MHC genotyping from rhesus macaque exome sequences

John R. Caskey[1], Roger W. Wiseman[1,2], Julie A. Karl[2], David A. Baker[1], Taylor Lee[1], Muthuswamy Raveendran[3], R. Alan Harris[3], Jianhong Hu[3], Donna M. Muzny[3],Jeffrey Rogers[3], 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 (doconnor@primate.wisc.edu; +1 608 890 0845)

Author	ORCID
Caskey	0000-0001-5665-524X
Wiseman	0000-0002-7682-7085
Karl	0000-0002-4447-4721
Baker	0000-0002-6811-8238
Lee	0000-0002-8768-4358
Raveendran	0000-0001-6185-4059
Jianghong	0000-0002-6318-8595
Harris	0000-0002-7333-4752
Rogers	0000-0002-7374-6490
O'Connor	0000-0003-2139-470X

Keywords: Macaca mulatta, exome, major histocompatibility complex

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 introduce macaque exome sequence (MES) genotyping, in which MHC class I and class II genotypes are determined with high confidence using target-enrichment probes that are enriched for MHC sequences. 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.

30 # Introduction

31 The major histocompatibility complex (MHC) is an intensively studied set of genes in macaques (Shiina et al. 2017;
32 Wiseman et al. 2013). The genomic MHC region contains clusters of genes that encode the MHC class I complex
33 and the MHC class II complex. Cells use MHC class I molecules to present intracellular peptides to immune cells
34 like the CD8+ T cell or natural killer cells (Garcia and Adams 2005). MHC class I molecules accommodate
35 intracellular peptides of varying specificity by having diverse amino acid sequences in the α1 and α2 subunits,
36 which form the peptide-binding cleft (Silver and Watkins 2017; Loffredo et al. 2009). These α 1 and α 2 subunits
37 correspond to exons 2 and 3, respectively, of an MHC class I gene (Malissen et al. 1982). Most of the
38 polymorphisms that distinguish MHC class I alleles are concentrated in exons 2 and 3 (Williams 2001). Thousands
39 of individual MHC allelic variants have been identified in the three most widely used macaque species for
40 biomedical research: rhesus (*Macaca mulatta*), cynomolgus (*Macaca fascicularis)*, and pig-tailed macaques
41 (*Macaca nemestrina*) (Semler et al. 2018; Karl et al. 2017; Maccari et al. 2017).

42 In humans, the MHC is also termed the human leukocyte antigen complex (HLA). HLA class I has a single copy of
43 the *HLA-A*, *HLA-B*, and *HLA-C* genes on each chromosome (Daza-Vamenta et al. 2004; Shiina et al. 2017). In
44 contrast, macaques have a variable number of genes on each chromosome that encode MHC class I MHC-A and
45 MHC-B proteins, and macaques lack an *HLA-C* orthologue (Daza-Vamenta et al. 2004; Shiina et al. 2017; Wiseman
46 et al. 2013). Initial approaches to genotype macaque MHC class I relied on using sequence-specific PCR
47 oligonucleotides to test for the presence or absence of individual alleles (Kaizu et al. 2007). More recently, deep
48 sequencing genomic DNA or complementary DNA PCR amplicons spanning a highly variable region of exon 2 has
49 become commonplace (Wiseman et al. 2009; Karl et al. 2013). Amplicon sequences can be used to genotype groups
50 of closely related MHC class I alleles, which are denoted as lineages. For example, an amplicon deep sequence that
51 corresponds to the rhesus macaque *Mamu-A1*001* lineage demonstrates that an animal possesses *Mamu-A1*001:01*,
52 *Mamu-A1*001:02*, or another closely related variant that has not yet been identified. This lineage-level reporting of
53 MHC class I genotypes can be sufficient for designing experiments where certain MHC class I genotypes need to be
54 matched between animals, balanced among experimental groups, or excluded entirely from a study (Loffredo et al.

55 2007; Loffredo et al. 2009; Muhl et al. 2002; Karl et al. 2013; Wiseman et al. 2009; Nomura et al. 2012; Mothe et al.
56 2003).

57 Amplicon deep sequencing can also be used for MHC class II genotyping. The human HLA class II genes *DQA1*,
58 *DQB1, DPA1*, and *DPB1* have direct orthologues in macaques (Otting et al. 2017). Both macaques and humans have
59 a variable number of MHC class II DRB genes on a single chromosome, while the DRA gene is oligomorphic
60 (Daza-Vamenta et al. 2004; Shiina et al. 2017), and typically is not used for genotyping purposes. MHC class II
61 molecules are members of the immunoglobulin superfamily, but they differ from MHC class I in several ways. The
62 MHC class II complex is comprised of an α and β subunit heterodimer, and separate genes encode each MHC class
63 II α and β subunit (Brown et al. 1993). Highly polymorphic regions that are diagnostic for MHC class II allele
64 lineages can be PCR amplified in a manner that is similar to the MHC class I genotyping. The most extensive
65 polymorphism among alleles of the MHC class II genes is found in exon 2 (Williams 2001). Each of the *DRB*,
66 *DQA1*, *DQB1*, *DPA1* and *DPB1* polymorphic MHC class II genes are sufficiently divergent that separate PCR
67 amplicons are required for deep sequencing (Karl et al. 2014).

68 Comprehensive MHC class I and class II genotyping of macaques by amplicon deep sequencing requires preparation
69 of six separate PCR amplicons for MHC class I and class II DRB, DQA1, DQB1, DPA1, and DPB1 (Karl et al.
70 2014; Karl et al. 2017). Despite this complexity, a major advantage to amplicon deep sequencing has been its cost
71 effectiveness. The output from a single MiSeq sequencing run can be used to determine the MHC class I and MHC
72 class II genotypes for up to 192 macaque samples. In recent years, improved sequencing hardware and software
73 have prompted new approaches to MHC genotyping. Instead of a utilizing PCR amplicons of variable gene regions
74 to genotype samples, researchers can use human whole exome sequencing (WES) and whole genome sequencing
75 (WGS) datasets to determine HLA genotypes (Xie et al. 2017; Kishikawa et al. 2019; Yang et a.l 2014; Posey et al.
76 2016). Likewise, target capture approaches have been described for HLA genotyping with next-generation
77 sequencing (Cao et al. 2013; Wittig et al. 2015). In contrast, MHC genotyping of macaques from WGS (Xue et al.
78 2016; Bimber et al. 2017; de Manuel et al. 2018) or whole exome sequence (Vallender 2011; Cornish et al. 2016)
79 datasets have not been reported to date. This reflects the challenges presented by mapping short sequence reads
80 against complex, duplicated gene families like the MHC class I genes of macaques where genomic reference
81 sequences and reference databases of known alleles are incomplete (Zimin et al. 2014; Maccari et al. 2017). In an

82 era where the per-base cost of sequencing is dropping rapidly, we explored the feasibility of obtaining whole exome
83 sequencing data while maintaining parity of results with the traditional MHC PCR amplicon approach.

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

90

91 # Methods

92 ## Animals

93 Twenty-seven whole blood samples were collected from Indian rhesus macaques (*Macaca mulatta*). Five of these
94 samples came from a breeding group of animals living at the Wisconsin National Primate Research Center
95 (WNPRC). The remaining 22 samples were provided by Dr. Michele Di Mascio from the National Institutes of
96 Health's National Institutes of Allergy and Infectious Diseases. Blood sampling was performed under anesthesia and
97 in accordance with the regulations and guidelines outlined in the Animal Welfare Act, the Guide for the Care and
98 Use of Laboratory Animals, and the Weatherall report (Animal Welfare Act 1966; Weatherall 2006).

99 ## Data

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

103 ## MHC class I and class II genotyping by amplicon deep sequencing

104 Genomic DNA was isolated from 250μL of whole blood using a Maxwell® 48 LEV Blood DNA Kit (Promega
105 Corporation, Fitchburg, WI). Following isolation, DNA concentrations were determined with a Nanodrop 2000 and
106 samples were normalized to 60 ng/ul. MHC class I and class II PCR amplicons were generated using exon 2-specific
107 primers with adapters (CS1 and CS2) necessary for 4-primer amplicon tagging with the Fluidigm Access Array™
108 System (Fluidigm, San Francisco, CA, USA) by previously described methods (Karl et al. 2017; Karl et al. 2014).
109 Pooled PCR products were purified using the AMPure XP beads (Agencourt Bioscience Corporation, Beverly, MA,
110 USA) and quantified using the Quant-iT dsDNA HS Assay kit with a Qubit fluorometer (Invitrogen, Carlsbad, CA,
111 USA), following the manufacturer's protocols. The MHC exon 2 genotyping amplicon pools were sequenced on an
112 Illumina MiSeq instrument (San Diego, CA, USA) as previously described (Karl et al. 2017).

113 Analysis of the MiSeq exon 2 genotyping amplicon sequences was performed using a custom Python workflow. The
114 workflow contained a step to remove oligonucleotide primers and sequencing adapters with bbduk, merge a step to

115 reads bbmerge, a step to identify unique sequences/remove chimeras with USEARCH, and finally mapping unique
116 reads against a deduplicated reference database of rhesus macaque partial MHC class I and class II exon 2 sequences
117 with bbmap (Bushnell et al. 2017; Edgar 2010). The database was derived from sequences in the IPD-MHC NHP
118 database downloaded from the European Bioinformatics Institute website (Maccari et al. 2017). For this publication,
119 we define "IPD exon 2" as this database of reference sequences. SAM output files from bbmap were parsed with the
120 Python package pandas to enumerate the reads from each animal that were identical to IPD exon 2 reference
121 sequences. *Mamu-A, -B, -DRB, -DQA1, -DQB1, -DPA1* and *-DPB1* lineage-level haplotypes were inferred for each
122 of the samples with a semi-automated custom workflow that identifies diagnostic alleles associated with previously
123 defined rhesus macaque haplotypes (Karl et al. 2013; Otting et al. 2017).

124 **MHC class I and class II genotyping by exome sequencing**

125 Genomic DNA was isolated as described above and shipped to the Human Genome Sequencing Center at the Baylor
126 College of Medicine. MHC and exon-containing genomic DNA fragments were selectively enriched using a custom
127 target-capture probeset. Genome-wide exons were captured using SeqCap EZ HGSC VCRome2.1, an optimized
128 human clinical exome probeset (Clark et al. 2013, Yang et al. 2013). SeqCap EZ HGSC VCRome2.1 contains
129 probes designed to enrich 23,585 human genes and 189,028 non-overlapping exons. A low coverage audit was
130 performed to identify rhesus macaque exons inferred from the reference genome rheMac2 that were not sufficiently
131 enriched (<20x coverage) with these human probes, and an additional 22,884 rhesus macaque exons were
132 incorporated into the genotyping probe design (Prall et al. 2017). Finally, and most importantly for MHC analyses,
133 we modified the SeqCap EZ Design: Human MHC Design to selectively enrich MHC class I and class II sequences.
134 This previous design was prepared by the Beijing Genome Institute in collaboration with Roche/Nimblegen and it
135 targeted the complete 4.97Mb HLA region with non-redundant probes designed against 8 fully sequenced HLA
136 haplotypes (Horton et al. 2008; Cao et al. 2013). For our macaque studies, we prepared a minimal MHC target
137 capture design using a subset of these probes that are based on all functional HLA class I (*HLA-A*, *-B*, *-C,* and *-E*)
138 and class II (*HLA-DRA*, *-DRB1*, *-DRB3*, *-DRB4*, *-DRB5*, *-DQA1*, *-DQB1*, *-DPA1* and *-DPB1*) genes. Probes were
139 included to capture complete gene sequences (exons + introns + 3' UTR) as well as ~1 kb of 5' upstream flanking
140 sequence. The BED file of rhesus rheMac2 target coordinates lifted over to rheMac8 was used to prepare this
141 combined minimal MHC and supplemental rhesus spike-in probe design. Because derivation of MHC results from

142 genotyping is paramount, we used a ratio of 2.5x spike-in probes to 1x VCRom2.1 probes. The supplemental probes
143 for MHC and rhesus macaque were a single reagent, and the MHC-specific probes only constituted 609 kb of the
144 37.9 Mb probes (Prall et al. 2017).

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

150 Enumeration of MHC reads in exome data

151 The effective enrichment of MHC reads in the exome datasets was calculated by mapping the reads of each sample's
152 exome dataset against an individual genomic reference file for each individual locus: *HLA-A* exons 2-3 (NCBI Gene
153 ID: 3105) and *HLA-E* exons 2-3 (Gene ID: 3133); *HLA-DPA1* exons 2-4 (Gene ID: 3113); *HLA-DPB1* exons 2-4
154 (Gene ID: 3115); *HLA-DQA1* exons 2-4 (Gene ID: 3117); *HLA-DQB1* exons 2-4 (Gene ID: 3119); *HLA-DRB1*
155 exons 2-4 (Gene ID: 3123), *HLA-DRB3* exons 2-4 (Gene ID: 3125), *HLA-DRB4* exons 2-4 (Gene ID: 3126) and
156 *HLA-DRB5* exons 2-4 (Gene ID: 3127). This mapping was done by using bbmap with default parameters, which
157 corresponds to a minimum alignment identity of approximately 76% (Bushnell et al. 2017). Empirically, these
158 mapping parameters are sufficient to map macaque MHC reads to their human orthologues. Mapped reads were
159 written to a new fastq file using bbmap's outm= parameter. To quantify the total number of reads in a sample and
160 the number of reads extracted with our reference file, we created a custom Python script, which is available to
161 download.

162 MHC genotyping from exome data

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

168 *Strategy 1: MHC genotyping using Diagnostic Sub-Region (DSR)*

169 The Diagnostic Sub-Region (DSR) was an intra-allelic region that encompassed polymorphisms, and these
170 polymorphisms were distinguishable from alleles with similar sequences. Therefore, this method ensured the DSR
171 was captured in at least one read for each called allele. MHC class I and class II reads initially were extracted from
172 each animal by mapping the FASTQ reads to HLA class I and class II reference sequences containing exons 2-3 and
173 exons 2-4, respectively, plus the intervening intron(s) as described above in 'Enumeration of MHC reads in exome
174 data'. Reads were mapped to these reference sequences using bbmap with default parameters and the parameter
175 (qtrim=lr) (Bushnell et al. 2017).

176 Following extraction of MHC reads from the total exome sequences, the MHC reads were prepared for assembly
177 using a modified version of a data pre-processing pipeline, which included tools from the BBTools package
178 (Bushnell et al. 2017). Briefly, optical duplicates and reads from low-quality regions of the sequencing run were
179 removed. Next, Illumina sequencing adapters were trimmed from the ends of sequencing reads. Any residual spike-
180 in or PhiX sequences that inadvertently survived mapping to HLA class I and class II were then removed. Three
181 rounds of error-correction and read merging were performed to create high-confidence merged reads that were well-
182 supported by common kmers found in the extracted MHC reads. The error-corrected reads were not merged with a
183 minimum overlap, but instead were separately mapped against the IPD exon 2 sequences using bbmapskimmer. The
184 default settings for the software tool bbmapskimmer were used with the following modified parameters
185 (semiperfectmode=t ambiguous=all ssa=t maxsites=50000 maxsites2=50000 expectedsites=50000) (Bushnell et al.
186 2017). The semiperfectmode setting accepted reads with perfect matches, as well as reads that extended off the end
187 of contigs for no more than half of the length of the mapped read segment. The ambiguous setting and the
188 'expectedsites' setting reported the first 50,000 matched read segments that met the 'semiperfectmode' filtering,
189 which was set exceedingly high in order to exhaustively map our IPD exon 2 reference file of ~1600 alleles. Using
190 semiperfect mode, these were the segment sequences with the longest matching length. This output file included all
191 mapped reads to all IPD exon 2 reference allele sequences.

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

200 *Strategy 2: De novo reconstruction of MHC sequences from exome reads*

201 Most *de novo* sequence assemblers have been optimized for resolving long contigs, and are tolerant of small
202 sequence mismatches that can otherwise fragment assemblies. In the case of MHC alleles, however, such closely
203 related sequences were often biologically distinct. Because of the gene duplications in the macaque MHC, there
204 were many valid MHC contigs that could be assembled from a single sample within the exome data. This is
205 conceptually similar to the computational challenge of assembling viral haplotypes, where rapidly evolving viruses
206 such as human immunodeficiency virus accumulate variants that frequently co-segregate as minor populations
207 within an infected person (Baaijens et al. 2017). Therefore, we utilized the overlap assembly algorithm SAVAGE,
208 originally designed to reconstruct viral haplotypes, to reconstruct MHC allele sequences from exome reads (Baaijens
209 et al. 2017). Similar to Strategy 1 above, HLA-mapped reads were pre-processed using the BBTools package to
210 remove low quality reads and optical duplicates. Next, adapters were trimmed, residual spike-in and PhiX sequences
211 were removed, three steps of error-correction took place, and reads were merged. These merged reads, as well as
212 high-confidence unmerged paired-end reads that could not be grouped into an overlapping merged read, were used
213 for SAVAGE assembly. The following workflow was implemented in a reproducible snakemake workflow that fully
214 documents parameter selection (Koster and Rahman 2012), and is available upon request.

215 SAVAGE is designed to construct individual haplotypes from the overlap graph of individual reads. We processed
216 the totality of sequence data in a single patch in order to maximize the sensitivity of the contiguous reconstruction
217 with parameter '--split 1', as well as the parameter '--revcomp' to handle reverse complement reads. The IPD exon 2

218 reference database that was used for MiSeq and DSR genotyping was also used to assess the quality of SAVAGE
219 genotypes. The IPD sequences were mapped to contigs produced by SAVAGE using sensitive parameters in bbmap
220 (minlength=100 vslow=t subfilter=0 indelfilter=0 lengthtag=t ignorefrequentkmers=t kfilter=100) designed to
221 identify sequences that perfectly match SAVAGE contigs. A post-processing script refined these mappings, and
222 only retained the mappings where the length of the mapped region was the same as the length of the IPD exon 2
223 sequence. These mappings indicated where the reference database sequence was fully and exactly contained within a
224 SAVAGE contig.

225

226

227

228 # Results

229 ## MHC reads are efficiently enriched using target-capture probes

230 We designed a custom target enrichment probeset that accounts for the extensive duplication of macaque MHC
231 genes (Prall et al. 2017). A tripartite target capture system was used in this study. The first component is SeqCap EZ
232 HGSC VCRome2.1, an optimized human clinical exome probeset. Since macaques are closely related to humans,
233 SeqCap EZ HGSC VCRome2.1 can also be used in macaques, though some sequences that are most divergent
234 between macaques and humans are not efficiently captured. The second component of the target capture system is
235 an additional 22,884 rhesus macaque exon sequences that were not effectively captured with the SeqCap EZ HGSC
236 VCRome2.1 reagent. The third component is a collection of probes designed to specifically enrich macaque MHC
237 class I and class II sequences. These probes span the full length of HLA class I and class II genes including introns,
238 3' UTRs, and approximately 1,000 bp of flanking 5' sequence.

239 We obtained a median coverage of 100X for the target exon sequences across the genome with >20X coverage for
240 94.97% of bases that were targeted in this study. As illustrated in **Table 1**, an average of 70,549,789 Illumina
241 sequence reads per sample were obtained for the 27 animals evaluated in this study. These reads were mapped
242 against reference files containing representative genomic HLA exons 2 -3 for *HLA-A,* and *HLA-E* exons 2 - 4 for
243 *HLA-DRA*, *-DRB*1, *-DRB3*, *-DRB4*, *-DRB5*, *-DQA1*, *-DQB1*, *-DPA1* and *-DPB1* sequences. We identified an
244 average of 269,057 MHC class I and class II sequence reads per sample which corresponds to an average of 0.37%
245 of the total sequence reads evaluated per sample (**Table 1**). In a previous study by Ericsen *et al*. (Ericsen et al.
246 2014), we found that MHC sequences only accounted for an average of 0.13% of the total Illumina sequence reads
247 that were evaluated per animal when the standard HGSC VCRome2.1 panel was used alone for target capture
248 (**Supplementary Table 1**). Thus, we achieved an almost three-fold increase of MHC genomic sequences after
249 inclusion of the spike-in probes for target capture compared to use of the VCRome2.1 probeset alone.

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

252 We hypothesized that MHC genotypes derived from target-enriched genomic sequence would be comparable in
253 accuracy to MHC genotypes derived from conventional amplicon deep sequencing. MHC class I and class II PCR
254 amplicons were generated from the same 27 animals and deep sequenced on an Illumina MiSeq. Representative
255 MHC class I and class II genotypes supported by the amplicon data are shown in **Figure 1** and **Figure 2**,
256 respectively. These figures illustrate genotypes from a pedigreed family of Indian rhesus macaques: a Sire, a Dam,
257 and three Progeny that are paternal half-siblings. Comprehensive genotypes for all 27 Indian rhesus macaques
258 evaluated in this study are shown in **Supplementary Figure 1**.

259 The superset of alleles supported by at least two of the methods described in this manuscript (MiSeq amplicon
260 genotyping, DSR genotyping of individual exome sequencing reads, or SAVAGE genotyping of contigs derived
261 from exome data), was considered to be the expected MHC genotypes for these animals. A complication for this
262 assertion was certain genotypes were difficult for MiSeq analysis to report correctly. As we have shown previously
263 (Karl et al. 2017), the number of reads supporting each genotyping call was highly variable, ranging from tens of
264 reads to thousands of reads per allele. A small subset of allelic variants contained nucleotide substitutions in their
265 sequences within the binding sites of the PCR oligonucleotides that interfere with efficient PCR amplification. This
266 result is exemplified when members of the *Mamu-B11L*01* allele lineage were closely examined. Members of the
267 *Mamu-B11L*01* allele lineage have two nucleotide substitutions relative to the 5' oligonucleotide that was used to
268 generate MHC class I amplicons, so these *Mamu-B11L*01* sequences were routinely absent in MiSeq genotypes
269 (**Figure 1**). Likewise, the *Mamu-DPA1*11:01* allele has four nucleotide substitutions relative to the oligonucleotide
270 pair used to generate *DPA1* amplicons for MiSeq genotyping (**Figure 2**).

271 These nucleotide substitutions do not fully account for all differences in the abundance of sequence reads for each
272 allele. False negatives were also noted for certain allele lineages, such as *Mamu-B*074* and *Mamu-B*098,* which
273 exhibited significantly diminished PCR efficiency despite being completely matched with the oligonucleotides used
274 for amplification (**Figure 1**). False positive genotyping calls were also noted in the MiSeq assay that resulted from
275 intermolecular recombination during the PCR process. This was exemplified by read support for the presence of
276 *Mamu-B*007:07* in Sire r07010, and Progeny 2 r17041 (**Figure 1**). *Mamu-B*007:07* only differs from the *Mamu-*

277 *B*007g1* allele group that was determined to be present in this pair of animals by a single nucleotide variant at the
278 extreme 5' end of the class I genotyping amplicon. Chimeric PCR products equivalent to the *Mamu-B*007:07*
279 sequence were formed between the 3' portion of the *Mamu-B*007g1* sequence and other allelic variants in these
280 animals with this 5' SNP. Taken together, these results illustrate that the MiSeq amplicon genotyping, while
281 generally reflective of an animal's MHC genotype, can yield both false positive and false negative results. When
282 compared to the expected genotypes for this dataset, MiSeq amplicon genotyping has 90.4% MHC class I and
283 97.8% MHC class II concordance (**Table 2**).

284 Two separate strategies were used to derive MHC genotypes from target-capture data. The DSR analysis was a
285 straightforward extension of the methodology used for deriving genotypes from MiSeq amplicons. MHC sequence
286 reads were extracted from the total exome dataset. A simplified workflow that mapped those MHC reads to
287 reference sequences, and then assigned the genotypes based on the presence of exome reads overlapping each
288 position of the reference sequence, was problematic. This simplified workflow was extremely vulnerable to false
289 positive genotypes, and the similarities among different MHC sequences often enabled those reads to map to
290 multiple different alleles (Wiseman et al. 2013). Two or more reads each could partially match a portion of a
291 sequence in the IPD exon 2 database, which could complement each other to provide support for an allele that was
292 not biologically relevant. As discussed in the Methods, the DSR encompassed polymorphisms that discriminated
293 among closely related alleles by requiring at least one mapped read to unambiguously map to the corresponding
294 allele within the IPD exon 2 sequences. This strategy can identify specific polymorphisms of interest among the IPD
295 exon 2 sequences to produce genotypes for MHC class I and MHC class II (**Figures 1**, **2** and **Supplementary**
296 **Figure 1**). For the 27 animals evaluated in this study, DSR was 97.3% and 100% concordant with expected MHC
297 class I and class II genotypes, respectively (**Table 2**).

298 The DSR strategy did not consider sequences that were not among the IPD exon 2 sequences and this lack of
299 consideration contributes to overcalled alleles. The apparent *Mamu-A1*059:01* allele in Dam r05029 was
300 erroneously derived from reads that were from *Mamu-A1*004g1* and an unknown allele that was not among the
301 sequences in the IPD exon 2 database. As a result of this unknown sequence not being among the known IPD exon

302 2 sequences, reads unambiguously mapped to *Mamu-A1*059:01*, which caused this allele to be overcalled. This type
303 of overcalling will be mitigated with the discovery of additional allelic variants, and their inclusion in future
304 iterations of the IPD database.

305 The second approach for determining genotypes performed *de novo* assembly on MHC class I and class II reads
306 from each sample. The resulting assembled contigs were then mapped against IPD exon 2 reference sequences to
307 define perfectly matching contigs. Because most assemblers were not tuned for the challenge of assembling large
308 numbers of contigs that differ from one another by as little as 1 bp, we relied on an assembler, SAVAGE, originally
309 designed to reconstruct viral sequencing haplotypes. As shown in **Figures 1, 2 and Supplementary Figure 1**,
310 contigs produced by SAVAGE matched the expected genotypes. Unlike the DSR method, SAVAGE assembled
311 contigs for downstream analyses. False positives and false negatives among closely-related variants were mitigated
312 by the inclusion of the filtering steps described under Strategy 2. Across all 27 samples, the SAVAGE contigs are
313 98.3% and 99.7% accurate with respect to the expected MHC genotypes (**Table 2**).

314 While this analysis focused on genotyping using the same exon 2 reference sequences that are commonly utilized for
315 MiSeq amplicon analyses, the SAVAGE contigs are frequently much longer than these reference sequences (**Figure**
316 **3**). These extended contigs frequently contain exons 2 through 4, plus the intervening introns, and could be used to
317 provide higher resolution genotyping than is possible using exon 2 sequence alone. Moreover, contigs that contain
318 complete sequences for exons 2-3 of MHC class I and exon 2 of MHC class II alleles meet the minimum criteria for
319 obtaining formal allele nomenclature for non-human primates from the IPD-MHC (Robinson et al. 2013).

320

321

322 # Discussion

323 Here we describe MES genotyping of Indian rhesus macaques. In our estimation, this method will supersede MiSeq
324 PCR amplicon genotyping as the most widely used macaque MHC genotyping assay in the future. This
325 methodology has a number of compelling advantages. Most importantly, exome sequencing dramatically improves
326 the overall quantity and quality of genomic information obtained from each sample. The same datasets used for
327 MHC analyses may be used to evaluate protein-coding genetic variation throughout the genome. The loss of start
328 and stop codons exome-wide can be obtained from the same datasets by modifying the workflows *in silico*, as
329 opposed to requiring the sequencing of multiple new sets of target genes (Yang et al. 2013). Exome-wide datasets
330 offer the promise of retrospectively identifying candidate DNA sequence variants that may be responsible for
331 unexpected experimental outcomes in studies with macaques and other nonhuman primates. The availability of
332 exome-wide sequences in conjunction with MHC genotypes may also increase the rigor of prospective macaque
333 experiments by enabling more sophisticated balancing of experimental groups, and exclusion of animals whose
334 genetics are likely to strongly bias experimental results (Loffredo et al. 2007; Reynolds et al. 2011; Haus et al.
335 2014).

336 These same exome datasets also offer the potential to improve the quality of MHC genotyping. Full-length long-read
337 MHC transcript sequencing offers exquisite resolution, but this technology can be labor intensive and difficult to
338 scale (Karl et al. 2017; Semler et al. 2018). The MiSeq exon 2 amplicon approach is limited in its allelic resolution,
339 but it is the current standard deep sequencing approach for high-throughput MHC genotyping in macaques. In this
340 report, we compare two novel strategies for MHC genotyping from MES datasets to this standard MiSeq amplicon
341 genotyping method. The results that we obtained with all three approaches show strong concordance with expected
342 MHC genotypes for the Indian rhesus macaques that were evaluated. As illustrated in **Figure 3a**, MHC class I
343 genomic contigs with an average length of approximately 1.1 kb can be assembled from sequence reads that were
344 initially extracted from the whole exome datasets.

345 Following the initial enrichment step, MHC class II genomic contigs averaging approximately 1.9 kb in length could
346 be assembled from whole exome sequence reads that mapped to *HLA-DRB1*,*-DRB4*, *-DRB5*, *-DQA1*, *-DQB1*, -
347 *DPA1* and *-DPB1* reference sequences, which contained exons 2 - 4, and the pairs of introns (**Figure 3b**). The

348 addition of HLA spike-in probes containing both exon and intron sequences for the target capture step, and the
349 introduction of 151 bp paired end reads for the Illumina NovaSeq platform greatly facilitated our ability to assemble
350 these extended MHC genomic contigs. Additional technological advances, including even longer sequence reads and
351 more efficient assembly algorithms, will undoubtedly increase genomic contig lengths as well as MHC allelic
352 resolution in future studies.

353 Both MHC genotyping approaches described here depend upon mapping exome sequence reads against a reference
354 database of MHC class I and class II allele sequences. Currently, the IPD database essentially is restricted to coding
355 regions for macaque MHC sequences, and many IPD entries are only partial transcript sequences that lack complete
356 coding regions. In addition, it is very challenging to correctly phase short Illumina sequence read that map to
357 different exons of a specific allelic variant and are separated by intronic sequences that span hundreds to thousands
358 of base pairs in genomic DNA. Our results with the SAVAGE workflow (**Figure 3**) demonstrate that exome
359 sequence reads that have been enriched with the enhanced MHC probe design described here can be assembled into
360 genomic contigs that span multiple exons and introns. These contigs, therefore, could also be used to improve MHC
361 reference databases, though this requires a major effort with more animals, and is beyond the scope of this
362 manuscript.

363 The current cost of genotyping is relatively high compared to amplicon deep sequencing, largely due to two
364 expenses. First, the amount of sequence data needed for a single sample can be high. Compared to conventional
365 MHC genotyping, where 192 macaques' data can be collected on a single instrument run, exome data acquisition is
366 much more expensive. On an Illumina Novaseq instrument, which has a much higher run cost than the MiSeq, only
367 70 exome samples can be sequenced simultaneously per lane of a S4 flow cell. However, this cost of sequencing is
368 rapidly decreasing, and as of early 2019, commercial providers have advertised sequencing for $9 USD per Gb of
369 whole genome sequence data[1]. Second, a major expense in MHC genotyping is the production, validation, and use
370 of target-capture arrays, as well as the development of *in silico* data analysis workflows. The approaches described
371 here, in particular MHC genotyping from SAVAGE contigs produced by *de novo* assembly of exome reads, is
372 flexible and should be adaptable as sequencing approaches evolve and improve.

[1] Comment via Twitter 03/18/2019 @albertvilella https://twitter.com/AlbertVilella/status/1107524501645000705

373 A major goal for future studies will be to attempt to extend these contigs to encompass full length genomic MHC
374 sequences using SAVAGE or other assembly software tools. The relatively compact genomic structure and
375 consistent length of MHC class I genes increase the attainability of this goal. Establishment of comprehensive
376 macaque MHC allele databases of extended genomic sequences will greatly facilitate mapping of exome sequence
377 reads since they will be contiguous with the reference sequences instead of being interrupted by intervening
378 sequences between each exon that are not included in current non-human primate IPD-MHC databases (Maccari et
379 al. 2017).

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

384

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 contracts 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.

405 # Figure Legends

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

408 **Table 2.** Summary of discordant results for each MHC genotyping method by animal.

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

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

424 **Figure 3**. Lengths of MHC sequence contigs assembled by the SAVAGE method for a representative breeding
425 group of rhesus macaques. Values in the body of this figure indicate contig lengths generated for these MHC
426 sequences in each animal. (**A**) MHC class I genotyping results are illustrated for the five animals in this breeding
427 group. Three false negative allele calls for SAVAGE versus the expected genotypes for these animals are
428 highlighted in magenta, e.g., *Mamu-B*030g1* in Sire r07010. Sequences highlighted in red are associated with the
429 maternal MHC Haplotype a that was inherited by Progeny 1 from its Dam. Sequences in Progeny 1 and Progeny 3

430 for their MHC Haplotype b that was inherited from Sire r07010 are highlighted in dark blue while the light blue
431 sequences represent the alternate paternal Haplotype c that was inherited by Progeny 2. Three unique extended
432 MHC Haplotypes in this breeding group are indicated with shades of grey (Haplotypes d - f). MHC allele groups
433 that are shared by both parental haplotypes are indicated by colored borders around filled cells. (**B**) MHC class II
434 genotyping results are illustrated for this same breeding group. (**C**) Abbreviated Mamu haplotype designations (Karl
435 et al. 2013) are summarized for the six extended MHC haplotypes identified by segregation in this breeding group.
436 For example, the twenty MHC sequences (red) that are associated with extended Haplotype a can be summarized by
437 the following string of abbreviated Mamu haplotype designations: Mamu-
438 A004/B048/DR04a/DQA01g1/DQB06:01/DPA02g1/DPB15g.

439

440

441 # Supplementary Materials

442 **Supplementary Table 1.** Fraction of total exome sequence reads corresponding to MHC class I and class II genes
443 after target capture with the VCRome2.1 probe design alone. This exome sequence dataset was described previously
444 by Ericsen and coworkers (Ericsen et al. 2014).

445 **Supplementary Figure 1.** Comparison of MHC class I results from MiSeq PCR amplicon versus whole exome
446 genotyping assays with the DSR and SAVAGE strategies for all 27 animals. Results for each of the three method are
447 provided side-by-side in the columns for each macaque. Each row indicates the detection of a specific MHC class I
448 allele or lineage group of closely related sequences that are ambiguous because they are identical over the IPD exon
449 2 database sequence. Values in the body of this figure indicate the number of sequence reads supporting each allele
450 call for the MiSeq and DSR methods while alleles supported by a SAVAGE contig are reported with a "1".
451 Discrepancies between the MiSeq (pink), DSR (yellow) or SAVAGE (blue) methods are highlighted by filled cells
452 with borders.

453

454

455

456

457 # References

458 Animal Welfare Act. 'The Animal Welfare Act - Public Law 89-544 Act of August 24, 1966. 1966.
459 https://www.nal.usda.gov/awic/animal-welfare-act-public-law-89-544-act-august-24-1966 . Accessed
460 05/01/2019

461 Baaijens, J. A., A. Z. E. Aabidine, E. Rivals, and A. Schonhuth. 2017. De novo assembly of viral quasispecies using
462 overlap graphs. *Genome Res* 27: 835-848. https://www.ncbi.nlm.nih.gov/pubmed/28396522

463 Bimber, B. N., R. Ramakrishnan, R. Cervera-Juanes, R. Madhira, S. M. Peterson, R. B. J. Norgren, and B. Ferguson.
464 2017. Whole genome sequencing predicts novel human disease models in rhesus macaques. *Genomics* 109:
465 214-220. https://www.ncbi.nlm.nih.gov/pubmed/28438488

466 Brown, J. H., T. S. Jardetzky, J. C. Gorga, L. J. Stern, R. G. Urban, J. L. Strominger, and D. C. Wiley. 1993. Three-
467 dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364: 33-39.
468 https://www.ncbi.nlm.nih.gov/pubmed/8316295

469 Bushnell, B., J. Rood, and E. Singer. 2017. BBMerge - Accurate paired shotgun read merging via overlap. *PLoS*
470 *One* 12: e0185056. https://www.ncbi.nlm.nih.gov/pubmed/29073143

471 Cao, H., J. Wu, Y. Wang, H. Jiang, T. Zhang, X. Liu, Y. Xu, D. Liang, P. Gao, Y. Sun, B. Gifford, M. D'Ascenzo,
472 X. Liu, L. C. Tellier, F. Yang, X. Tong, D. Chen, J. Zheng, W. Li, T. Richmond, X. Xu, J. Wang, and Y. Li.
473 2013. An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human
474 MHC region using targeted high-throughput sequencing. *PLoS One* 8: e69388.
475 https://www.ncbi.nlm.nih.gov/pubmed/23894464

476 Clark, M. J., R. Chen, and M. Snyder. 2013. Exome sequencing by targeted enrichment. *Curr Protoc Mol Biol*
477 Chapter 7: Unit7.12. https://www.ncbi.nlm.nih.gov/pubmed/23547016

478 Cornish, A. S., R. M. Gibbs, and R. B. J. Norgren. 2016. Exome screening to identify loss-of-function mutations in
479 the rhesus macaque for development of preclinical models of human disease. *BMC Genomics* 17: 170.
480 https://www.ncbi.nlm.nih.gov/pubmed/26935327

481 Daza-Vamenta, R., G. Glusman, L. Rowen, B. Guthrie, and D. E. Geraghty. 2004. Genetic divergence of the rhesus
482 macaque major histocompatibility complex. *Genome Res* 14: 1501-1515.
483 https://www.ncbi.nlm.nih.gov/pubmed/15289473

484 de Manuel, M., T. Shiina, S. Suzuki, N. Dereuddre-Bosquet, H. J. Garchon, M. Tanaka, N. Congy-Jolivet, A.
485 Aarnink, R. Le Grand, T. Marques-Bonet, and A. Blancher. 2018. Whole genome sequencing in the search
486 for genes associated with the control of SIV infection in the Mauritian macaque model. *Sci Rep* 8: 7131.
487 https://www.ncbi.nlm.nih.gov/pubmed/29739964
488 Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460-2461.
489 https://www.ncbi.nlm.nih.gov/pubmed/20709691
490 Ericsen, A. J., G. J. Starrett, J. M. Greene, M. Lauck, M. Raveendran, D. R. Deiros, M. S. Mohns, N. Vince, B. T.
491 Cain, N. H. Pham, J. T. Weinfurter, A. L. Bailey, M. L. Budde, R. W. Wiseman, R. Gibbs, D. Muzny, T. C.
492 Friedrich, J. Rogers, and D. H. O'Connor. 2014. Whole genome sequencing of SIV-infected macaques
493 identifies candidate loci that may contribute to host control of virus replication. *Genome Biol* 15: 478.
494 https://www.ncbi.nlm.nih.gov/pubmed/25418588
495 Garcia, K. C., and E. J. Adams. 2005. How the T cell receptor sees antigen--a structural view. *Cell* 122: 333-336.
496 https://www.ncbi.nlm.nih.gov/pubmed/16096054
497 Haus, T., B. Ferguson, J. Rogers, G. Doxiadis, U. Certa, N. J. Rose, R. Teepe, G. F. Weinbauer, and C. Roos. 2014.
498 Genome typing of nonhuman primate models: implications for biomedical research. *Trends Genet* 30: 482-
499 487. https://www.ncbi.nlm.nih.gov/pubmed/24954183
500 Horton, R., R. Gibson, P. Coggill, M. Miretti, R. J. Allcock, J. Almeida, S. Forbes, J. G. Gilbert, K. Halls, J. L.
501 Harrow, E. Hart, K. Howe, D. K. Jackson, S. Palmer, A. N. Roberts, S. Sims, C. A. Stewart, J. A. Traherne,
502 S. Trevanion, L. Wilming, J. Rogers, P. J. de Jong, J. F. Elliott, S. Sawcer, J. A. Todd, J. Trowsdale, and S.
503 Beck. 2008. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project.
504 *Immunogenetics* 60: 1-18. https://www.ncbi.nlm.nih.gov/pubmed/18193213
505 Kaizu, M., G. J. Borchardt, C. E. Glidden, D. L. Fisk, J. T. Loffredo, D. I. Watkins, and W. M. Rehrauer. 2007.
506 Molecular typing of major histocompatibility complex class I alleles in the Indian rhesus macaque which
507 restrict SIV CD8+ T cell epitopes. *Immunogenetics* 59: 693-703.
508 https://www.ncbi.nlm.nih.gov/pubmed/17641886
509 Karl, J. A., P. S. Bohn, R. W. Wiseman, F. A. Nimityongskul, S. M. Lank, G. J. Starrett, and D. H. O'Connor. 2013.
510 Major histocompatibility complex class I haplotype diversity in Chinese rhesus macaques. *G3 (Bethesda)* 3:
511 1195-1201. https://www.ncbi.nlm.nih.gov/pubmed/23696100

512 Karl, J. A., M. E. Graham, R. W. Wiseman, K. E. Heimbruch, S. M. Gieger, G. G. Doxiadis, R. E. Bontrop, and D.
513 H. O'Connor. 2017. Major histocompatibility complex haplotyping and long-amplicon allele discovery in
514 cynomolgus macaques from Chinese breeding facilities. *Immunogenetics* 69: 211-229.
515 https://www.ncbi.nlm.nih.gov/pubmed/28078358
516 Karl, J. A., K. E. Heimbruch, C. E. Vriezen, C. J. Mironczuk, D. M. Dudley, R. W. Wiseman, and D. H. O'Connor.
517 2014. Survey of major histocompatibility complex class II diversity in pig-tailed macaques. *Immunogenetics*
518 66: 613-623. https://www.ncbi.nlm.nih.gov/pubmed/25129472
519 Kishikawa, T., Y. Momozawa, T. Ozeki, T. Mushiroda, H. Inohara, Y. Kamatani, M. Kubo, and Y. Okada. 2019.
520 Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci Rep* 9:
521 1784. https://www.ncbi.nlm.nih.gov/pubmed/30741997
522 Koster, J., and S. Rahmann. 2012. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* 28: 2520-
523 2522. https://www.ncbi.nlm.nih.gov/pubmed/22908215
524 Loffredo, J. T., J. Maxwell, Y. Qi, C. E. Glidden, G. J. Borchardt, T. Soma, A. T. Bean, D. R. Beal, N. A. Wilson,
525 W. M. Rehrauer, J. D. Lifson, M. Carrington, and D. I. Watkins. 2007. Mamu-B*08-positive macaques
526 control simian immunodeficiency virus replication. *J Virol* 81: 8827-8832.
527 https://www.ncbi.nlm.nih.gov/pubmed/17537848
528 Loffredo, J. T., J. Sidney, A. T. Bean, D. R. Beal, W. Bardet, A. Wahl, O. E. Hawkins, S. Piaskowski, N. A. Wilson,
529 W. H. Hildebrand, D. I. Watkins, and A. Sette. 2009. Two MHC class I molecules associated with elite
530 control of immunodeficiency virus replication, Mamu-B*08 and HLA-B*2705, bind peptides with sequence
531 similarity. *J Immunol* 182: 7763-7775. https://www.ncbi.nlm.nih.gov/pubmed/19494300
532 Maccari, G., J. Robinson, K. Ballingall, L. A. Guethlein, U. Grimholt, J. Kaufman, C. S. Ho, N. G. de Groot, P.
533 Flicek, R. E. Bontrop, J. A. Hammond, and S. G. Marsh. 2017. IPD-MHC 2.0: an improved inter-species
534 database for the study of the major histocompatibility complex. *Nucleic Acids Res* 45: D860-D864.
535 https://www.ncbi.nlm.nih.gov/pubmed/27899604
536 Malissen, M., B. Malissen, and B. R. Jordan. 1982. Exon/intron organization and complete nucleotide sequence of
537 an HLA gene. *Proc Natl Acad Sci U S A* 79: 893-897. https://www.ncbi.nlm.nih.gov/pubmed/6461010
538 Mothe, B. R., J. Weinfurter, C. Wang, W. Rehrauer, N. Wilson, T. M. Allen, D. B. Allison, and D. I. Watkins. 2003.
539 Expression of the major histocompatibility complex class I molecule Mamu-A*01 is associated with control

540 of simian immunodeficiency virus SIVmac239 replication. *J Virol* 77: 2736-2740.
541 https://www.ncbi.nlm.nih.gov/pubmed/12552014
542 Muhl, T., M. Krawczak, P. Ten Haaft, G. Hunsmann, and U. Sauermann. 2002. MHC class I alleles influence set-
543 point viral load and survival time in simian immunodeficiency virus-infected rhesus monkeys. *J Immunol*
544 169: 3438-3446. https://www.ncbi.nlm.nih.gov/pubmed/12218167
545 Nomura, T., H. Yamamoto, T. Shiino, N. Takahashi, T. Nakane, N. Iwamoto, H. Ishii, T. Tsukamoto, M. Kawada,
546 S. Matsuoka, A. Takeda, K. Terahara, Y. Tsunetsugu-Yokota, N. Iwata-Yoshikawa, H. Hasegawa, T. Sata, T.
547 K. Naruse, A. Kimura, and T. Matano. 2012. Association of major histocompatibility complex class I
548 haplotypes with disease progression after simian immunodeficiency virus challenge in burmese rhesus
549 macaques. *J Virol* 86: 6481-6490. https://www.ncbi.nlm.nih.gov/pubmed/22491464
550 Otting, N., M. K. van der Wiel, N. de Groot, A. J. de Vos-Rouweler, N. G. de Groot, G. G. Doxiadis, R. W.
551 Wiseman, D. H. O'Connor, and R. E. Bontrop. 2017. The orthologs of HLA-DQ and -DP genes display
552 abundant levels of variability in macaque species. *Immunogenetics* 69: 87-99.
553 https://www.ncbi.nlm.nih.gov/pubmed/27771735
554 Posey, J. E., J. A. Rosenfeld, R. A. James, M. Bainbridge, Z. Niu, X. Wang, S. Dhar, W. Wiszniewski, Z. H.
555 Akdemir, T. Gambin, F. Xia, R. E. Person, M. Walkiewicz, C. A. Shaw, V. R. Sutton, A. L. Beaudet, D.
556 Muzny, C. M. Eng, Y. Yang, R. A. Gibbs, J. R. Lupski, E. Boerwinkle, and S. E. Plon. 2016. Molecular
557 diagnostic experience of whole-exome sequencing in adult patients. *Genet Med* 18: 678-685.
558 https://www.ncbi.nlm.nih.gov/pubmed/26633545
559 Prall, T. M., M. E. Graham, J. A. Karl, R. W. Wiseman, A. J. Ericsen, M. Raveendran, R. Alan Harris, D. M.
560 Muzny, R. A. Gibbs, J. Rogers, and D. H. O'Connor. 2017. Improved full-length killer cell immunoglobulin-
561 like receptor transcript discovery in Mauritian cynomolgus macaques. *Immunogenetics* 69: 325-339.
562 https://www.ncbi.nlm.nih.gov/pubmed/28343239
563 Reynolds, M. R., J. B. Sacha, A. M. Weiler, G. J. Borchardt, C. E. Glidden, N. C. Sheppard, F. A. Norante, P. A.
564 Castrovinci, J. J. Harris, H. T. Robertson, T. C. Friedrich, A. B. McDermott, N. A. Wilson, D. B. Allison, W.
565 C. Koff, W. E. Johnson, and D. I. Watkins. 2011. The TRIM5{alpha} genotype of rhesus macaques affects
566 acquisition of simian immunodeficiency virus SIVsmE660 infection after repeated limiting-dose intrarectal
567 challenge. *J Virol* 85: 9637-9640. https://www.ncbi.nlm.nih.gov/pubmed/21734037

568 Robinson, J., J. A. Halliwell, H. McWilliam, R. Lopez, and S. G. Marsh. 2013. IPD--the Immuno Polymorphism
569 Database. *Nucleic Acids Res* 41: D1234-40. https://www.ncbi.nlm.nih.gov/pubmed/23180793
570 Semler, M. R., R. W. Wiseman, J. A. Karl, M. E. Graham, S. M. Gieger, and D. H. O'Connor. 2018. Novel full-
571 length major histocompatibility complex class I allele discovery and haplotype definition in pig-tailed
572 macaques. *Immunogenetics* 70: 381-399. https://www.ncbi.nlm.nih.gov/pubmed/29134258
573 Shiina, T., A. Blancher, H. Inoko, and J. K. Kulski. 2017. Comparative genomics of the human, macaque and mouse
574 major histocompatibility complex. *Immunology* 150: 127-138.
575 https://www.ncbi.nlm.nih.gov/pubmed/27395034
576 Silver, Z. A., and D. I. Watkins. 2017. The role of MHC class I gene products in SIV infection of macaques.
577 *Immunogenetics* 69: 511-519. https://www.ncbi.nlm.nih.gov/pubmed/28695289
578 Vallender, E. J. 2011. Expanding whole exome resequencing into non-human primates. *Genome Biol* 12: R87.
579 https://www.ncbi.nlm.nih.gov/pubmed/21917143
580 Weatherall, D. 2006. The use of non-human primates in research. 147. https://mrc.ukri.org/documents/pdf/the-use-
581 of-non-human-primates-in-research/ . Accessed 04/30/2019
582 Williams, T. M. 2001. Human leukocyte antigen gene polymorphism and the histocompatibility laboratory. *J Mol*
583 *Diagn* 3: 98-104. https://www.ncbi.nlm.nih.gov/pubmed/11486048
584 Wiseman, R. W., J. A. Karl, B. N. Bimber, C. E. O'Leary, S. M. Lank, J. J. Tuscher, A. M. Detmer, P. Bouffard, N.
585 Levenkova, C. L. Turcotte, E. J. Szekeres, C. Wright, T. Harkins, and D. H. O'Connor. 2009. Major
586 histocompatibility complex genotyping with massively parallel pyrosequencing. *Nat Med* 15: 1322-1326.
587 https://www.ncbi.nlm.nih.gov/pubmed/19820716
588 Wiseman, R. W., J. A. Karl, P. S. Bohn, F. A. Nimityongskul, G. J. Starrett, and D. H. O'Connor. 2013. Haplessly
589 hoping: macaque major histocompatibility complex made easy. *ILAR J* 54: 196-210.
590 https://www.ncbi.nlm.nih.gov/pubmed/24174442
591 Wittig, M., J. A. Anmarkrud, J. C. Kassens, S. Koch, M. Forster, E. Ellinghaus, J. R. Hov, S. Sauer, M. Schimmler,
592 M. Ziemann, S. Gorg, F. Jacob, T. H. Karlsen, and A. Franke. 2015. Development of a high-resolution NGS-
593 based HLA-typing and analysis pipeline. *Nucleic Acids Res* 43: e70.
594 https://www.ncbi.nlm.nih.gov/pubmed/25753671

595 Xie, C., Z. X. Yeo, M. Wong, J. Piper, T. Long, E. F. Kirkness, W. H. Biggs, K. Bloom, S. Spellman, C. Vierra-
596 Green, C. Brady, R. H. Scheuermann, A. Telenti, S. Howard, S. Brewerton, Y. Turpaz, and J. C. Venter.
597 2017. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc Natl*
598 *Acad Sci U S A* 114: 8059-8064. https://www.ncbi.nlm.nih.gov/pubmed/28674023
599 Xue, C., M. Raveendran, R. A. Harris, G. L. Fawcett, X. Liu, S. White, M. Dahdouli, D. Rio Deiros, J. E. Below, W.
600 Salerno, L. Cox, G. Fan, B. Ferguson, J. Horvath, Z. Johnson, S. Kanthaswamy, H. M. Kubisch, D. Liu, M.
601 Platt, D. G. Smith, B. Sun, E. J. Vallender, F. Wang, R. W. Wiseman, R. Chen, D. M. Muzny, R. A. Gibbs, F.
602 Yu, and J. Rogers. 2016. The population genomics of rhesus macaques (Macaca mulatta) based on whole-
603 genome sequences. *Genome Res* 26: 1651-1662. https://www.ncbi.nlm.nih.gov/pubmed/27934697
604 Yang, Y., D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu,
605 M. Hardison, R. Person, M. R. Bekheirnia, M. S. Leduc, A. Kirby, P. Pham, J. Scull, M. Wang, Y. Ding, S. E.
606 Plon, J. R. Lupski, A. L. Beaudet, R. A. Gibbs, and C. M. Eng. 2013. Clinical whole-exome sequencing for
607 the diagnosis of mendelian disorders. *N Engl J Med* 369: 1502-1511.
608 https://www.ncbi.nlm.nih.gov/pubmed/24088041
609 Yang, Y., D. M. Muzny, F. Xia, Z. Niu, R. Person, Y. Ding, P. Ward, A. Braxton, M. Wang, C. Buhay, N.
610 Veeraraghavan, A. Hawes, T. Chiang, M. Leduc, J. Beuten, J. Zhang, W. He, J. Scull, A. Willis, M.
611 Landsverk, W. J. Craigen, M. R. Bekheirnia, A. Stray-Pedersen, P. Liu, S. Wen, W. Alcaraz, H. Cui, M.
612 Walkiewicz, J. Reid, M. Bainbridge, A. Patel, E. Boerwinkle, A. L. Beaudet, J. R. Lupski, S. E. Plon, R. A.
613 Gibbs, and C. M. Eng. 2014. Molecular findings among patients referred for clinical whole-exome
614 sequencing. *JAMA* 312: 1870-1879. https://www.ncbi.nlm.nih.gov/pubmed/25326635
615 Zimin, A. V., A. S. Cornish, M. D. Maudhoo, R. M. Gibbs, X. Zhang, S. Pandey, D. T. Meehan, K. Wipfler, S. E.
616 Bosinger, Z. P. Johnson, G. K. Tharp, G. Marcais, M. Roberts, B. Ferguson, H. S. Fox, T. Treangen, S. L.
617 Salzberg, J. A. Yorke, and R. B. J. Norgren. 2014. A new rhesus macaque assembly and annotation for next-
618 generation sequencing analyses. *Biol Direct* 9: 20. https://www.ncbi.nlm.nih.gov/pubmed/25319552
619
620
621

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

Animal ID	SRA Accession	Total Raw Reads	MHC-I Reads Extracted	DRB1/3/4/5 Reads Extracted	DQA1 Reads Extracted	DQB1 Reads Extracted	DPA1 Reads Extracted	DPB1 Reads Extracted	MHC Reads Extracted	MHC % of total
r05029	SAMN11131055	87,345,430	79,712	215,654	24,266	39,188	29,604	112,376	500,800	0.57
r17099	SAMN11131056	67,632,236	32,406	89,946	9,694	16,782	11,214	47,974	208,016	0.31
r07010	SAMN11173514	81,062,856	71,694	204,352	22,528	38,382	25,392	111,626	473,974	0.58
r17041	SAMN11131058	80,004,430	30,852	244,326	9,746	16,310	12,644	151,376	465,254	0.58
r17061	SAMN11131059	78,428,820	70,082	204,694	20,776	41,018	19,440	95,538	451,548	0.58
J8R	SAMN11282383	80,868,782	45,164	87,864	11,878	19,736	13,810	57,146	235,598	0.29
J01	SAMN11282382	61,009,650	34,342	87,124	9,050	16,628	10,958	46,964	205,066	0.34
ZC08	SAMN11282384	60,693,394	31,168	81,586	10,392	15,660	10,172	45,748	194,726	0.32
DGKG	SAMN11282378	80,530,772	48,658	130,256	12,788	22,682	13,372	62,990	290,746	0.36
CF18	SAMN11282368	68,593,710	37,792	96,842	10,864	18,226	13,118	51,556	228,398	0.33
DE1AA	SAMN11282369	65,712,570	39,334	108,002	10,060	17,124	11,840	51,134	237,494	0.36
DEG8	SAMN11282370	68,490,344	35,566	116,824	10,404	19,278	12,294	55,086	249,452	0.36
DEXX	SAMN11282371	67,451,626	33,900	128,128	10,442	18,898	9,778	53,258	254,404	0.38
DF24	SAMN11282372	67,768,724	41,362	96,172	10,606	18,176	10,976	50,612	227,904	0.34
DF64	SAMN11282373	65,762,818	37,426	107,974	11,222	15,632	11,446	51,214	234,914	0.36
DF6T	SAMN11282374	71,472,222	41,160	101,102	11,052	18,794	13,492	51,626	237,226	0.33
DFET	SAMN11282375	77,040,660	49,068	128,862	12,588	20,952	12,586	58,238	282,294	0.37
DFV0	SAMN11282377	73,877,184	41,924	98,866	11,630	17,690	13,952	52,394	236,456	0.32
DJ94	SAMN11282379	66,051,412	39,616	88,196	9,910	15,892	10,978	45,972	210,564	0.32
HIH	SAMN11282380	76,491,578	41,308	85,732	11,148	18,366	14,102	50,746	221,402	0.29
DFJT	SAMN11282376	62,249,064	37,728	78,860	9,666	15,268	11,426	43,086	196,034	0.31
HLJ	SAMN11282381	68,288,558	41,748	94,570	9,960	17,240	12,756	48,356	224,630	0.33
37360	SAMN11282363	62,310,824	38,200	97,452	9,866	15,024	10,320	46,140	217,002	0.35
0L7	SAMN11282364	64,145,834	35,302	128,778	9,310	19,992	10,686	54,196	258,264	0.40
0R1	SAMN11282365	64,789,208	36,462	86,772	10,656	17,258	11,982	46,124	209,254	0.32
0R8	SAMN11282366	72,836,490	38,286	118,184	10,500	22,248	13,170	56,142	258,530	0.35
0TI	SAMN11282367	80,730,750	41,552	106,356	13,364	20,714	14,860	57,744	254,590	0.32
	Average	**70,549,789**	**42,660**	**119,018**	**12,014**	**20,487**	**13,569**	**61,310**	**269,057**	**0.37**

Table 2. Summary of discordant results for each MHC genotyping method by animal

	MHC class I				MHC class II			
Animal ID	**Total Alleles Expected**	**# MiSeq Discordance**	**# DSR Discordance**	**# SAVAGE Discordance**	**Total Alleles Expected**	**# MiSeq Discordance**	**# DSR Discordance**	**# SAVAGE 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	0.3
Concordance (%)		90.4	97.3	98.3		97.8	100.0	99.7

Relationship Animal_ID	Dam r05029			Progeny 1 r17099			Sire r07010			Progeny 2 r17041			Progeny 3 r17061			
Assay	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	MiSeq	DSR	SAVAGE	Ambiguous Allele Group
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_alleles*
*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_alleles*
*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_alleles,B*063:02:02_allele*
*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_alleles,B11L*01:04,B11L*01:05*
*Mamu-B16*01:01*				44	34	1	72	61	1				110	29	1	
*Mamu-B16*01g1*	181	74	1							176	23	1				*B16*01_alleles*
*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*

Fig. 1 Comparison of MHC class I results from MiSeq PCR amplicon versus whole exome genotyping strategies for a representative breeding group of rhesus macaques

Relationship	**Dam**			**Progeny 1**			**Sire**			**Progeny 2**			**Progeny 3**			
Animal_ID	**r05029**			**r17099**			**r07010**			**r17041**			**r17061**			
Assay	MiSeq DSR SAVAGE			MiSeq DSR SAVAGE			MiSeq DSR SAVAGE			MiSeq DSR SAVAGE			MiSeq DSR SAVAGE			Ambiguous Allele Group
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*04:05
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*02:04:02
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	

Fig. 2 Comparison of MHC class II results from MiSeq PCR amplicon versus whole exome genotyping strategies for a representative breeding group of rhesus macaques

A

Relationship		Dam	Progeny 1	Sire	Progeny 2	Progeny 3	
GS ID	Ave Length	r05029	r17099	r07010	r17041	r17061	Ambiguous Allele Groups
MHC Haplotype 1		a	a	b	c	b	
MHC Haplotype 2		d	b	c	e	f	
Mamu-A Major Alleles							
*Mamu-A1*001g2*	1,290		1208	1397		1266	*A1*001:01,A1*001:02*
*Mamu-A1*002:01*	1,421	1486				1356	
*Mamu-A1*004g1*	1,264	1271	1257				*A1*004_6_alleles*
*Mamu-A1*008g1*	1,189			1248	1130		*A1*008_5_alleles*
*Mamu-A1*023_A1*106g1*	1,123				1123		*A1*023:01,A1*023:02,A1*106:01*
Mamu-A Minor Alleles							
*Mamu-A2*05g1*	1,235		1293	789	1342	1515	*A2*05_32_alleles*
*Mamu-A3*13_A4*14g1*	1,366	1251				1481	*A3*13_7_alleles,A4*14_1_allele*
*Mamu-A3*13g1*	973			717	1228		*A3*13:03,A3*13:11*
*Mamu-A4*14g1*	1,264	1379	1197		1215		*A4*14_14_alleles*
Mamu-E Alleles							
*Mamu-E*02g1*	1,044	935	1515	1214	628	928	*E*02_9_alleles*
*Mamu-E*02g4*	768	750				785	*E*02:13,E*02:21,E*02:29*
*Mamu-E*02g5*	985	973				997	*E*02:04,E*02:06*
*Mamu-E*02g7*	1,408			1579	1236		*E*02:12:01,E*02:28*
*Mamu-E*02g8*	1,326	1445	1459	1394		1006	*E*02:20,E*02:30*
Mamu-B Major Alleles							
*Mamu-B*001g1*	1,228			1382	1074		*B*001:01:01,B*001:04,B*001:05,B*001:06*
*Mamu-B*007g1*	1,310			1369	1251		*B*007_6_alleles*
*Mamu-B*012:01*	1,210					1210	
Mamu-B_022g1	1,247					1247	*B*022:01,B*022:02*
*Mamu-B*027g1*	1,209				1209		*B*027:03,B*027:04*
*Mamu-B*030g1*	629		550	-	459	878	*B*030_6_alleles*
*Mamu-B*030g2*	636			797	475		*B*030:02,B*030:05*
*Mamu-B*031:03*	820		563	1382		514	
*Mamu-B*031g1*	867					867	*B*031:01,B*031:02*
*Mamu-B*041g1*	1,227	1302	1152				*B*041:01,B*041:02*
*Mamu-B*043:01*	1,284		1256	1276	1291	1311	
*Mamu-B*048:01*	1,325	1427	1223				
*Mamu-B*052:01*	1,434	1434					
*Mamu-B*055:01*	1,214	1214					
*Mamu-B*058:02*	1,250	1250					
*Mamu-B*064:01*	1,332	1441	1222				
*Mamu-B*073g1*	1,209		1211	1287		1130	*B*073:01,B*073:02*
*Mamu-B*074g1*	1,188	1357			1011	1197	*B*074:01,B*074:02,B*074:03*
Mamu-B Minor Alleles							
*Mamu-B*035_B*049g1*	1,363	1398			1230	1460	*B*035:01,B*035:02,B*049:01*
*Mamu-B*053g1*	1,275					1275	*B*053:02,B*053:03*
*Mamu-B*054:01*	1,341	1341					
*Mamu-B*054:04*	1,377				1377		
*Mamu-B*054:05*	1,177					1177	
*Mamu-B*057:01*	718					718	
*Mamu-B*057:06*	1,039		1157	1290		671	
*Mamu-B*057g1*	1,165				1165		*B*057:02,B*057:03,B*057:04,B*057:05*
*Mamu-B*063:03*	1,259	1285	1257	1298	1197		
*Mamu-B*063g1*	1,376	1376					*B*063:01:01,B*063:01:02,B*063:02:02*
*Mamu-B*070g2*	1,354	1395				1313	*B*070:01,B*070:06*
*Mamu-B*072g1*	1,301		1294	1443	1222	1246	*B*072_8_alleles*
*Mamu-B*092g1*	1,293		1249	1306	1245	1372	*B*092:01,B*092:02,B*092:03*
*Mamu-B*098g1*	1,322	1427			1180	1358	*B*098_9_alleles*
*Mamu-B*109:06*	1,198	1315	1080				
*Mamu-B*134g2*	1,044	1390	698				*B*134:03,B*134:04,B*134:06*
*Mamu-B*188g1*	1,303			1274	1331		*B*188:01,B*188:02*
*Mamu-B11L*01:12*	1,188		-	1254		1121	
*Mamu-B11L*01g1*	958	1363	551	1268	650		*B11L*01:01:02,B11L*01:02,B11L*01:04,B11L*01:05*
*Mamu-B16*01:01*	621		593	525		746	
*Mamu-B16*01g1*	988	1052			924		*B16*01:04,B16*01:05,B16*01:07,B16*01:08*
*Mamu-B16*01g2*	636			636	-		*B16*01:02,B16*01:09*
*Mamu-B17*01g1*	513				409	616	*B17*01_14_alleles*
*Mamu-B17*01g2*	1,035	1446	962	1358	372		*B17*01_9_alleles*
*Mamu-B20*01:02*	615				615		
*Mamu-B20*01g1*	1,126	1379	1183	1378	565		*B20*01:01,B20*01:04*
Mamu-I Alleles							
*Mamu-I*01g1*	944	1397	1327	780	413	803	*I_01_66_alleles*
*Mamu-I*01g3*	293				293		*I*01:19,I*01:20:01*

B

Relationship		Dam	Progeny 1	Sire	Progeny 2	Progeny 3	
GS ID	Ave Length	r05029	r17099	r07010	r17041	r17061	Ambiguous Allele Groups
MHC Haplotype 1		a	a	b	c	b	
MHC Haplotype 2		d	b	c	e	f	
Mamu-DRB Alleles							
*Mamu-DRB1*03:09*	1,763	3384	761	1053	559	3057	
*Mamu-DRB1*03g1*	866		1131	768		700	*DRB1*03:06,DRB1*03:26*
*Mamu-DRB1*10:03*	653		631	710		618	
*Mamu-DRB*W2g1*	591	655	520	556	613	612	*DRB*W2:01,DRB*W2:05*
*Mamu-DRB*W3g1*	2,213	1430			2996		*DRB*W3:03:01,DRB*W3:03:02*
*Mamu-DRB*W4:01*	691	644			737		
Mamu-DQA/DQB Alleles							
*Mamu-DQA1*01:02*	1,942		2087	1475		2264	
*Mamu-DQA1*01g1*	1,677	1467	1988	1429	1980	1522	*DQA1*01:04:01,DQA1*01:04:02*
*Mamu-DQA1*23:01*	2,073	2070			2076		
*Mamu-DQB1*06:01*	2,820	2726	1334	3199	4063	2780	
*Mamu-DQB1*06g2*	2,436		1076	3257		2975	*DQB1*06:05,DQB1*06:18*
*Mamu-DQB1*18g4*	4,332	4336			4328		*DQB1*18:02,DQB1*18:27*
Mamu-DPA/DPB Alleles							
*Mamu-DPA1*02g1*	1,897	1989	1795	1929	1873		*DPA1*02_9_alleles*
*Mamu-DPA1*04g1*	1,985		1894	2051		2011	*DPA1*04:03:01,DPA1*04:03:02,DPA1*04:05*
*Mamu-DPA1*11:01*	1,842					1842	
*Mamu-DPB1*02g1*	3,693		3946	3073		4059	*DPB1*02:01,DPB1*02:04:01,DPB1*02:04:02*
*Mamu-DPB1*15g1*	2,687	2470	2506	3190	2581		*DPB1*15:01,DPB1*15:02*
*Mamu-DPB1*16:01*	827					827	

C

MHC Haplotype	a	b	c	d	e	f
Mamu-A Haplotype	A004	A001	A008	A002a	A023	A002a
Mamu-B Haplotype	B048	B043b	B001a	B055	B043a	B012b
Mamu-DRB Haplotype	DR04a	DR03f	DR04a	DR06	DR06	DR04a
Mamu-DQA	01g1	01:02	01g1	23:01	23:01	01g1
Mamu-DQB	06:01	06g2	06:01	18g4	18g4	06:01
Mamu-DPA	02g1	04g	02g1	02g1	02g1	11:01
Mamu-DPB	15g	02g	15g	15g	15g	16:01

Fig. 3 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 a 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 Haplotypes in this breeding group are indicated with shades of grey (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.
(C) Abbreviated Mamu haplotype designations are summarized for the sx extended MHC haplotypes identified by segregation in this breeding group. For example, the twenty MHC sequences (red) that are associated with extended Haplotype a can be summarized by the following string of abbreviated Mamu haplotype designations: Mamu-A004/B048/DR04a/DQA01g1/DQB06:01/DPA02g1/DPB15g (Karl et al 2013).

Supplementary Table 1. Fraction of exome sequence reads corresponding to MHC class I and class II genes after target capture with the VCRome2.1 probe design alone

GS ID	SRA Accession	Total Raw Reads	MHC-I Reads Extracted	DRB1/3/4/5 Reads Extracted	DQA1 Reads Extracted	DQB1 Reads Extracted	DPA1 Reads Extracted	DPB1 Reads Extracted	MHC Reads Extracted	MHC % of total
cy0642	**SAMN11281734**	79,795,404	4,848	26,822	2,568	2,274	900	81,340	118,752	0.15
cy0643	**SAMN11281735**	86,265,844	4,408	26,930	2,542	2,320	936	74,370	111,506	0.13
cy0646	**SAMN11281736**	109,909,444	4,724	25,476	2,382	2,196	910	72,206	107,894	0.1
cy0648	**SAMN11281737**	106,315,304	5,558	34,442	2,966	2,714	1,092	97,254	144,026	0.14
cy0649	**SAMN11281738**	78,575,038	6,022	33,742	3,176	2,970	1,134	90,556	137,600	0.18
cy0651	**SAMN11281739**	102,890,862	6,686	33,194	3,388	3,094	1,294	94,532	142,188	0.14
cy0652	**SAMN11281740**	89,074,544	4,652	25,314	2,660	2,478	976	69,976	106,056	0.12
cy0654	**SAMN11281741**	99,169,798	5,294	30,328	2,924	2,756	1,042	79,664	122,008	0.12
	Average	**93,999,530**	**5,274**	**29,531**	**2,826**	**2,600**	**1,036**	**82,487**	**123,754**	**0.13**