Immunogenetics: Open Access

Caskey et al., Immunogenet Open Access 2019, 4:1

Research Article Open Access

MHC Genotyping from Rhesus Macaque Exome Sequences

John R Caskey[1], Roger W Wiseman[1,2], Julie A Karl[2], David A Baker[2], Taylor Lee[2], Robert J Maddox[1], Muthuswamy Raveendran[3], R Alan Harris[3], Jianhong Hu[3], Donna M Muzny[3], Jeffrey Rogers[3] and David H O'Connor[1,2*]

[1]*Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, WI 53715, USA*

[2]*Department of Pathology and Laboratory Medicine, University of Wisconsin-Madison, Madison, WI 53705, USA*

[3]*Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA*

[*]**Corresponding author:** Dr David H O'Connor, Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, WI 53715, USA, Tel: +1 608 890 0845; E-mail: doconnor@primate.wisc.edu

Received date: June 27, 2019; **Accepted date:** July 15, 2019; **Published date:** July 29, 2019

Abstract

Indian rhesus macaque major histocompatibility complex (MHC)-variation can influence the outcomes of transplantation and infectious disease studies. Frequently, rhesus macaques are MHC genotyped to identify variants that could account for unexpected results. Since the MHC is only one region in the genome where variation could impact experimental outcomes, strategies for simultaneously profiling variation in the macaque MHC and the remainder of the protein coding genome would be useful. Here we determine MHC class I and class II genotypes using target-capture probes enriched for MHC sequences, a method we term macaque exome sequence (MES) genotyping. For a cohort of 27 Indian rhesus macaques, we describe two methods for obtaining MHC genotypes from MES data and demonstrate that the MHC class I and class II genotyping results obtained with these methods are 98.1% and 98.7% concordant, respectively, with expected MHC genotypes. In contrast, conventional MHC genotyping results obtained by deep sequencing of short multiplex PCR amplicons were only 92.6% concordant with expectations for this cohort.

Keywords: Macaca mulatta; Exome; Major histocompatibility complex

Introduction

The major histocompatibility complex (MHC) is an intensively studied set of genes in macaques [1,2]. The genomic MHC region contains clusters of genes that encode the MHC class I complex and the MHC class II complex. Cells use MHC class I molecules to present intracellular peptides to immune cells like the CD8+ T cell or natural killer cells [3]. MHC class I molecules accommodate intracellular peptides of varying specificity by having diverse amino acid sequences in the α1 and α2 subunits, which form the peptide-binding cleft [4,5]. These α1 and α2 subunits correspond to exons 2 and 3 respectively, of an MHC class I gene [6]. Most of the polymorphisms that distinguish MHC class I alleles are concentrated in exons 2 and 3 [7]. Thousands of individual MHC allelic variants have been identified in the three most widely used macaque species for biomedical research: rhesus (*Macaca mulatta*), cynomolgus (*Macaca fascicularis*), and pig-tailed macaques (*Macaca nemestrina*) [8-10].

The human MHC is termed the human leukocyte antigen complex (HLA). HLA class I has a single copy of the *HLA-A, HLA-B*, and *HLA-C* genes on each chromosome [1,11]. In contrast, macaques have a variable number of genes on each chromosome that encode MHC class I MHC-A and MHC-B proteins, and macaques lack an *HLA-C* orthologue. Initial approaches to genotype macaque MHC class I relied on using sequence-specific PCR oligonucleotides to test for the presence or absence of individual alleles [12]. More recently, deep sequencing genomic DNA or complementary DNA PCR amplicons spanning a highly variable region of exon 2 has become commonplace [13,14].

Amplicon sequences can be used to genotype groups of closely related MHC class I alleles, which are denoted as lineages. For example, an amplicon deep sequence that corresponds to the rhesus macaque *Mamu-A1*001* lineage demonstrates that an animal possesses *Mamu-A1*001:01*, *Mamu-A1*001:02*, or another closely related variant that has not yet been identified. This lineage-level reporting of MHC class I genotypes can be sufficient for designing experiments where specific MHC class I genotypes need to be matched between animals, balanced among experimental groups, or excluded entirely from a study [5,13-18]. Amplicon deep sequencing can also be used for MHC class II genotyping. The human HLA class II genes DQA1, DQB1, DPA1, and DPB1 have direct orthologues in macaques [19]. Both macaques and humans have a variable number of MHC class II DRB genes on a single chromosome, while the DRA gene is oligomorphic [1,11] and typically is not used for genotyping purposes. MHC class II molecules are members of the immunoglobulin superfamily, but they differ from MHC class I in several ways. A and β subunits comprise the MHC class II heterodimer, with separate genes encoding α and β subunits [20]. Highly polymorphic regions that are diagnostic for MHC class II allele lineages can be PCR amplified in a manner that is similar to the MHC class I genotyping. Exon 2 contains the most extensive polymorphism among MHC class II alleles [7]. Each of the *DRB, DQA1, DQB1, DPA1,* and *DPB1* polymorphic MHC class II genes are sufficiently divergent that separate PCR amplicons are required for deep sequencing [21].

Comprehensive MHC class I and class II genotyping of macaques by amplicon deep sequencing requires preparation of six separate PCR amplicons for MHC class I and class II DRB, DQA1, DQB1, DPA1, and DPB1 [9,21]. Despite this complexity, a major advantage to amplicon deep sequencing has been its cost effectiveness. The output from a single MiSeq sequencing run can be used to determine the MHC class I and MHC class II genotypes for up to 192 macaque

samples. In recent years, improved sequencing hardware and software have prompted new approaches to MHC genotyping. Instead of utilizing PCR amplicons of variable gene regions to genotype samples, researchers can use human whole exome sequencing (WES) and whole genome sequencing (WGS) datasets to determine HLA genotypes [22-25]. Likewise, target capture approaches have been described for HLA genotyping with next-generation sequencing datasets [26,27]. In contrast, MHC genotyping of macaques using WGS [28-30] or WES [31,32] have not been reported to date. The macaque genome is being actively updated, but the reference database of known alleles is still incomplete [10,33]. The unfinished state of the macaque reference database and the inherent intricacies that exist when attempting to map short sequence reads against complex, duplicated gene families both illustrate the significant challenges researchers can face with macaque genotyping.

In an era where the per-base cost of sequencing is dropping rapidly, we explored the feasibility of obtaining whole exome sequencing data while maintaining parity of results with the traditional MHC PCR amplicon approach. WES datasets provide investigators with the capability to examine sequence variants across the entire coding region of the genome. Although SIV researchers have traditionally focused on specific MHC genotypes that are associated with exceptional control of viral replication, there is also interest in evaluating variation in host restriction factor genes, such as *TRIM5*, *Tetherin*, and *APOBEC3G* receptor genes [34-37]. Exome sequencing will also provide an opportunity for investigators in multiple fields to evaluate variants from a wide variety of pathways that are important for biomedical research [38]. A key advantage of WES over WGS remains to be the cost per sample of generating sequence assays. Since WES only targets the coding regions, which are 1-2% of the genome, the amount of sequencing that is required is dramatically reduced. At our institutions, currently WES can be completed for less than 20% of the cost of a comparable WGS study. The ability to evaluate at least five times more animals for the same cost opens the possibility of characterizing much larger non-human primate cohorts, such as with breeding colonies of macaques, or complete retrospective experimental studies [30].

Here we introduce MHC genotyping via macaque exome sequencing (MES), which is an exome sequencing-based workflow for comprehensive MHC class I and class II genotyping in macaques. This workflow uses a commercially available human exome target-capture enrichment kit in conjunction with specialized spike-in target-capture probes to cope with the high copy number of macaque MHC genes. We show that accuracy of Indian rhesus macaque MHC class I and class II results from this workflow are comparable to conventional MiSeq genotyping when using exon 2 reference sequences.

Methods

Animals

Twenty-seven whole blood samples were collected from Indian rhesus macaques (*Macaca mulatta*). Five of these samples came from a breeding group of animals living at the Wisconsin National Primate Research Center (WNPRC). The remaining 22 samples were provided by Dr. Michele Di Mascio from the National Institutes of Health's National Institutes of Allergy and Infectious Diseases. Blood sampling was performed under anesthesia and in accordance with the regulations and guidelines outlined in the Animal Welfare Act, the Guide for the Care and Use of Laboratory Animals, and the Weatherall report [39,40].

Data

Exome sequence datasets have been deposited in the sequence read archive (SRA) under BioProjects PRJNA527214 and PRJNA529708. Fasta reference sequences used for MHC genotyping, and sequence analysis scripts are available from https://go.wisc.edu/jb0926.

MHC class I and class II genotyping by amplicon deep sequencing

Genomic DNA was isolated from 250 μL of whole blood using a Maxwell® 48 LEV Blood DNA Kit (Promega Corporation, Fitchburg, WI). Following isolation, DNA concentrations were determined with a Nanodrop 2000 and samples were normalized to 60 ng/ul. MHC class I and class II PCR amplicons were generated using exon 2-specific primers with adapters (CS1 and CS2) necessary for 4-primer amplicon tagging with the Fluidigm Access Array™ System (Fluidigm, San Francisco, CA, USA) by previously described methods [9,21]. Pooled PCR products were purified using the AMPure XP beads (Agencourt Bioscience Corporation, Beverly, MA, USA) and quantified using the Quant-iT dsDNA HS Assay kit with a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA), following the manufacturer's protocols. The MHC exon 2 genotyping amplicon pools were sequenced on an Illumina MiSeq instrument (San Diego, CA, USA) as previously described [9]. Analysis of the MiSeq exon 2 genotyping amplicon sequences was performed using a custom Python workflow. The workflow contained a step to remove oligonucleotide primers and sequencing adapters with bbduk, a step to merge reads using bbmerge, a step to identify unique sequences/remove chimeras with USEARCH, and finally mapping unique reads against a deduplicated reference database of rhesus macaque partial MHC class I and class II exon 2 sequences with bbmap [41,42]. The IPD-MHC NHP database on the European Bioinformatics Institute website is continuously updated as new information from researchers is submitted to it, and we elected to download the release that was available on September 8, 2018 [10]. By doing so, we created a fixed snapshot of the reference sequences for our data analyses, and this ensured consistency for us when we tested our experimental hypotheses. For this publication, we define "IPD exon 2" as this database of reference sequences. SAM output files from bbmap were parsed with the Python package pandas to enumerate the reads from each animal that were identical to IPD exon 2 reference sequences. Mamu-A, -B, -DRB, -DQA1, -DQB1, -DPA1 and -DPB1 lineage-level haplotypes were inferred for each of the samples with a semi-automated custom workflow that identifies diagnostic alleles associated with previously defined rhesus macaque haplotypes [14,19].

MHC class I and class II genotyping by exome sequencing

Genomic DNA was isolated as described above and shipped to the Human Genome Sequencing Center at the Baylor College of Medicine. MHC and exon-containing genomic DNA fragments were selectively enriched using a custom target-capture probeset. Genome-wide exons were captured using SeqCap EZ HGSC VCRome2.1, an optimized human clinical exome probeset [43,44]. SeqCap EZ HGSC VCRome2.1 contains probes designed to enrich 23,585 human genes and 189,028 non-overlapping exons. A low coverage audit was performed to identify rhesus macaque exons inferred from the reference genome rheMac2 that were not sufficiently enriched (<20x coverage) with these human probes, and an additional 22,884 rhesus macaque exons were incorporated into the genotyping probe design [45]. Finally, and most importantly for MHC analyses, we modified the SeqCap EZ Design: Human MHC Design to selectively enrich MHC class I and class II

sequences. This previous design was prepared by the Beijing Genome Institute in collaboration with Roche/Nimblegen and it targeted the complete 4.97 Mb HLA region with non-redundant probes designed against 8 fully sequenced HLA haplotypes [26,46]. For our macaque studies, we prepared a minimal MHC target capture design using a subset of these probes that are based on all functional HLA class I (*HLA-A, -B, -C,* and *-E*) and class II (*HLA-DRA, -DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1* and *-DPB1*) genes. Probes were included to capture complete gene sequences (exons+introns+3'UTR) as well as ~1 kb of 5' upstream flanking sequence. The BED file of rhesus rheMac2 target coordinates lifted over to rheMac8 was used to prepare this combined minimal MHC and supplemental rhesus spike-in probe design. Because derivation of MHC results from genotyping is paramount, we used a ratio of 2.5x spike-in probes to 1x VCRom2.1 probes. The supplemental probes for MHC and rhesus macaque were a single reagent, and the MHC-specific probes only constituted 609 kb of the 37.9 Mb probes.

An Illumina paired-end pre-capture library was constructed with 750 nanograms of DNA, as described by the Baylor College of Medicine Human Genome Sequencing Center. Pre-capture libraries were pooled into 10-plex library pools for target capture according to the manufacturer's protocol. Samples were pooled in 10-plex sequence capture library pools for 151 bp paired-end sequencing in a single lane of an S4 flow cell on an Illumina NovaSeq 6000 at the Baylor College of Medicine Human Genome Sequencing Center.

Enumeration of MHC reads in exome data

The effective enrichment of MHC reads in the exome datasets was calculated by mapping the reads of each sample's exome dataset against an individual genomic reference file for each individual locus: *HLA-A* exons 2-3 (NCBI Gene ID: 3105) and HLA-E exons 2-3 (Gene ID: 3133); *HLA-DPA1* exons 2-4 (Gene ID: 3113); *HLA-DPB1* exons 2-4 (Gene ID: 3115); *HLA-DQA1* exons 2-4 (Gene ID: 3117); *HLA-DQA2* exons 2-4 (Gene ID: 3118); *HLA-DQB1* exons 2-4 (Gene ID: 3119); *HLA-DQB2* exons 2-4 (Gene ID: 3120); *HLA-DRB1* exons 2-4 (Gene ID: 3123), *HLA- DRB3* exons 2-4 (Gene ID: 3125), *HLA-DRB4* exons 2-4 (Gene ID: 3126) and *HLA-DRB5* exons 2-4 (Gene ID: 3127). This mapping was done by using bbmap with default parameters, which corresponds to a minimum alignment identity of approximately 76% [41]. Empirically, these mapping parameters are sufficient to map macaque MHC reads to their human orthologues. Mapped reads were written to a new fastq file using bbmap's outm=parameter. To quantify the total number of reads in a sample and the number of reads extracted with our reference file,we created a custom Python script, which is available to download.

MHC genotyping from exome data

Two complementary data analysis strategies were employed to analyze the exome sequence data, and to verify reproducibility and confidence in the MHC genotyping results. For accuracy and quantification purposes, the expected MHC genotypes for each animal were established based on concordance among at least two out of the three described strategies and biological plausibility, e.g., no more than two alleles per Mamu-*DQA1, -DQB1, - DPA1*, or-*DPB1* locus.

Strategy 1: MHC genotyping using Diagnostic Sub-Region (DSR): The Diagnostic Sub-Region (DSR) was an intra-allelic region that encompassed polymorphisms, and these polymorphisms were distinguishable from alleles with similar sequences. Therefore, this method ensured the DSR was captured in at least one read for each called allele. MHC class I and class II reads initially were extracted from each animal by mapping the FASTQ reads to HLA class I and class II reference sequences containing exons 2-3 and exons 2-4, respectively, plus the intervening intron(s) as described above in 'Enumeration of MHC reads in exome data'. Reads were mapped to these reference sequences using bbmap with default parameters and the parameter (qtrim=lr).

Following extraction of MHC reads from the total exome sequences, the MHC reads were prepared for assembly using a modified version of a data pre-processing pipeline, which included tools from the BBTools package. Briefly, optical duplicates and reads from low-quality regions of the sequencing run were removed. Next, Illumina sequencing adapters were trimmed from the ends of sequencing reads. Any residual spike-in or PhiX sequences that inadvertently survived mapping to HLA class I and class II were then removed. Three rounds of error-correction and read merging were performed to create high-confidence merged reads that were well- supported by common kmers found in the extracted MHC reads. The error-corrected reads were not merged with a minimum overlap, but instead were separately mapped against the IPD exon 2 sequences using bbmapskimmer. The default settings for the software tool bbmapskimmer were used with the following modified parameters (semiperfectmode=t ambiguous=all ssa=t maxsites=50000 maxsites2=50000 expectedsites=50000). The semiperfectmode setting accepted reads with perfect matches, as well as reads that extended off the end of contigs for no more than half of the length of the mapped read segment. The ambiguous setting and the 'expectedsites' setting reported the first 50,000 matched read segments that met the 'semiperfectmode' filtering. These settings were set exceedingly high in order to exhaustively map the reads to our IPD exon 2 reference file of approximately 874 alleles. Using semiperfect mode, these were the segment sequences with the longest matching length. This output file included all mapped reads to all IPD exon 2 reference allele sequences.

We then used samtools mpileup with the settings (-A -a --ff UNMAP -x -B -q 0 -Q 0) on the bbmapskimmer output to calculate a depth of coverage at each position for each IPD exon 2 sequence. Next, we removed any aligned reads to the IPD exon 2 database sequences that contained less than a minimum depth of coverage of two across the entire reference sequence. For ambiguously-mapped reads, alignments with the longest-matching region were selected for further analysis; ties among alignments were counted multiple times. To reduce the number of false positives, we used Python pandas to only report database sequence matches that had at least one unambiguously mapped read. Based on these mapping parameters, unambiguously-mapped reads must span the DSR. The MHC genotypes from this method were reported as the minimum depth of coverage for each IPD exon 2 database sequence per animal.

Strategy 2: *De novo* **reconstruction of MHC sequences from exome reads:** Most *de novo* sequence assemblers have been optimized for resolving long contigs, and are tolerant of small sequence mismatches that can otherwise fragment assemblies. In the case of MHC alleles, however, such closely related sequences were often biologically distinct. Because of the gene duplications in the macaque MHC, there were many valid MHC contigs that could be assembled from a single sample within the exome data. This is conceptually similar to the computational challenge of assembling viral haplotypes, where rapidly evolving viruses such as human immunodeficiency virus accumulate variants that frequently co-segregate as minor populations within an

infected person [47]. Therefore, we utilized the overlap assembly algorithm SAVAGE, originally designed to reconstruct viral haplotypes, to reconstruct MHC allele sequences from exome reads. Similar to Strategy 1 above, HLA-mapped reads were pre-processed using the BBTools package to remove low quality reads and optical duplicates. Next, adapters were trimmed, residual spike-in and PhiX sequences were removed, three steps of error-correction took place, and reads were merged. These merged reads, as well as high-confidence unmerged paired-end reads that could not be grouped into an overlapping merged read, were used for SAVAGE assembly. The following workflow was implemented in a reproducible snakemake workflow that fully documents parameter selection [48] and is available upon request.

SAVAGE is designed to construct individual haplotypes from the overlap graph of individual reads. We processed the totality of sequence data in a single patch in order to maximize the sensitivity of the contiguous reconstruction with parameter '--split 1', as well as the parameter '--revcomp' to handle reverse complement reads. The IPD exon 2 reference database that was used for MiSeq and DSR genotyping was also used to assess the quality of SAVAGE genotypes. The IPD sequences were mapped to contigs produced by SAVAGE using sensitive parameters in bbmap (minlength=100 vslow=t subfilter=0 indelfilter=0 lengthtag=t ignorefrequentkmers=t kfilter=100) designed to identify sequences that perfectly match SAVAGE contigs. A post-processing script refined these mappings, and only retained the mappings where the length of the mapped region was the same as the length of the IPD exon 2 sequence. These mappings indicated where the reference database sequence was fully and exactly contained within a SAVAGE contig.

Results

MHC reads are efficiently enriched using target-capture probes

We designed a custom target enrichment probeset that accounts for the extensive duplication of macaque MHC genes [45]. A tripartite target capture system was used in this study. The first component is SeqCap EZ HGSC VCRome2.1, an optimized human clinical exome probeset. Since macaques are closely related to humans, SeqCap EZ HGSC VCRome 2.1 can also be used in macaques, though some sequences that are most divergent between macaques and humans are not efficiently captured. The second component of the target capture system is an additional 22,884 rhesus macaque exon sequences that were not effectively captured with the SeqCap EZ HGSCVCRome2.1 reagent. The third component is a collection of probes designed to specifically enrich macaque MHC class I and class II sequences. These probes span the full length of HLA class I and class II genes including introns, 3' UTRs, and approximately 1,000 bp of flanking 5' sequence.

We obtained a median coverage of 100X for the target exon sequences across the genome with >20X coverage for 94.97% of bases that were targeted in this study. As illustrated in Table 1, an average of 70,549,789 Illumina sequence reads per sample were obtained for the 27 animals evaluated in this study. These reads were mapped against reference files containing representative genomic HLA exons 2 -3 for HLA-A and *HLA-E*, and exons 2-4 for *HLA-DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1* and -DPB1 sequences. We identified an average of 269,057 MHC class I and class II sequence reads per sample which corresponds to an average of 0.37% of the total sequence reads evaluated per sample (Table 1). In a previous study by Ericsen et al. [37] we found that MHC sequences only accounted for an average of 0.13% of the total Illumina sequence reads that were evaluated per animal when the standard HGSC VCRome2.1 panel was used alone for target capture (Supplementary Table 1). Thus, we achieved an almost three-fold increase of MHC genomic sequences after inclusion of the spike-in probes for target capture compared to use of the VCRome2.1 probeset alone.

Animal ID	SRA Accession	Total Raw Reads	MHC-1 Reads Extracted	DRB1/3/4/5 Reads Extracted	DQA1 Reads Extracted	DQB1 Reads Extracted	DPA1 Reads Extracted	DPB1 Reads Extracted	MHC Reads Extracted	HC % of total
r05029	SAMN11131055	8,73,45,430	79,712	2,15,654	24,266	39,188	29,604	1,12,376	5,00,800	0.57
r17099	SAMN11131056	6,76,32,236	32,406	89,946	9,694	16,782	11,214	47,974	2,08,016	0.31
r07010	SAMN11173514	8,10,62,856	71,694	2,04,352	22,528	38,382	25,392	1,11,626	4,73,974	0.58
r17041	SAMN11131058	8,00,04,430	30,852	2,44,326	9,746	16,310	12,644	1,51,376	4,65,254	0.58
r17061	SAMN11131059	7,84,28,820	70,082	2,04,694	20,776	41,018	19,440	95,538	4,51,548	0.58
J8R	SAMN11282383	8,08,68,782	45,164	87,864	11,878	19,736	13,810	57,146	2,35,598	0.29
J01	SAMN11282382	6,10,09,650	34,342	87,124	9,050	16,628	10,958	46,964	2,05,066	0.34
ZC08	SAMN11282384	6,06,93,394	31,168	81,586	10,392	15,660	10,172	45,748	1,94,726	0.32
DGKG	SAMN11282378	8,05,30,772	48,658	1,30,256	12,788	22,682	13,372	62,990	2,90,746	0.36
CF18	SAMN11282368	6,85,93,710	37,792	96,842	10,864	18,226	13,118	51,556	2,28,398	0.33
DE1AA	SAMN11282369	6,57,12,570	39,334	1,08,002	10,060	17,124	11,840	51,134	2,37,494	0.36

DEG8	SAMN11282370	6,84,90,344	35,566	1,16,824	10,404	19,278	12,294	55,086	2,49,452	0.36
DEXX	SAMN11282371	6,74,51,626	33,900	1,28,128	10,442	18,898	9,778	53,258	2,54,404	0.38
DF24	SAMN11282372	6,77,68,724	41,362	96,172	10,606	18,176	10,976	50,612	2,27,904	0.34
DF64	SAMN11282373	6,57,62,818	37,426	1,07,974	11,222	15,632	11,446	51,214	2,34,914	0.36
DF6T	SAMN11282374	7,14,72,222	41,160	1,01,102	11,052	18,794	13,492	51,626	2,37,226	0.33
DFET	SAMN11282375	7,70,40,660	49,068	1,28,862	12,588	20,952	12,586	58,238	2,82,294	0.37
DFV0	SAMN11282377	7,38,77,184	41,924	98,866	11,630	17,690	13,952	52,394	2,36,456	0.32
DJ94	SAMN11282379	6,60,51,412	39,616	88,196	9,910	15,892	10,978	45,972	2,10,564	0.32
HIH	SAMN11282380	7,64,91,578	41,308	85,732	11,148	18,366	14,102	50,746	2,21,402	0.29
DFJT	SAMN11282376	6,22,49,064	37,728	78,860	9,666	15,268	11,426	43,086	1,96,034	0.31
HLJ	SAMN11282381	6,82,88,558	41,748	94,570	9,960	17,240	12,756	48,356	2,24,630	0.33
37360	SAMN11282363	6,23,10,824	38,200	97,452	9,866	15,024	10,320	46,140	2,17,002	0.35
0L7	SAMN11282364	6,41,45,834	35,302	1,28,778	9,310	19,992	10,686	54,196	2,58,264	0.4
0R1	SAMN11282365	6,47,89,208	36,462	86,772	10,656	17,258	11,982	46,124	2,09,254	0.32
0R8	SAMN11282366	7,28,36,490	38,286	1,18,184	10,500	22,248	13,170	56,142	2,58,530	0.35
0TI	SAMN11282367	8,07,30,750	41,552	1,06,356	13,364	20,714	14,860	57,744	2,54,590	0.32
	Average	7,05,49,789	42,660	1,19,018	12,014	20,487	13,569	61,310	2,69,057	0.37

Table 1: Fraction of total exome sequence reads corresponding to exons 2-3 of MHC class I and exons 2-4 of MHC class II genes.

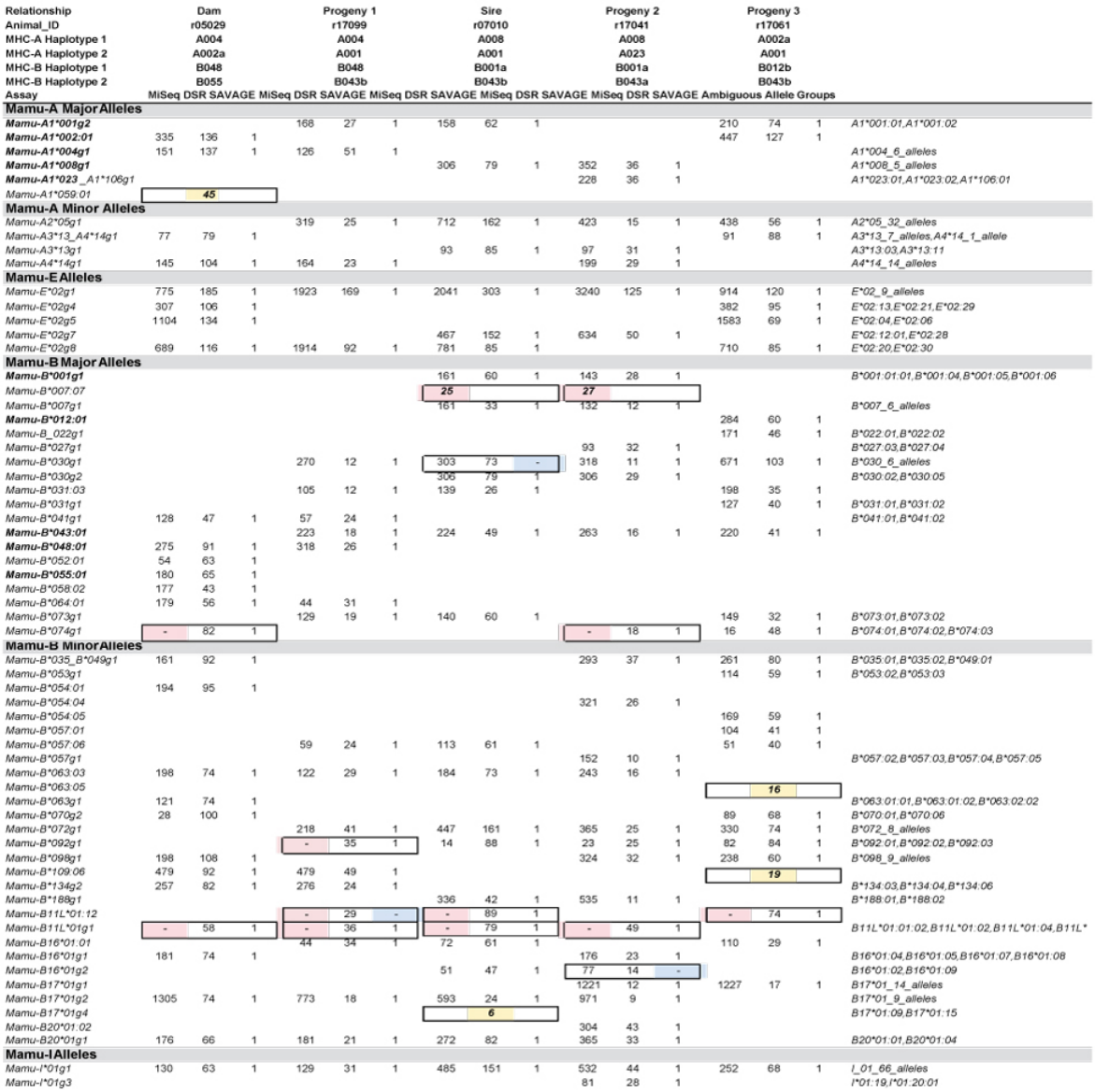
MHC genotypes defined from target-enriched genomic DNA are comparable to those obtained by amplicon deep sequencing

We hypothesized that MHC genotypes derived from target-enriched genomic sequence would be comparable in accuracy to MHC genotypes derived from conventional amplicon deep sequencing. In order to compare the accuracy of MHC genotyping results between the methods described here, the expected genotypes for each animal were defined by concordance among at least two out of the three described strategies, as well as biological plausibility. For example, no more than two alleles should be present for each of the MHC class II Mamu-*DQA1, -DQB1, -DPA1,* or *-DPB1* loci. MHC class I and class II PCR amplicons were deep sequenced on an Illumina MiSeq from the same 27 animals that were evaluated by MES analysis. Representative MHC class I and class II genotypes supported by the amplicon data are shown in Figures 1 and 2 respectively.

These figures illustrate genotypes from a pedigreed family of Indian rhesus macaques: a sire, a dam, and three progeny that are paternal half-siblings. Although segregation data confirmed all expected genotypes that were shared among the directly related individuals in this breeding group, this was generally not possible, since pedigree information was unavailable for 22 of 27rhesus macaques examined in this study. Supplementary Figures 1 and 2 show comprehensive genotyping results for all 27 animals.

The superset of alleles supported by at least two of the methods described in this manuscript (MiSeq amplicon genotyping, DSR genotyping of individual exome sequencing reads, or SAVAGE genotyping of contigs derived from exome data), was considered to be the expected MHC genotypes for these animals. A complication for this assertion was specific genotypes were difficult for MiSeq analysis to report correctly. As we have shown previously, the number of reads supporting each genotyping call was highly variable, ranging from tens of reads to thousands of reads per allele.

A small subset of allelic variants contained nucleotide substitutions in their sequences within the binding sites of the PCR oligonucleotides that interfere with efficient PCR amplification. Members of the Mamu-B11L*01 allele lineage exemplify this issue; they have two nucleotide substitutions relative to the 5' oligonucleotide that was used to generate MHC class I amplicons. These Mamu-B11L*01 sequences were routinely absent in MiSeq genotypes (Figure 1). Likewise, the Mamu-DPA1*11:01 allele has four nucleotide substitutions relative to the oligonucleotide pair used to generate DPA1 amplicons for MiSeq genotyping (Figure 2).



Relationship	Dam			Progeny 1			Sire			Progeny 2			Progeny 3			
Animal_ID	r05029			r17099			r07010			r17041			r17061			
MHC-A Haplotype 1	A004			A004			A008			A008			A002a			
MHC-A Haplotype 2	A002a			A001			A001			A023			A001			
MHC-B Haplotype 1	B048			B048			B001a			B001a			B012b			
MHC-B Haplotype 2	B055			B043b			B043b			B043a			B043b			
Assay	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	Ambiguous Allele Groups
Mamu-A Major Alleles																
Mamu-A1*001g2				168	27	1	158	62	1				210	74	1	*A1*001:01,A1*001:02*
Mamu-A1*002:01	335	136	1										447	127	1	
Mamu-A1*004g1	151	137	1	126	51	1										*A1*004_6_alleles*
Mamu-A1*008g1							306	79	1	352	36	1				*A1*008_5_alleles*
Mamu-A1*023 _A1*106g1										228	36	1				*A1*023:01,A1*023:02,A1*106:01*
*Mamu-A1*059:01*		45														
Mamu-A Minor Alleles																
*Mamu-A2*05g1*				319	25	1	712	162	1	423	15	1	438	56	1	*A2*05_32_alleles*
*Mamu-A3*13_A4*14g1*	77	79	1										91	88	1	*A3*13_7_alleles,A4*14_1_allele*
*Mamu-A3*13g1*							93	85	1	97	31	1				*A3*13:03,A3*13:11*
*Mamu-A4*14g1*	145	104	1	164	23	1				199	29	1				*A4*14_14_alleles*
Mamu-E Alleles																
*Mamu-E*02g1*	775	185	1	1923	169	1	2041	303	1	3240	125	1	914	120	1	*E*02_9_alleles*
*Mamu-E*02g4*	307	106	1										382	95	1	*E*02:13,E*02:21,E*02:29*
*Mamu-E*02g5*	1104	134	1										1583	69	1	*E*02:04,E*02:06*
*Mamu-E*02g7*							467	152	1	634	50	1				*E*02:12:01,E*02:28*
*Mamu-E*02g8*	689	116	1	1914	92	1	781	85	1				710	85	1	*E*02:20,E*02:30*
Mamu-B Major Alleles																
Mamu-B*001g1							161	60	1	143	28	1				*B*001:01:01,B*001:04,B*001:05,B*001:06*
*Mamu-B*007:07*							25			27						
*Mamu-B*007g1*							161	33	1	132	12	1				*B*007_6_alleles*
Mamu-B*012:01													284	60	1	
Mamu-B_022g1													171	46	1	*B*022:01,B*022:02*
*Mamu-B*027g1*										93	32	1				*B*027:03,B*027:04*
*Mamu-B*030g1*				270	12	1	303	73	-	318	11	1	671	103	1	*B*030_6_alleles*
*Mamu-B*030g2*							306	79	1	306	29	1				*B*030:02,B*030:05*
*Mamu-B*031:03*				105	12	1	139	26	1				198	35	1	
*Mamu-B*031g1*													127	40	1	*B*031:01,B*031:02*
*Mamu-B*041g1*	128	47	1	57	24	1										*B*041:01,B*041:02*
Mamu-B*043:01				223	18	1	224	49	1	263	16	1	220	41	1	
Mamu-B*048:01	275	91	1	318	26	1										
*Mamu-B*052:01*	54	63	1													
Mamu-B*055:01	180	65	1													
*Mamu-B*058:02*	177	43	1													
*Mamu-B*064:01*	179	56	1	44	31	1										
*Mamu-B*073g1*				129	19	1	140	60	1				149	32	1	*B*073:01,B*073:02*
*Mamu-B*074g1*	-	82	1							-	18	1	16	48	1	*B*074:01,B*074:02,B*074:03*
Mamu-B Minor Alleles																
*Mamu-B*035_B*049g1*	161	92	1							293	37	1	261	80	1	*B*035:01,B*035:02,B*049:01*
*Mamu-B*053g1*													114	59	1	*B*053:02,B*053:03*
*Mamu-B*054:01*	194	95	1													
*Mamu-B*054:04*										321	26	1				
*Mamu-B*054:05*													169	59	1	
*Mamu-B*057:01*													104	41	1	
*Mamu-B*057:06*				59	24	1	113	61	1				51	40	1	
*Mamu-B*057g1*										152	10	1				*B*057:02,B*057:03,B*057:04,B*057:05*
*Mamu-B*063:03*	198	74	1	122	29	1	184	73	1	243	16	1				
*Mamu-B*063:05*														16		
*Mamu-B*063g1*	121	74	1													*B*063:01:01,B*063:01:02,B*063:02:02*
*Mamu-B*070g2*	28	100	1										89	68	1	*B*070:01,B*070:06*
*Mamu-B*072g1*				218	41	1	447	161	1	365	25	1	330	74	1	*B*072_8_alleles*
*Mamu-B*092g1*				-	35	1	14	88	1	23	25	1	82	84	1	*B*092:01,B*092:02,B*092:03*
*Mamu-B*098g1*	198	108	1							324	32	1	238	60	1	*B*098_9_alleles*
*Mamu-B*109:06*	479	92	1	479	49	1								19		
*Mamu-B*134g2*	257	82	1	276	24	1										*B*134:03,B*134:04,B*134:06*
*Mamu-B*188g1*							336	42	1	535	11	1				*B*188:01,B*188:02*
*Mamu-B11L*01:12*				-	29	-	-	89	1				-	74	1	
*Mamu-B11L*01g1*	-	58	1	-	36	1	-	79	1	-	49	1				*B11L*01:01:02,B11L*01:02,B11L*01:04,B11L**
*Mamu-B16*01:01*				44	34	1	72	61	1				110	29	1	
*Mamu-B16*01g1*	181	74	1							176	23	1				*B16*01:04,B16*01:05,B16*01:07,B16*01:08*
*Mamu-B16*01g2*							51	47	1	77	14	-				*B16*01:02,B16*01:09*
*Mamu-B17*01g1*										1221	12	1	1227	17	1	*B17*01_14_alleles*
*Mamu-B17*01g2*	1305	74	1	773	18	1	593	24	1	971	9	1				*B17*01_9_alleles*
*Mamu-B17*01g4*								6								*B17*01:09,B17*01:15*
*Mamu-B20*01:02*										304	43	1				
*Mamu-B20*01g1*	176	66	1	181	21	1	272	82	1	365	33	1				*B20*01:01,B20*01:04*
Mamu-I Alleles																
*Mamu-I*01g1*	130	63	1	129	31	1	485	151	1	532	44	1	252	68	1	*I_01_66_alleles*
*Mamu-I*01g3*										81	28	1				*I*01:19,I*01:20:01*

Figure 1: Comparison of MHC class I results from MiSeq PCR amplicon versus whole exome genotyping strategies for a representative breeding group of rhesus macaques. Results for each of the three methods are provided side-by- side in the columns for each macaque. Each row indicates the detection of a specific MHC class I allele or lineage group of closely related sequences that are ambiguous because they are identical over the IPD exon 2 database sequence. Values in the body of this figure indicate the number of sequence reads supporting each allele call for the MiSeq and DSR methods while alleles supported by a SAVAGE contig are reported with a "1". Discrepancies between the MiSeq (pink), DSR (yellow) or SAVAGE (blue) methods are highlighted by filled cells with borders.

These nucleotide substitutions do not fully account for all differences in the abundance of sequence reads for each allele. Allele lineages, such as *Mamu-B*074* and *Mamu-B*098*, exhibited significantly diminished PCR efficiency despite being perfectly matched with the oligonucleotides used for amplification (Figure 1). False positive genotyping calls were also noted in the MiSeq assay that resulted from intermolecular recombination during the PCR process [49-51].

Relationship	Dam			Progeny 1			Sire			Progeny 2			Frogeny 3			
Animal_ID	r05029			r17099			r07010			r17041			r17061			
MHC-DRB Haplotype 1	DR04a			DR04a			DR04a			DR04a			DR04a			
MHC-DRB Haplotype 2	DR06			DR03f			DR03f			DR06			DR03f			
MHC-DQA Haplotype 1	01g1			01g1			01g1			01g1			01g1			
MHC-DQA Haplotype 2	23_01			01_02			01_02			23_01			01_02			
MHC-DQB Haplotype 1	06_01			06_01			06_01			06_01			06_01			
MHC-DQB Haplodtype 2	18g4			06g2			06g2			18g4			06g2			
MHC-DPA Haplotype 1	02g1			02g1			02g1			02g1			11_01			
MHC-DPA Haplotype 2	02g1			04g			04g			02g1			04g			
MHC-DPB Haplotype 1	15g			15g			15g			15g			16_01			
MHC-DPB Haplotype 2	15g			02g			02g			15g			02g			
Assay	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	Ambiguous Allele Groups
Mamu-DRB Alleles																
*Mamu-DRB1*03:09*	198	370	1	72	131	1	170	305	1	272	107	1	156	200	1	
*Mamu-DRB1*03g1*				2789	217	1	2714	497	1				3052	499	1	*DRB1*03:06,DRB1*03:26*
*Mamu-DRB1*10:03*				1761	168	1	1952	358	1				2711	310	1	
*Mamu-DRB*W2g1*	154	451	1	72	183	1	132	410	1	206	176	1	107	398	1	*DRB*W2:01,DRB*W2:05*
*Mamu-DRB*W3g1*	1936	546	1							3097	217	1				*DRB*W3:03:01,DRB*W3.03:02*
*Mamu-DRB*W4:01*	1777	414	1							2631	158	1				
Mamu-DQA/DQB Alleles																
*Mamu-DQA1*01:02*				2384	98	1	2444	256	1				2964	154	1	
*Mamu-DQA1*01g1*	2298	318	1	2046	96	1	2276	288	1	4317	107	1	3360	210	1	*DQA1*01:04:01,DQA1*01:04:02*
*Mamu-DQA1*23:01*	2883	334	1							5079	108	1				
*Mamu-DQB1*06:01*	3803	160	1	3980	41	1	3892	152	1	5757	33	1	3892	122	1	
*Mamu-DQB1*06g2*				5229	173	1	4784	345	1				6400	282	1	*DQB1*06:05,DQB1*06:18*
*Mamu-DQB1*18g4*	2656	329	1							4504	108	1				*DQB1*18:02,DQB1*18:27*
Mamu-DPA/DPB Alleles																
*Mamu-DPA1*02g1*	5450	1765	1	5301	316	1	5153	840	1	8727	758	1				*DPA1*02_9_alleles*
*Mamu-DPA1*04g1*				435	181	1	489	385	1				1075	350	1	*DPA1*04:03:01,DPA1*04:03:02,DPA1*0*
*Mamu-DPA1*11:01*													-	225	1	
*Mamu-DPB1*02g1*				106	25	1	87	73	1				342	47	1	*DPB1*02:01,DPB1*02:04:01,DPB1*02:0*
*Mamu-DPB1*15g1*	254	217	1	286	31	1	207	129	1	3084	60	1				*DPB1*15:01,DPB1*15:02*
*Mamu-DPB1*16:01*													1769	98	1	

Figure 2: Comparison of MHC class II results from MiSeq PCR amplicon versus whole exome genotyping strategies for a representative breeding group of rhesus macaques. Results for each of the three methods are provided side-by-side in the columns for these five related macaques. Each row indicates detection of a specific MHC class II allele or lineage group of closely related sequences that are ambiguous because they are identical over the IPD exon 2 database sequence. Values in the body of this figure indicate the number of sequence reads supporting each allele call for the MiSeq and DSR methods, while alleles supported by a SAVAGE contig are reported with a "1". The Mamu-DPA1*11:01 allele missed by the MiSeq assay due to multiple mismatches versus the amplification primers is highlighted in pink.

Artifacts from PCR amplification, dubbed 'PCR chimeras', arise when an incompletely extended PCR product serve as a primer with a partially mismatched template during subsequent cycles of PCR (Supplementary Figure 3). This was exemplified by read support for the presence of *Mamu-B*007:07* in sire r07010, and progeny 2 r17041 (Figure 1). *Mamu- B*007:07* only differs from the *Mamu-B*007*g1 allele group that was determined to be present in this pair of animals by a single nucleotide variant at the extreme 5' end of the class I genotyping amplicon. Chimeric PCR products, equivalent to the *Mamu-B*007:07* sequence, were formed between the 3' portion of the *Mamu-B*007g1* sequence and other allelic variants in these animals with this 5' SNP such as *Mamu-B20*01g1*. These results illustrate that the MiSeq amplicon genotyping, while generally reflective of an animal's MHC genotype, can yield both false positive and false negative results. When compared to the expected genotypes for this dataset, MiSeq amplicon genotyping has 90.4% MHC class I and 97.8% MHC class II concordance (Table 2).

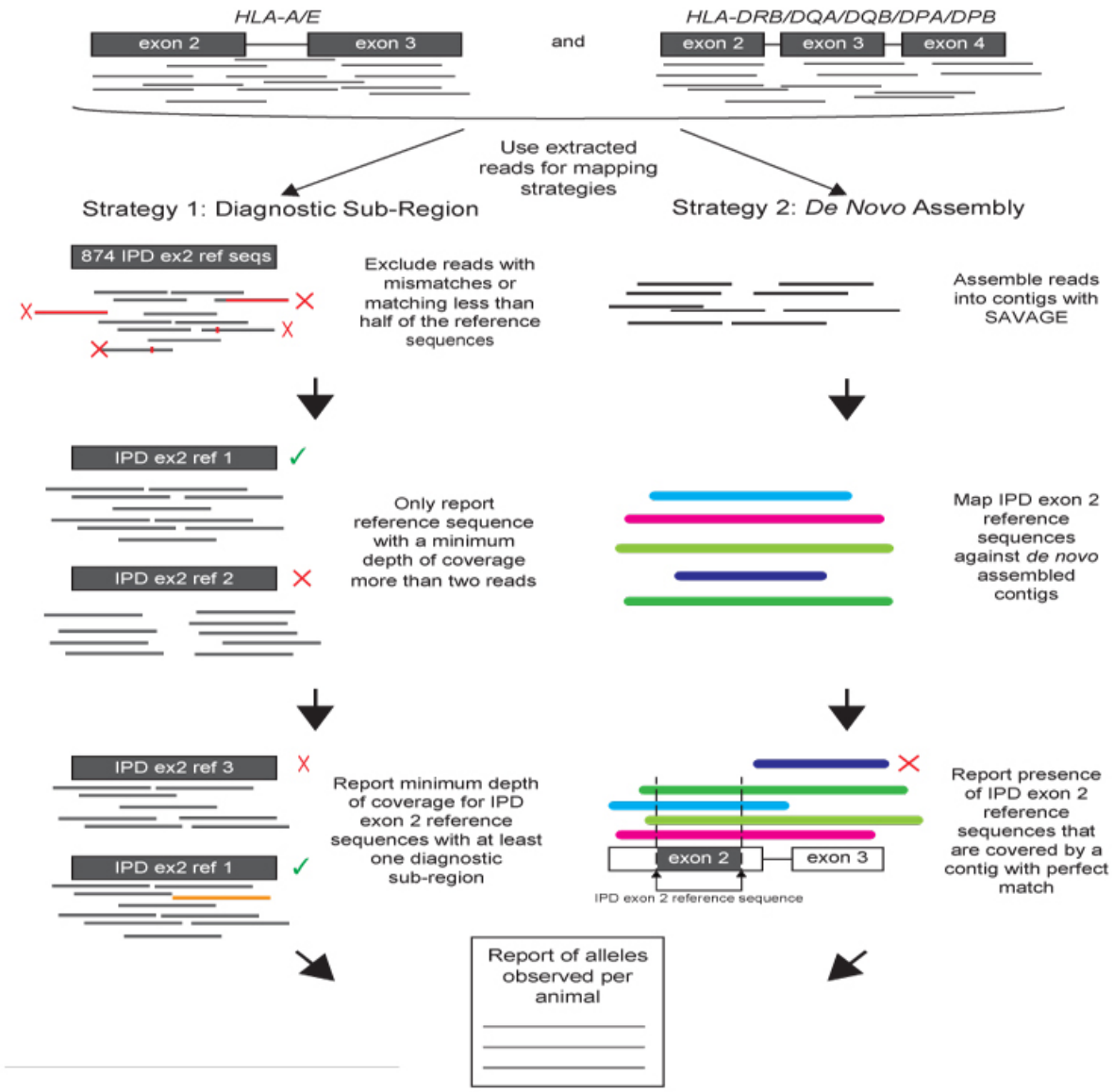


Figure 3: Description of strategies employed for MHC genotyping from macaque exome sequences. Exome sequencing reads are extracted by mapping to HLA-A/E exons 2-3 and HLA-DRB/DQA/DQB/DPA/DPB exons 2-4. These extracted sequences are then used for both analysis strategies. Strategy 1 maps the extracted reads to IPD exon 2 reference sequences. Only reads with perfect matches that extended off the end of contigs for no more than half of the length of the mapped read segment were retained. Next, each alignment of mapped sequences is checked to exclude those with a coverage that is less than 2 reads. The mapped alignments with this coverage are used to form a diagnostic sub-read, which is used to match alignments to the reference sequences. The minimum depth of coverage of the mapped alignments was reported in the output. Strategy 2 conducts a *de novo* assembly of the extracted reads using the computational tool SAVAGE. Then, the assembled contigs are mapped to IPD exon 2 reference sequences. To determine correct mapping, any IPD exon 2 reference sequences that fully cover and perfectly match a contig are reported in the output.

Two separate strategies were used to derive MHC genotypes from target-capture data using data from macaque exome sequencing. These strategies were outlined in Figure 3, and can be described as follows. The DSR analysis was a straightforward extension of the methodology used for deriving genotypes from MiSeq amplicons. MHC sequence reads were extracted from the total exome dataset. A simplified workflow that mapped those MHC reads to reference sequences, and then assigned the genotypes based on the presence of exome reads overlapping each position of the reference sequence, was problematic.

	MHC Class I				**MHC Class II**			
Animal ID	Alleles	Discordance	Discordance	Discordance	Alleles	Discordance	Discordance	Discordance
r05029	28	2	1		10			
r17099	25	3		1	12			
r07010	26	3	1	1	12			
r17041	31	3		1	10			
r17061	29	1	2		12	1		
J8R	29	4	1		12	2		
J01	30	3	1	1	11	1		
ZC08	24	2		1	12			
DGKG	28	4	1		13			
CF18	28	3			12			
DE1AA	30	5	1		13			
DEG8	27	3	1		13			
DEXX	24	3	1	1	14			
DF24	32	4		1	12			
DF64	30	3	1	1	13			
DF6T	31	2		1	10			
DFET	30	3	1	1	13	1		1
DFV0	28	2			6			
DJ94	34	3	2	1	12			
HIH	30	3			10			
DFJT	28	2	3		11			
HLJ	29	2	1	1	11			
37360	29	2	1	1	13	1		
0L7	20	2			14			
0R1	17	1			12			
0R8	25	2			13	1		
0TI	25	2	1		12			
Total	747	72	20	13	318	7	0	1
Discordance (%)		9.6	2.7	1.7		2.2	0	0.3
Concordance (%)		90.4	97.3	98.3		97.8	100	99.7

Table 2: Fraction of total exome sequence reads corresponding to exons 2-3 of MHC class I and exons 2-4 of MHC class II genes.

This simplified workflow was extremely vulnerable to false positive genotypes, and the similarities among different MHC sequences often enabled those reads to map to multiple different alleles. Two or more reads each could partially match a portion of a sequence in the IPD exon 2 database, which could complement each other to provide support for an allele that was not biologically relevant. As discussed in the Methods, the DSR encompassed polymorphisms that discriminated among closely related alleles by requiring at least one mapped read to unambiguously map to the corresponding allele within the IPD exon 2 sequences. This strategy can identify specific

polymorphisms of interest among the IPD exon 2 sequences to produce genotypes for MHC class I and MHC class II (Figures 1, 2 and Supplementary Figures 1 and 2). For the 27 animals evaluated in this study, DSR was 97.3% and 100% concordant with expected MHC class I and class II genotypes, respectively (Table 2).

Figure 4: Lengths of MHC sequence contigs assembled by the SAVAGE method for a representative breeding group of rhesus macaques. Values in the body of this figure indicate contig lengths generated for these MHC sequences in each animal. (A) MHC class I genotyping results are illustrated for the five animals in this breeding group. Three false negative allele calls for SAVAGE versus the expected genotypes for these animals are highlighted in magenta, e.g., Mamu-B*030g1 in sire r07010. Sequences highlighted in red are associated with the maternal MHC Haplotype that was inherited by progeny 1 from its dam. Sequences in progeny 1 and progeny 3 for their MHC Haplotype b that was inherited from sire r07010 are highlighted in dark blue while the light blue sequences represent the alternate paternal Haplotype c that was inherited by Progeny 2. Three unique extended MHC Haplotypes in this breeding group are indicated with shades of grey (Haplotypes d-f). MHC allele groups that are shared by both parental haplotypes are indicated by colored borders around filled cells. (B) MHC class II genotyping results are illustrated for this same breeding group.

The DSR strategy did not consider sequences that were not among the IPD exon 2 sequences, and this lack of consideration contributes to overcalled alleles. The apparent Mamu-A1*059:01 allele in dam r05029 was erroneously derived from reads that were from Mamu-A1*004g1 and an unknown allele that was not among the sequences in the IPD exon 2 database. As a result of this unknown sequence not being among the known IPD exon 2 sequences, reads unambiguously mapped to Mamu-A1*059:01, which caused this allele to be overcalled. This type of overcalling will be mitigated with the discovery of additional allelic variants and their inclusion in future iterations of the IPD database.

The second approach for determining genotypes performed *de novo* assembly on MHC class I and class II reads from each sample. The resulting assembled contigs were then mapped against IPD exon 2 reference sequences to define perfectly matching contigs. Because most assemblers were not tuned for the challenge of assembling large numbers of contigs that differ from one another by as little as 1 bp, we relied on an assembler, SAVAGE, originally designed to reconstruct viral sequencing haplotypes. As shown in Figures 1, 2 and Supplementary Figure 1, contigs produced by SAVAGE matched the expected genotypes. Unlike the DSR method, SAVAGE assembled sequence contigs that may be useful for downstream analyses such as higher resolution MHC genotyping. False positives and false negatives among closely-related variants were mitigated by the inclusion of the filtering steps described under Strategy 2. Across all 27 samples, the SAVAGE contigs are 98.3% and 99.7% accurate with respect to the expected MHC genotypes (Table 2).

While this analysis focused on genotyping using the same exon 2 reference sequences that are commonly utilized for MiSeq amplicon analyses, the SAVAGE contigs are frequently much longer than these reference sequences (Figure 4). These extended contigs frequently contain exons 2 through 4, plus the intervening introns, and could be used to provide higher resolution genotyping than is possible using exon 2 sequence alone. Moreover, contigs that contain complete sequences for exons 2-3 of MHC class I and exon 2 of MHC class II alleles meet the minimum criteria for obtaining formal allele nomenclature for non-human primates from the IPD-MHC [52].

Discussion

Here we describe MES genotyping of Indian rhesus macaques. In our estimation, this method will supersede MiSeq PCR amplicon genotyping as the most widely used macaque MHC genotyping assay in the future. This methodology has several compelling advantages. Most importantly, exome sequencing dramatically improves the overall quantity and quality of genomic information obtained from each sample. The same datasets used for MHC analyses may be used to evaluate protein-coding genetic variation throughout the genome. The loss of start and stop codons exome-wide can be obtained from the same datasets by modifying the workflows *in silico*, as opposed to requiring the sequencing of multiple new sets of target genes [44]. Exome-wide datasets offer the promise of retrospectively identifying candidate DNA sequence variants that may be responsible for unexpected experimental outcomes in studies with macaques and other nonhuman primates. The availability of exome-wide sequences in conjunction with MHC genotypes may also increase the rigor of prospective macaque experiments by enabling more sophisticated balancing of experimental groups, and exclusion of animals whose genetics are likely to strongly bias experimental results [15,34,38].

These same exome datasets also offer the potential to improve the quality of MHC genotyping. Full-length long-read MHC transcript sequencing offers the highest resolution, but this technology can be labor intensive and difficult to scale [8,9]. The MiSeq exon 2 amplicon approach is limited in its allelic resolution, but it is the current standard deep sequencing approach for high-throughput MHC genotyping in macaques. In this report, we compare two novel strategies for MHC genotyping from MES datasets to this standard MiSeq amplicon genotyping method. The results that we obtained with all three approaches show strong concordance with expected MHC genotypes. As illustrated in Figure 4a, MHC class I genomic contigs with an average length of approximately 1.1 kb can be assembled from sequence reads that were initially extracted from the whole exome datasets.

Following the initial enrichment step, MHC class II genomic contigs averaging approximately 1.9 kb in length could be assembled from whole exome sequence reads that mapped to *HLA-DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1* and *-DPB1* reference sequences, which contained exons 2 - 4, and the pairs of introns (Figure 4b). The addition of HLA spike-in probes containing both exon and intron sequences for the target capture step, and the introduction of 151 bp paired end reads for the Illumina NovaSeq platform greatly facilitated our ability to assemble these extended MHC genomic contigs. Additional technological advances, including even longer sequence reads and more efficient assembly algorithms, will undoubtedly increase genomic contig lengths as well as MHC allelic resolution in future studies.

Both MHC genotyping approaches described here depend upon mapping exome sequence reads against a reference database of MHC class I and class II allele sequences. Currently, the IPD database mostly is restricted to codin regions for macaque MHC sequences, and many IPD entries are only partial transcript sequences that lack complete coding regions. It is also very challenging to correctly phase short Illumina sequence reads. The reads map to different exons of a specific allelic variant, and are separated by intronic sequences that span hundreds to thousands of base pairs in genomic DNA. Our results with the SAVAGE workflow (Figure 4) demonstrate that exome sequence reads that have been enriched with the enhanced MHC probe design described here can be assembled into genomic contigs that span multiple exons and introns. These contigs, therefore, could also be used to improve MHC reference databases, though this requires a major effort with more animals, and is beyond the scope of this manuscript.

The current cost of genotyping is relatively high compared to amplicon deep sequencing, primarily due to two expenses. First, the amount of sequence data needed for a single sample can be high. Compared to conventional MHC genotyping, where 192 macaques' data can be collected on a single instrument run, exome data acquisition is much more expensive. On an Illumina NovaSeq instrument, which has a much higher run cost than the MiSeq, only 70 exome samples can be sequenced simultaneously per lane of a S4 flow cell. However, this cost of sequencing is rapidly decreasing, and as of early 2019, commercial providers have advertised sequencing for $9 USD per Gb of whole genome sequence data1. Second, a major expense in MHC genotyping is the production, validation, and use of target-capture arrays, as well as the development of *in silico* data analysis workflows. The approaches described here, in particular MHC genotyping from SAVAGE contigs produced by *de novo* assembly of exome reads, is flexible and should be adaptable as sequencing approaches evolve and improve.

A major goal for future studies will be to attempt to extend these contigs to encompass full length genomic MHC sequences using SAVAGE or other assembly software tools. The relatively compact genomic structure and consistent length of MHC class I genes increase the attainability of this goal. Establishment of comprehensive macaque MHC allele databases of extended genomic sequences will greatly facilitate mapping of exome sequence reads since they will be contiguous with the reference sequences instead of being interrupted by intervening sequences between each exon that are not included in current non-human primate IPD-MHC databases.

Conclusion

These results demonstrate that MHC genotypes can be obtained by analyzing genomic DNA selectively enriched for MHC and protein-coding gene sequences. This represents an important advance for characterizing MHC genetics in macaques, and this suggests that analyses of whole exome and whole genome data will become the predominant method for studying macaque genetics in the coming decade.

Acknowledgements

We gratefully acknowledge Michele Di Mascio and his group at the National Institute of Allergy and Infectious Diseases of the National Institutes of Health for providing rhesus macaque samples used in this study. We also gratefully thank Brian Bushnell for assistance with the BBTools software, and the WNPRC for providing samples from five related macaques.

This research was supported by contract HHSN272201600007C from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health. This work was also supported in part by the Office of Research Infrastructure Programs/OD (P51OD011106) awarded to the Wisconsin National Primate Research Center at the University of Wisconsin-Madison. This research was also supported in part by grant R24-OD011173 from the National Institutes of Health. This research was conducted in part at a facility constructed with support from Research Facilities Improvement Program grants RR15459-01 and RR020141-01.

References

1. Shiina T, Blancher A, Inoko H, Kulski JK (2017) Comparative genomics of the human, macaque and mouse major histocompatibility complex. Immunology 150: 127-138.
2. Wiseman RW, Karl JA, Bohn PS, Nimityongskul FA, Starrett GJ, et al. (2013) Haplessly hoping: Macaque major histocompatibility complex made easy. ILARJ 54: 196-210.
3. Garcia KC, Adams EJ (2005) How the T cell receptor sees antigen-A structural view. Cell 122: 333-336.
4. Silver ZA, Watkins DI (2017) The role of MHC class I gene products in SIV infection of macaques. Immunogenetics 69: 511-519.
5. Loffredo JT, Sidney J, Bean AT, Beal DR, Bardet W, et al. (2009) Two MHC class I molecules associated with elite control of immunodeficiency virus replication, Mamu-B*08 and HLA-B*2705, bind peptides with sequence similarity. J Immunol 182: 7763-7775.
6. Malissen M, Malissen B, Jordan BR (1982) Exon/intron organization and complete nucleotide sequence of an HLA gene. Proc Natl Acad Sci USA 79: 893-897.
7. Williams TM (2001) Human leukocyte antigen gene polymorphism and the histocompatibility laboratory. J Mol Diagn 3: 98-104.
8. Semler MR, Wiseman RW, Karl JA, Graham ME, Gieger SM, et al. (2018) Novel full-length major histocompatibility complex class I allele discovery and haplotype definition in pig-tailed macaques. Immunogenetics 70: 381-399.
9. Karl JA, Graham ME, Wiseman RW, Heimbruch KE, Gieger SM, et al. (2017) Major histocompatibility complex haplotyping and long-amplicon allele discovery in cynomolgus macaques from chinese breeding facilities. Immunogenetics 69: 211-229.
10. Maccari G, Robinson J, Ballingall K, Guethlein LA, Grimholt U, et al. (2017) IPD-MHC 2.0: An improved inter-species database for the study of the major histocompatibility complex. Nucleic Acids Res 45: D860-D864.
11. Daza-Vamenta R, Glusman G, Rowen L, Guthrie B, Geraghty DE (2004) Genetic divergence of the rhesus macaque major histocompatibility complex. Genome Res 14: 1501-1515.
12. Kaizu M, Borchardt GJ, Glidden CE, Fisk DL, Loffredo JT (2007) Molecular typing of major histocompatibility complex class I alleles in the Indian rhesus macaque which restrict SIV CD8+ T cell epitopes. Immunogenetics 59: 693-703.
13. Wiseman RW, Karl JA, Bimber BN, O'Leary CE, Lank SM, et al. (2009) Major histocompatibility complex genotyping with massively parallel pyrosequencing. Nat Med 15: 1322-1326.
14. Karl JA, Bohn PS, Wiseman RW, Nimityongskul FA, Lank SM, et al. (2013) Major histocompatibility complex class I haplotype diversity in chinese rhesus macaques. G3 (Bethesda) 3: 1195-1201.
15. Loffredo JT, Maxwell J, Qi Y, Glidden CE, Borchardt GJ, et al. (2007) Mamu-B*08-positive macaques control simian immunodeficiency virus replication. J Virol 81: 8827-8832.
16. Muhl T, Krawczak M, Ten Haaft P, Hunsmann G, Sauermann U (2002) MHC class I alleles influence set-point viral load and survival time in simian immunodeficiency virus-infected rhesus monkeys. J Immunol 169: 3438-3446.
17. Nomura T, Yamamoto H, Shiino T, Takahashi N, Nakane T, et al. (2012) Association of major histocompatibility complex class I haplotypes with disease progression after simian immunodeficiency virus challenge in burmese rhesus macaques. J Virol 86: 6481-6490.
18. Mothe BR, Weinfurter J, Wang C, Rehrauer W, Wilson N, et al. (2003) Expression of the major histocompatibility complex class I molecule Mamu-A*01 is associated with control of simian immunodeficiency virus SIVmac239 replication. J Virol 77: 2736-2740.
19. Otting N, Van Der Wiel MK, De Groot N, De Vos-Rouweler AJ, de Groot NG, et al. (2017) The orthologs of HLA-DQ and -DP genes display abundant levels of variability in macaque species. Immunogenetics 69: 87-99.
20. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, et al. (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364: 33-39.
21. Karl JA, Heimbruch KE, Vriezen CE, Mironczuk CJ, Dudley DM, et al. (2014) Survey of major histocompatibility complex class II diversity in pig-tailed macaques. Immunogenetics 66: 613-623.
22. Xie C, Yeo ZX, Wong M, Piper J, Long T, et al. (2017) Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. Proc Natl Acad Sci USA 114: 8059-8064.
23. Kishikawa T, Momozawa Y, Ozeki T, Mushiroda T, Inohara H, et al. (2019) Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. Sci Rep 9: 1784.
24. Yang Y, Muzny DM, Xia F, Niu Z, Person R, et al. (2014) Molecular findings among patients referred for clinical whole-exome sequencing. JAMA 312: 1870-1879.
25. Posey JE, Rosenfeld JA, James RA, Bainbridge M, Niu Z, et al. (2016) Molecular diagnostic experience of whole-exome sequencing in adult patients. Genet Med 18: 678-685.
26. Cao H, Wu J, Wang Y, Jiang H, Zhang T, et al. (2013) An integrated tool to study MHC region: Accurate SNV detection and HLA genes typing in human MHC region using targete. PLoS ONE 8: e69388.

27. Wittig M, Anmarkrud JA, Kassens JC, Koch S, Forster M, et al. (2015) Development of a high-resolution NGS-based HLA-typing and analysis pipeline. Nucleic Acids Res 43: e70.
28. Xue C, Raveendran M, Harris RA, Fawcett GL, Liu X, et al. (2016) The population genomics of rhesus macaques (Macaca mulatta) based on whole-genome sequences. Genome Res 26: 1651-1662.
29. Bimber BN, Ramakrishnan R, Cervera Juanes R, Madhira R, Peterson SM, et al. (2017) Whole genome sequencing predicts novel human disease models in rhesus macaques. Genomics 109: 214-220.
30. De Manuel M, Shiina T, Suzuki S, Dereuddre Bosquet N, Garchon HJ, et al. (2018) Whole genome sequencing in the search for genes associated with the control of SIV infection in the Mauritian macaque model. Sci Rep 8: 7131.
31. Vallender EJ (2011) Expanding whole exome resequencing into non-human primates. Genome Biol 12: R87.
32. Cornish AS, Gibbs RM, Norgren RBJ (2016) Exome screening to identify loss-of-function mutations in the rhesus macaque for development of preclinical models of human disease. BMC Genomics 17: 170.
33. Zimin AV, Cornish AS, Maudhoo MD, Gibbs RM, Zhang X, et al. (2014) A new rhesus macaque assembly and annotation for next-generation sequencing analyses. Biol Direct 9: 20.
34. Reynolds MR, Sacha JB, Weiler AM, Borchardt GJ, Glidden CE, et al. (2011) The TRIM5{alpha} genotype of rhesus macaques affects acquisition of simian immunodeficiency virus SIVsmE660 infection after repeated limiting-dose intrarectal challenge. J Virol 85: 9637-9640.
35. Janaka SK, Tameh AT, Neidermyer WJJ, Serra Moreno R, Hoxie JA, et al. (2018) Polymorphisms in rhesus macaque tetherin are associated with differences in acute viremia in simian immunodeficiency virus deltanef-infected animals. J Virol 92.
36. Krupp A, McCarthy KR, Ooms M, Letko M, Morgan JS, et al. (2013) APOBEC3G polymorphism as a selective barrier to cross-species transmission and emergence of pathogenic SIV and AIDS in a primate host. PLoS Pathog 9: e1003641.
37. Ericsen AJ, Starrett GJ, Greene JM, Lauck M, Raveendran M, et al. (2014) Whole genome sequencing of SIV-infected macaques identifies candidate loci that may contribute to host control of virus replication. Genome Biol 15: 478.
38. Haus T, Ferguson B, Rogers J, Doxiadis G, Certa U, et al. (2014) Genome typing of nonhuman primate models: Implications for biomedical research. Trends Genet 30: 482-487.
39. Animal Welfare Act. 'The animal welfare act-public law 89-544' Act of August 24, 1966. 1966.
40. Weatherall D (2006) The use of non-human primates in research. 147.
41. Bushnell B, Rood J, Singer E (2017) BBMerge-Accurate paired shotgun read merging via overlap. PLoS One 12: e0185056.
42. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460-2461.
43. Clark MJ, Chen R, Snyder M (2013) Exome sequencing by targeted enrichment. Curr Protoc Mol Biol 7(7):12.
44. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369: 1502-1511.
45. Prall TM, Graham ME, Karl JA, Wiseman RW, Ericsen AJ, et al. (2017) Improved full-length killer cell immunoglobulin-like receptor transcript discovery in mauritian cynomolgus macaques. Immunogenetics 69: 325-339.
46. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, et al. (2008) Variation analysis and gene annotation of eight MHC haplotypes: The MHC haplotype project. Immunogenetics 60: 1-18.
47. Baaijens JA, Aabidine AZE, Rivals E, Schonhuth A (2017) *De novo* assembly of viral quasispecies using overlap graphs. Genome Res 27: 835-848.
48. Koster J, Rahmann S (2012) Snakemake-A scalable bioinformatics workflow engine. Bioinformatics 28: 2520-2522.
49. Ennis PD, Zemmour J, Salter RD, Parham P (1990) Rapid cloning of HLA-A,B cDNA by using the polymerase chain reaction: Frequency and nature of errors produced in amplification. Proc Natl Acad Sci USA 87: 2833-2837.
50. Fichot EB, Norman RS (2013) Microbial phylogenetic profiling with the pacific biosciences sequencing platform. Microbiome 1 10: 1.
51. Von Wintzingerode F, Gobel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21: 213-229.
52. Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SG (2013) IPD-The immuno polymorphism database. Nucleic Acids Res 41: D1234-D1240.