sequences generated by the 530 bp amplicon; there is no scaffold to build the full-length alleles using MiSeq sequences from any transcripts missed by this amplicon. Although Roche/454 pyrosequencing of this 530 bp amplicon was previously validated as a high-throughput and accurate genotyping approach, we wanted to further establish that this amplicon would detect most of the transcribed MHC class I alleles in each animal (Fernandez et al. 2011; Wiseman et al. 2009; Karl et al. 2013; Wiseman et al. 2013). The 530 bp amplicon was sequenced three times from each Chinese rhesus macaque on three independent Roche/454 GS Junior sequencing runs. For each sequence run, 57 amplicons were pooled together from different animals for multiple projects, including the 6 Chinese rhesus macaques presented here. An average of 743 sequences (range: 400-1119 sequences) were obtained for each animal in each 454 GS Junior run. After sequence reads were compiled together from all three runs into one file using the concatenate command in the command line, each animal was represented by an average of 2230 sequences (range: 1199-3357 sequences). Using the compiled data we detected an average of 13 MHC class I *Mamu-A* and *Mamu-B* transcripts per animal. Of note, because we start with mRNA rather than genomic DNA, we only detect transcribed sequences while excluding the pseudogenes that comprise a significant fraction of the MHC in nonhuman primates. Also, to reduce the potential of analyzing sequencing artifacts as alleles, we set the minimum threshold of sequences at five sequence reads for the generation of a consensus sequence representing an allelic variant transcript, thereby limiting detection of alleles transcribed at very low levels. Previous work has suggested that there are two to four functional *Mamu-A* alleles and seven to fourteen functional *Mamu-B* transcripts per haplotype in individual rhesus macaques (Daza-Vamenta et al. 2004; Wiseman et al. 2009; Doxiadis et al. 2013). We detected an average of 4 distinct *Mamu-A* transcripts (range 3-5) and 9 *Mamu-B* transcripts per animal (range: 7-12). These data indicate that our Roche/454 method is detecting the majority of the functionally active and highly transcribed alleles in our Chinese rhesus macaques. Furthermore, this method detects many more minor transcripts than the four to six *Mamu-A* and *Mamu-B* transcripts typically identified when sequencing 96 clones per sample using traditional cloning and sequencing techniques. Lastly, this method was directly validated in a seventh Chinese rhesus macaque sample for which four major MHC class I transcripts were previously identified using traditional Sanger-based cloning and sequencing methods. All four alleles were successfully recreated using this method (data not shown).

MHC class I alleles are transcribed at different levels and there is some evidence to suggest that those alleles most abundantly transcribed may be most functionally relevant in immune responses against disease (Budde et al. 2011). We have previously defined abundant transcripts, or major alleles, as those whose steady state levels are >4% of the total MHC class I transcripts, as quantitated by the number of sequence reads representing each allelic variant relative to the total number of MHC class I sequence reads obtained per animal by Roche/454 sequencing (Karl et al. 2013). To determine the number of sequence reads required to pick up both major and minor transcripts for allele discovery, we compared the allelic variants found in each animal when analyzing reads from a single GS Junior run to those found when sequences from all three runs were analyzed together. We found an average of two additional allelic variants per animal when using data from all three runs concatenated together (data not shown). The additional alleles were generally less transcriptionally abundant (< 1%), indicating that more Roche/454 sequences allow better detection of less abundantly transcribed MHC class I alleles. Based on this information, we recommend pooling no more than 20 animals per Roche/454 GS Junior run to optimize discovery of less abundantly transcribed alleles by sequencing each animal more deeply.

Illumina MiSeq sequencing of full-length MHC class I

A total of 17 samples from different projects were pooled together for the MiSeq run, including the 6 Chinese rhesus macaque MHC class I amplicon libraries described here. The cluster density of the MiSeq run was 966k/mm² and a total of 33,590,174 paired-end sequence reads were obtained. An average of 1,500,000 sequences were obtained per animal after merging forward and reverse paired-end reads together. As expected, the number of sequences aligning to each allele identified by Roche/454 pyrosequencing varied based on the relative transcription levels of each allele. Transcriptionally abundant (>4%) major alleles had a mean coverage at each nucleotide site across the coding region of 10,000-20,000 reads, while less abundant alleles were typically represented by a mean of hundreds of sequences at each nucleotide site.

The number of sequences representing each nucleotide across the coding region varied within each transcript, but in general showed a pattern of lower coverage nearing the ends and higher coverage in the middle of the transcript. This reduction in coverage near the ends is due to the nature of using transposon-based fragmentation. We fragmented our amplicons using a modified transposon (Illumina Nextera XT kit) that cleaves the DNA and leaves behind a target sequence that is used in a limited cycle PCR to add Illumina adaptors and indices onto each fragment for cluster formation and multiplex sequencing. However, only fragments that have this transposon sequence at each end can incorporate adaptors and be sequenced on the Illumina MiSeq. It is relatively rare for transposons to integrate into the DNA at the very ends of the primary full-length cDNA amplicons, which results in lower sequence coverage of these regions. This is particularly important to note because the MHC class I start and stop codons are located near the ends of our full-length amplicon. To improve the coverage of the start and stop codon sequences without altering our optimized primer locations, we used fusion primers from previous work that contained a Roche/454 adaptor and MID sequences. These 35 bp adaptor/MID sequences, which could be replaced by any random sequence, provides a sufficient buffer on each end of the amplicon for some transposons to integrate into the gene outside of the start and stop codons. Overall, coverage of the 5' end including the start codon was lower than the 3' end containing the stop codon. However, the coverage of the start codon averaged 791 sequences (range 3-4829) across the different alleles from all 6 animals and is more than sufficient to generate confident sequences encompassing the start and stop codons. It should also be reiterated that regions of the MHC class I gene represented by smaller numbers of sequences are still represented by high quality sequences due to the stringent trimming parameters used during the analysis.

process. Altogether, this data shows that transposon-based fragmentation of the amplicons generated with buffer sequences on the ends in conjunction with multiplexing up to 17 samples in a MiSeq run can yield enough sequence reads to build out full-length sequences of MHC class I coding regions with sufficient coverage of the start and stop codon.

Characterization and identification of MHC class I alleles found in 6 Chinese rhesus macaques

After assembling the MiSeq reads to the Roche/454 530 bp scaffold, a consensus sequence representing each transcript was compared to a curated database of all known rhesus macaque MHC class I alleles using BLAST. While several *Mamu-I* transcripts were identified and submitted to IPD and GenBank from these animals, for the purpose of this work we focused on the classical *Mamu-A* and *Mamu-B* transcripts due to their known importance in disease and transplant studies. Allelic variants that contained at least the leader peptide region through the stop codon were considered full-length sequences for this analysis. Full leader peptide sequence is not always required for generation of biological reagents. For example, the leader peptide sequence is often removed prior to tetramer generation. Ninety percent (65) of the 72 transcripts found in our 6 Chinese rhesus macaques were full-length using this criterion. Eighty five percent of these sequences contained both a start and stop codon. Of the 65 unique full-length *Mamu-A* and *Mamu-B* transcripts identified by our combined Roche/454 and Illumina approach, 19 (29.2%) sequences were previously known with little additional sequence information added by our methods (Table 2). In this category, up to nine new nucleotides could be added to the known allele, typically expanding the sequence near the start codon from one methionine to another putative start codon three amino acids upstream. Seventeen (26.2%) of the 65 transcripts we identified added sequence (30+ bp) to known alleles, referred to as extensions of known alleles (Table 2). Lastly, we discovered 29 (44.6%) novel full-length transcripts that were not present in either GenBank or the IPD database (Table 3).

The breakdown of known alleles, extensions to known alleles, and novel alleles for *Mamu-A* and *Mamu-B* allelic variants in our six Chinese rhesus macaques are shown in Fig. 3. Combining the extensions to known alleles and novel transcripts, we obtained new sequence information for a total of 46 allelic variants (71%) in just six animals, 91% of which are full-length sequences. In addition, one of the novel alleles discovered (*Mamu-B*193:01*) represents a new nonhuman primate lineage (Table 3). This expands the 272 known, full-length Chinese rhesus MHC class I alleles that have been officially named by IPD, by 15% after examining just 6 animals.

This method holds promise of quickly expanding MHC class I databases of nonhuman primate populations commonly used in biomedical research. These databases allow the development of rapid genotyping tools to characterize animals prospectively and retrospectively for research studies. This characterization is important to design well-powered studies to test the function of particular alleles in disease progression or in a vaccine modality. In other contexts, distributing diversity of the MHC class I haplotypes of animals between control and test groups may be imperative to ensure that results are not biased due to uneven distribution of a disease-influencing allele or alleles between groups. In addition, tissue transplant studies may need to match the level of MHC class I disparity between or within study groups. Lastly, tools required to delineate specific CD8+ T-cell responses and MHC class I epitopes important in disease or vaccine models, such as tetramers, monoclonal antibodies, and cell lines can be more easily developed given the full-length sequences of alleles that this method provides. Altogether, generating comprehensive full-length MHC class I allele databases for nonhuman primates used in biomedical research

will vastly improve these disease models by increasing the capacity for important immunological studies.

In conclusion, we have described a new approach that utilizes two distinct next-generation sequencing platforms to define full-length MHC class I transcripts in Chinese rhesus macaques. Used in conjunction, these two platforms overcome difficulties found using either platform alone. This approach is becoming evermore feasible as it is increasingly common to find both of these benchtop sequencing platforms in individual labs or genome centers and the multiplexing capabilities of each sequencing method reduces the cost to that below traditional Sanger-based sequencing even with two sequence runs per sample. The semi-automated analysis pipeline allows quick and consistent unbiased assemblies that generally lead to full-length allele sequences. These can then be used to generate reagents such as tetramers and antibodies that are required to determine the role of MHC class I in a variety of disease and tissue transplant models. We are currently using this technique to describe new MHC class I alleles in cynomolgus macaque populations from multiple geographic origins as well as pigtailed macaques to increase our knowledge base of MHC class I in these species. This approach is also ideal for full-length KIR characterization where differences in sizes of the full-length KIRs within each animal make balancing pyrosequencing runs extremely difficult. Lastly, these methods could easily be adapted to other highly polymorphic loci such as T-cell receptors and immunoglobulin variable chains along with the less complex MHC class I, class II, and KIR genes of humans, as well as similar loci in other model organisms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank Gabriel Starrett for bioinformatics support. We thank Batelle Biomedical Research Center for providing us with Chinese rhesus macaque samples. This research was supported by the Department of Health and Human Services Public Health Service grant from the National Institutes of Health under award R24 OD011048.

References

- Budde ML. Transcriptionally abundant major histocompatibility complex class I alleles are fundamental to nonhuman primate simian immunodeficiency virus-specific CD8+ T cell responses. *J Virol.* 2011; 85:3250–3261. [PubMed: 21270169]
- Budde ML, Wiseman RW, Karl JA, Hanczaruk B, Simen BB, O'Connor DH. Characterization of Mauritian cynomolgus macaque major histocompatibility complex class I haplotypes by high-resolution pyrosequencing. *Immunogenetics.* 2010; 62:773–780. [PubMed: 20882385]
- Chen S, Lai C, Wu X, Lu Y, Han D, Guo W, Fu L, Andrieu JM, Lu W. Variability of bio-clinical parameters in Chinese-origin Rhesus macaques infected with simian immunodeficiency virus: a nonhuman primate AIDS model. *PLoS One.* 2011; 6:e23177. [PubMed: 21850259]
- Chen Y. Pathological lesions and viral localization of influenza A (H5N1) virus in experimentally infected Chinese rhesus macaques: implications for pathogenesis and viral transmission. *Arch Virol.* 2009; 154:227–233. [PubMed: 19130169]
- Choi HJ, Kim MK, Lee HJ, Ko JH, Jeong SH, Lee JI, Oh BC, Kang HJ, Wee WR. Efficacy of pig-to-rhesus lamellar corneal xenotransplantation. *Invest Ophthalmol Vis Sci.* 2011; 52:6643–6650. [PubMed: 21743020]
- Daza-Vamenta R, Glusman G, Rowen L, Guthrie B, Geraghty DE. Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Res.* 2004; 14:1501–1515. [PubMed: 15289473]

- de Groot NG. Nomenclature report on the major histocompatibility complex genes and alleles of Great Ape, Old and New World monkey species. *Immunogenetics*. 2012; 64:615–631. [PubMed: 22526602]
- Doxiadis GG. Haplotype diversity generated by ancient recombination-like events in the MHC of Indian rhesus macaques. *Immunogenetics*. 2013; 65:569–584. [PubMed: 23715823]
- Erlich HA, Valdes AM, McDevitt SL, Simen BB, Blake LA, McGowan KR, Todd JA, Rich SS, Noble JA. Next Generation Sequencing Reveals the Association of DRB3*02:02 With Type 1 Diabetes. *Diabetes*. 2013; 62:2618–2622. [PubMed: 23462545]
- Fernandez CS, Reece JC, Saepuloh U, De Rose R, Ishkandriati D, O'Connor DH, Wiseman RW, Kent SJ. Screening and confirmatory testing of MHC class I alleles in pig-tailed macaques. *Immunogenetics*. 2011; 63:511–521. [PubMed: 21556859]
- Hutnick NA. An optimized SIV DNA vaccine can serve as a boost for Ad5 and provide partial protection from a high-dose SIVmac251 challenge. *Vaccine*. 2012; 30:3202–3208. [PubMed: 22406458]
- Karl JA, Bohn PS, Wiseman RW, Nimityongskul FA, Lank SM, Starrett GJ, O'Connor DH. Major Histocompatibility Complex Class I Haplotype Diversity in Chinese Rhesus Macaques. G3 (Bethesda). 2013
- Kloverpris HN. HLA-B*57 Micropolymorphism shapes HLA allele-specific epitope immunogenicity, selection pressure, and HIV immune control. *J Virol*. 2012; 86:919–929. [PubMed: 22090105]
- Li Q, Ruan Z, Zhang H, Peng N, Zhao S, Qin L, Chen X. Characterization of peripheral blood T lymphocyte subsets in Chinese rhesus macaques with repeated or long-term infection with *Plasmodium cynomolgi*. *Parasitol Res*. 2012; 110:961–969. [PubMed: 21842385]
- Ling B, Rogers LB, Johnson AM, Piatak M, Lifson J, Veazey R. Effect of Combination Antiretroviral Therapy on Chinese Rhesus Macaques of SIV infection. *AIDS Res Hum Retroviruses*. 2013
- Mallal S. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet*. 2002; 359:727–732. [PubMed: 11888582]
- Miguel SA. HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc Natl Acad Sci U S A*. 2000; 97:2709–2714. [PubMed: 10694578]
- Mumbauer A, Gettie A, Blanchard J, Cheng-Mayer C. Efficient mucosal transmissibility but limited pathogenicity of R5 SHIVSF162P3N in Chinese origin rhesus macaques. *J Acquir Immune Defic Syndr*. 2013
- Otting N, Heijmans CM, Noort RC, de Groot NG, Doxiadis GG, van Rood JJ, Watkins DI, Bontrop RE. Unparalleled complexity of the MHC class I region in rhesus macaques. *Proc Natl Acad Sci U S A*. 2005; 102:1626–1631. [PubMed: 15665097]
- Profaizer T, Eckels D. HLA alleles and drug hypersensitivity reactions. *Int J Immunogenet*. 2012; 39:99–105. [PubMed: 22136512]
- Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SG. IPD--the Immuno Polymorphism Database. *Nucleic Acids Res*. 2013; 41:D1234–D1240. [PubMed: 23180793]
- Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE. IMGT/HLA - a sequence database for the human major histocompatibility complex. *Tissue Antigens*. 2000; 55:280–287. [PubMed: 10777106]
- MF SCHaS. Population status of nonhuman primates in Asia, with emphasis on rhesus macaques in India. *American Journal of Primatology*. 1994; 34(1):51–59.
- Stocchi L, Cascella R, Zampatti S, Pirazzoli A, Novelli G, Giardina E. The Pharmacogenomic HLA Biomarker Associated to Adverse Abacavir Reactions: Comparative Analysis of Different Genotyping Methods. *Curr Genomics*. 2012; 13:314–320. [PubMed: 23204921]
- Wei Q, Sun Z, He X, Tan T, Lu B, Guo X, Su B, Ji W. Derivation of rhesus monkey parthenogenetic embryonic stem cells and its microRNA signature. *PLoS One*. 2011; 6:e25052. [PubMed: 21966410]
- Wiseman RW. Major histocompatibility complex genotyping with massively parallel pyrosequencing. *Nat Med*. 2009; 15:1322–1326. [PubMed: 19820716]

Wiseman RW, Karl JA, Bohn PS, Nimityongskul FA, Starrett GJ, O'Connor DH. Haplessly Hoping: Macaque Major Histocompatibility Complex Made Easy. ILAR Journal. 2013 In press.

Fig. 1.

Preparation and sequencing of samples. a) Whole blood from each animal is separated using Ficoll-Paque centrifugation and PBMCs are collected. b) Total RNA is isolated from PBMCs using standard RNA isolation methods and cDNA is synthesized. c) The MHC class I cDNAs are amplified as a full-length PCR product or a fragment containing parts of exons 2-4. d) The exon 2-4 amplicon containing MID tags and adaptors added during the primary PCR is sequenced by Roche/454 pyrosequencing using a GS Junior. e) The full-length PCR amplicon is fragmented using Nextera XT transposons. f) Fragmented full-length amplicon libraries containing indices added during fragmentation undergoes paired-end Illumina MiSeq sequencing.

Fig. 2.

Sequence analysis overview. a) Forward and reverse sequence reads obtained from Roche/454 sequencing are joined based on unique SNPs in the overlap region. b) The 530 bp Roche/454 joined sequences are used as a scaffold to assemble the full-length fragmented sequences obtained from Illumina MiSeq analysis. c) Paired-end reads lying outside of the 530 bp scaffold are used to expand the 530 bp sequence. These newly elongated sequences are used as the backbone to expand out the MHC class I transcript over multiple iterations. d) The full-length molecule is generated when iterations no longer expand the sequence length

Fig. 3. Distribution of *Mamu-A* and *Mamu-B* transcripts identified in 6 Chinese rhesus macaques. a) Percentage of unique novel, extensions to known, and known *Mamu-A* transcripts found in six animals. b) Percentage of unique novel, extensions to known, and known *Mamu-B* transcripts found in six animals. n, the number of transcripts found in each category

Table 1

MHC class I *Mamu-A* and *Mamu-B* haplotypes^a of 6 Chinese rhesus macaques.

Animal	ChrH01	ChrH02	ChrH03	ChrH04	ChrH05	ChrH06
A-haplotypes	A026	A010	A004	A004	A018b	A018a
	A056b	A074	A092a	A004	A019	A056b
B-haplotypes	B056a	B039a	B056b	B010b	B023	B013b
	B150	B056a	B085	B045a	B136	B015b

^a Haplotypes were previously determined based on MHC class I exons 2-4 sequencing (Karl et al. 2013).

Table 2

Previously identified (known) alleles and extensions of known alleles present in the IPD database^{a,b} that were found in 6 Chinese rhesus macaques using our combined Roche/454 and Illumina sequencing approach.

Known <i>Mamu-A</i> alleles	Known <i>Mamu-B</i> alleles	Extensions of <i>Mamu-A</i> alleles	Extensions of <i>Mamu-B</i> alleles
<i>Mamu-A1*026:03</i>	<i>Mamu-B*007:03</i>	<i>Mamu-A1*004:01:02</i>	<i>Mamu-B*013:04</i>
<i>Mamu-A1*056:02:01</i>	<i>Mamu-B*010:01</i>	<i>Mamu-A1*018:02</i>	<i>Mamu-B*014:01</i>
<i>Mamu-A2*01:02</i>	<i>Mamu-B*015:02</i>	<i>Mamu-A2*05:02:01</i>	<i>Mamu-B*016:01</i>
<i>Mamu-A4*14:03:01</i>	<i>Mamu-B*036:02</i>	<i>Mamu-A5*30:03</i>	<i>Mamu-B*023:01</i>
	<i>Mamu-B*037:01</i>		<i>Mamu-B*035:01</i>
	<i>Mamu-B*039:01</i>		<i>Mamu-B*044:03</i>
	<i>Mamu-B*045:02</i>		<i>Mamu-B*056:03</i>
	<i>Mamu-B*056:01</i>		<i>Mamu-B*062:01</i>
	<i>Mamu-B*060:04</i>		<i>Mamu-B*065:03:01</i>
	<i>Mamu-B*065:02</i>		<i>Mamu-B*067:01</i>
	<i>Mamu-B*066:01</i>		<i>Mamu-B*068:01:01</i>
	<i>Mamu-B*068:02</i>		<i>Mamu-B*072:01:01</i>
	<i>Mamu-B*085:02</i>		<i>Mamu-B*082:06</i>
	<i>Mamu-B*087:01</i>		
	<i>Mamu-B*116:01</i>		

^a(Robinson et al. 2013)

^b(Robinson et al. 2000)

Table 3

Novel transcripts found in 6 Chinese rhesus macaques using our combined Roche/454 and Illumina sequencing approach.

Novel Mamu-A transcripts	Novel Mamu-B transcripts
<i>Mamu-A1*010:03</i>	<i>Mamu-B*019:07</i>
<i>Mamu-A1*019:09</i>	<i>Mamu-B*045:08</i>
<i>Mamu-A1*090:02</i>	<i>Mamu-B*045:09</i>
<i>Mamu-A1*090:03</i>	<i>Mamu-B*050:03</i>
<i>Mamu-A1*092:03</i>	<i>Mamu-B*051:09</i>
<i>Mamu-A4*14:13</i>	<i>Mamu-B*056:05</i>
<i>Mamu-A4*14:03:03</i>	<i>Mamu-B*060:05</i>
<i>Mamu-A6*01:06</i>	<i>Mamu-B*060:01:02</i>
	<i>Mamu-B*067:01:02</i>
	<i>Mamu-B*074:03</i>
	<i>Mamu-B*074:03 (splice variant)</i>
	<i>Mamu-B*082:07</i>
	<i>Mamu-B*088:01:02</i>
	<i>Mamu-B*098:05</i>
	<i>Mamu-B*098:06</i>
	<i>Mamu-B*193:01^a</i>
	<i>Mamu-B*002:03</i>
	<i>Mamu-B*142:02N</i>
	<i>Mamu-B*150:02</i>
	<i>Mamu-B*151:01</i>
	<i>Mamu-B*167:01</i>

^a New lineage.