1 Title: No assembly required: full-length MHC class I allele discovery by PacBio circular
2 consensus sequencing

3
4 Author names and affiliations:

5 Catherine J. Westbrook^a, Julie A. Karl^b, Roger W. Wiseman^b, Suzanne Mate^a, Galina
6 Koroleva^a, Karla Garcia^a, Mariano Sanchez-Lockhart^{a,d}, David H. O'Connor^{b,c}, and
7 Gustavo Palacios^a

8 ^aU.S. Medical Research Institute of Infectious Diseases, 1425 Porter Street, Fort
9 Detrick, Maryland 21702.

10 ^bWisconsin National Primate Research Center, 555 Science Drive, University of
11 Wisconsin-Madison, Madison, Wisconsin 53711.

12 ^cDepartment of Pathology and Laboratory Medicine, 1685 Highland Ave., University of
13 Wisconsin-Madison, Madison, Wisconsin 53705.

14 ^dCorresponding author: Mariano Sanchez-Lockhart

15 Mariano.sanchez-lockhart.ctr@mail.mil

16 1425 Porter St. Frederick, MD 21702.

17 Phone: (301) 619-3279.

Abstract

Single-molecule real-time (SMRT) sequencing technology with the Pacific Biosciences (PacBio) RS II platform offers the potential to obtain full-length coding regions (~1,100-bp) from MHC class I cDNAs. Despite the relatively high error rate associated with SMRT technology, high quality sequences can be obtained by circular consensus sequencing (CCS) due to the random nature of the error profile. In the present study we first validated the ability of SMRT-CCS to accurately identify class I transcripts in Mauritian-origin cynomolgus macaques (*Macaca fascicularis*) that have been characterized previously by cloning and Sanger-based sequencing as well as pyrosequencing approaches. We then applied this SMRT-CCS method to characterize 60 novel full-length class I transcript sequences expressed by a cohort of cynomolgus macaques from China. The SMRT-CCS method described here provides a straightforward protocol for characterization of unfragmented single-molecule cDNA transcripts that will potentially revolutionize MHC class I allele discovery in nonhuman primates and other species.

Abbreviations: SMRT – single-molecule real-time; PacBio – Pacific Biosciences; CCS – circular consensus sequencing; MCM – Mauritian cynomolgus macaques; ChCM – Chinese cynomolgus macaques; IPD – Immuno Polymorphism Database

Keywords: MHC class I; cynomolgus macaques; single-molecule real-time circular consensus sequencing (SMRT-CCS); PacBio

1. Introduction

Cynomolgus macaques (*Macaca fascicularis*) and other related macaque species are important animal models for infectious disease research [1-3]. Because the expression of specific major histocompatibility complex (MHC) class I alleles have been associated

with susceptibility or resistance to certain infectious diseases, significant effort has been invested in characterizing and investigating the role of MHC class I alleles in macaque populations used in research protocols [4-6]. The proteins encoded by MHC class I genes perform the critical function of presenting foreign peptides of intracellular origin to cytotoxic T cells [7]. Humans express three highly polymorphic classical class I genes per haplotype (*HLA-A*, *HLA-B* and *HLA-C*). Though macaques lack an apparent *HLA-C* equivalent, the gene counterparts to *HLA-A* and *HLA-B* have undergone multiple rounds of duplication, deletion, and recombination resulting in a significant expansion of the MHC class I loci. Macaque *MHC-A* and *MHC-B* haplotypes contain variable numbers of loci that exhibit a wide range of steady state RNA levels [8, 9]. This makes the macaque MHC class I region more difficult to characterize than its human counterpart.

The advent of massively parallel DNA sequencing technologies has revolutionized the ability to identify the complement of expressed MHC class I alleles in an individual macaque [10-13]. However, the short read lengths produced by most next-generation sequencing platforms remains a barrier to obtaining the full-length, ~1,100-bp MHC class I open reading frame from a single molecule. MHC class I macaque sequencing assays developed on the Roche/454 platform with a 568-bp cDNA amplicon [12,13] and on the Illumina MiSeq with a 195-bp amplicon [4] provide accurate lineage-level genotyping of macaques by sequencing the highly polymorphic peptide-binding region. However, since there are often multiple allelic sequence variants within a lineage group [13,14] and there is evidence that specific allelic variants may have an impact on disease susceptibility and progression [15-17], it is important to expand the MHC class I full-length allele databases for macaque animal models. Previously described methods to characterize the full-length MHC class I allele sequences in macaques involve cloning and Sanger-based sequencing cDNA-PCR products [9,18] or assembling overlapping Roche/454 amplicons or fragmented Illumina MiSeq reads extended from a Roche/454 scaffold [11,13]. Although successful, these methods are resource intensive and will be obsolete when Roche ends support of the 454 platform in 2016.

Here we present a novel full-length MHC class I allele discovery approach using circular consensus sequencing (CCS) with single-molecule real-time (SMRT) technology on the

PacBio RS II [19, 20]. Initially, we validated the accuracy of the PacBio SMRT-CCS method by analyzing Mauritian cynomolgus macaques (MCM) that express previously characterized, full-length MHC class I transcripts. Although the PacBio RS II has a relatively high inherent sequencing error rate, the errors are random and can therefore be overcome by high sequence redundancy. After validation, we utilized the PacBio approach to characterize novel full-length MHC class I allele transcripts expressed in a cohort of cynomolgus macaques from a Chinese breeding facility (ChCM). We believe SMRT-CCS on the PacBio RS II system provides a clear advantage over other next generation sequencing platforms for characterizing full-length MHC class I cDNA transcripts because sequencing is achieved in one continuous read. Since the ORF amplicon is not fragmented prior to sequencing, this method eliminates the ambiguity introduced by assembly of short, highly-similar allele fragments.

2. Materials and methods

2.1. *Animals*

Five representative MCM were selected for the validation study based on microsatellite analysis [4]. These individuals were homozygous or heterozygous for the four most common MHC class I haplotypes in this population (M1-M4), and full-length MHC class I transcripts have been characterized previously by both Sanger-based cloning and sequencing and Roche/454 overlapping amplicons [4, 11]. Similarly, ten ChCM were selected for full-length MHC-I allele discovery based on Roche/454 genotyping results identifying a high number of potential novel allelic variants (see Section 2.2). Although these cynomolgus macaques were bred in a Chinese facility, mitochondrial sequencing results were consistent with mainland Southeast Asian origin for their ancestors [21]. All animals were cared for according to the regulations and guidelines of the Institutional Care and Use Committees at their respective institutions.

2.2. MHC genotyping

Total cellular RNA was isolated from fresh whole blood samples and converted to cDNA using a Superscript III First-Strand Synthesis System (Life Technologies). For the MCM samples, cDNA-PCR was performed with high-fidelity Phusion polymerase (New England BioLabs) and gene-specific primers targeting conserved sequences that flank the highly polymorphic peptide binding domain encoded by exon 2, allowing simultaneous amplification of 195-bp amplicons from all MHC class I transcripts expressed by each macaque (Supplementary Figure 1a). The class I-specific primers shared common sequences on their 5' ends for a 4-primer amplicon tagging strategy; the MHC-specific amplicons with incorporated consensus sequences were combined during PCR with an 8-fold excess of sample-specific primer pairs containing one of 384 unique 10-bp barcoding sequences and the Illumina PE1 or PE2 adaptor sequences required for cluster generation by bridge PCR [22]. After purification with AMPure PB beads (Agencourt), amplicons were normalized to equimolar concentrations and pooled for 250-bp paired-end sequencing on a MiSeq instrument (Illumina).

ChCM samples were genotyped by pyrosequencing 568-bp cDNA-PCR amplicons on a GS FLX instrument (Roche/454) as described previously [4, 12]. MHC genotypes were determined using custom pipelines with FLEXBAR [23] and FLASH [24] followed by Bowtie [25] mapping of the processed sequencing reads to a curated reference database of cynomolgus macaque MHC sequences (Mafa_MHC_mRNA-allseq-13.09.01.fasta) that is publicly available on the Nonhuman Primate MHC Contract Web Portal (<http://go.wisc.edu/173j30>). Sequence reads identical to known cynomolgus macaque alleles over the region covered by each amplicon were enumerated and exported to Excel for further analyses. These workflows are implemented in BASH scripts with Applescript wrappers and are available upon request.

2.3 PacBio RS II sequencing

To amplify the complete MHC class I ORF, cDNA-PCR was conducted using gene-specific primer sequences shown in Supplementary Figure 1b [11]. Two independent

cDNA-PCRs were performed for each animal with distinct forward primers plus a pool of three reverse primers in the 3'UTR (Supplementary Figure 1b). A unique 16-bp PacBio barcode was added to the 5' end of each primer so that all amplicons from a specific animal were identified with the same 5' to 3' barcode sequence, from a set of 384 barcodes custom designed for the PacBio system (<http://www.smrtcommunity.com/Share/Protocol?id=a1q7000000J4m5AAC&strRecordTypeName=Protocol>).

Amplified DNA was purified using a 1X volume AMPure PB Beads, and yields and size distributions were measured using the dsDNA BR assay kit on the Qubit Fluorometer (Life Technologies) and Agilent 2100 Bioanalyzer DNA7500 kit. Amplicons were processed without shearing and SMRTbell libraries were constructed following the PacBio 2-kb Template Preparation protocol for CCS, allowing a single insert to be sequenced multiple times [20]. Briefly, primary PCR products were end-repaired and hairpin adapters were added via blunt end ligation to produce SMRTbell templates using the PacBio DNA Template Prep Kit 2.0. Exonucleases III and VII were added to remove failed ligation products and SMRTbells were purified with a 1X volume of AMPure PB beads. Final library assessments for each barcoded sample were again obtained using the dsDNA BR assay and the DNA7500 kit. Five to ten individual barcoded samples were then pooled in equimolar concentrations for multiplex sequencing. The ratio of sequencing primer and polymerase was determined by a PacBio calculator, entering in SMRTbell concentrations and the ~1,200-bp insert size. The sequencing primer was annealed to the single-stranded loop region of the SMRTbell template, and primer-annealed templates were then bound to DNA polymerase P4. The polymerase-template complexes diffused into and were immobilized at the bottom of the zero-mode waveguides. The pooled SMRTbell libraries for each cohort of barcoded samples were sequenced on a PacBio RS II system with C2 sequencing chemistry, using multiple SMRT cells run at two different on-plate loading concentrations (150pM and 200pM). A 2-kb collection protocol was utilized with a 180 minute movie window for signal detection.

Base calling of the raw sequence data, filtering for length and quality, and CCS read

generation was performed through the SMRT Portal using PacBio's SMRTAnalysis software (v2.2.0). Reads were separated by barcode, and only CCS reads making five or greater full passes around the closed loop SMRTbell amplicon were retained (~99% accuracy). Allele characterization was performed with Geneious Pro software (v7.1.4) (Biomatters Limited). Fastq files of the CCS reads were imported into Geneious and sequences between 1,000-bp and 1,200-bp were extracted. CCS reads from each animal were de novo assembled at a minimum overlap identity of 99% over 400-bp, allowing gaps, with 1% maximum mismatches per read. BLAST was used to compare consensus sequences from the resulting contigs of 4 or more reads against a custom database containing all known cynomolgus macaque MHC alleles (see Section 2.2). Novel alleles and extensions of known alleles identified for each animal were submitted to GenBank (accession nos. **KP307842-KP307910**) and the Immuno Polymorphism Database (IPD) to obtain official nomenclature designations [14, 26].

3. Results and Discussion

3.1. *MCM validation study*

Since the MCM population arose from a very small number of founders introduced to a geographically isolated island, these macaques have extremely restricted MHC diversity [4]. The class I diversity of the entire population can be described by seven distinct haplotypes, and the full-length sequences for the transcripts associated with each haplotype are known [4, 11]. Five MCM samples containing the four most abundant haplotypes, M1-M4, were used to validate the accuracy of the SMRT-CCS method.

Circular consensus sequencing of the MHC class I transcripts of the five barcoded MCM on the PacBio RS II produced 111,324 reads that passed initial filtering (length ≥ 50 and quality ≥ 0.75), and 22,028 reads represented five or more passes of the insert with an average pass number of 11.5 and a quality score of 99.4. Sequence yields per barcode are shown in Table 1. After de novo assembly and alignment, a total of 11,848 CCS reads were identified. These sequences were identical to the expected MHC class I alleles for these individuals. The total reads representing each specific allele are summarized in Figure 1 as percentages of total MHC class I PacBio reads in each

sample, along with the expected allele content of each sample as determined by Illumina MiSeq genotyping.

The MiSeq results provide a semi-quantitative estimate of relative steady-state transcript abundance for the alleles on each MCM haplotype. As previously stated, macaque *MHC-A* and *MHC-B* haplotypes contain variable numbers of loci with different levels of transcript expression; MHC class I alleles can be operationally categorized as either “major” or “minor” alleles, with “major” alleles typically exceeding an average 4% transcript abundance for MHC heterozygous individuals in the population [12]. Accurate full-length MHC class I allele sequences were recovered by the SMRT-CCS method for all 15 of the “major” alleles present in the four MCM haplotypes shown in Figure 1. Ten additional “minor” alleles were also recovered in at least one of the five MCM samples. As summarized in Figure 1, the MiSeq platform yielded close to a ten-fold increase in read number with 99,792 MHC class I reads produced from the same cohort of MCMs. This greater depth of coverage allowed for detection of more of the “minor” alleles with the MiSeq compared to the SMRT-CCS protocol; it should be possible to detect more “minor” alleles with PacBio CCS by deeper sequencing of the same samples. Likewise, additional improvements in the PacBio polymerase and chemistry hold the promise of more comprehensive coverage compared to the P4C2 reagents available for this study. Most important to the MCM validation study is that the 25 full-length MHC class I alleles recovered were identical to the previous Sanger and Roche/454 results.

3.2. ChCM full-length MHC class I allele discovery

With the accuracy of the SMRT-CCS method confirmed, PacBio RS II sequencing was performed on ten ChCM samples selected to express multiple novel class I transcripts. Circular consensus sequencing produced 216,565 reads that passed initial filtering and 55,446 reads representing five or more passes of the insert (average pass number of 12.2 with a quality score of 99.1). Yields per individual barcode are shown in Table 1. The MHC class I genotypes determined by Roche/454 pyrosequencing for these ChCM samples are shown in Figure 2. As described in Section 3.1, these results are

normalized as the percentage of total MHC class I reads identified for each allele within an animal and provide a semi-quantitative estimate of relative transcript abundance. Alleles that were also detected in four or more reads with the SMRT-CCS method are highlighted with grey boxes in Figure 2.

Overall, 70 functionally full-length sequences containing at least the leader peptide region through the stop codon were obtained from the PacBio RS II for these ten individuals. Sixty of these unique *Mafa-A* and *Mafa-B* variants, listed in Table 2, are either newly reported, match previously described sequences that lacked official allele nomenclature, or are extensions of prior named alleles that were not formerly full-length. The remaining ten transcripts were not included in Table 2 since they were identical to previously described full-length sequences already in GenBank.

As with the MCM cohort, essentially all of the “major” alleles and several of the “minor” alleles were detected by the SMRT-CCS method at the depth of coverage obtained in this study. Additional optimization, sequencing the same samples on additional SMRT cells, or improvements to the PacBio polymerase and chemistry may increase depth of coverage and enhance the detection of alleles with lower transcriptional abundance. There is evidence, however, that CD8⁺ T cell responses are primarily restricted by the most highly transcribed MHC class I alleles in macaques [6], so full characterization of all “minor” alleles expressed in a population of macaques may not be required for most studies.

The SMRT-CCS method described here provides a straightforward approach for unambiguously sequencing unfragmented full-length single cDNA molecules like these MHC class I transcripts. Complete full-length MHC class I allele databases will allow for the development of more accurate genotyping tools for macaques and other model organisms. Genotyping individuals prior to the start of a study allows for balancing of MHC genotypes among experimental groups, while retrospective genotyping can inform data interpretation. Additionally, full-length MHC class I cDNA sequences are required for the creation of reagents such as MHC-peptide tetramers, which will enable researchers to assess the range and functionality of CD8⁺ T cell responses. This method could also be used for full-length sequencing of MHC class II and killer-cell

immunoglobulin-like receptor transcripts [27- 29], as well as other highly polymorphic immune loci.

Acknowledgements

This work was supported by Defense Threat Reduction Agency under USAMRIID project number 3174512. This study was also supported by the National Institute of Allergy and Infectious Diseases (HHSN272201100013C) and was conducted in part at a facility constructed with support from the Research Facilities Improvement Program (RR15459-01, RR020141-01).

The content of this publication does not necessarily reflect the views or policies of the U.S. Department of the Army, the U.S. Department of Defense, or the U.S. Department of Health and Human Services or of the institutions and companies with which the authors are affiliated.

Research was conducted under an IACUC approved protocol in compliance with the Animal Welfare Act, PHS Policy, and other Federal statutes and regulations relating to animals and experiments involving animals. The facility where this research was conducted is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care, International and adheres to principles stated in the Guide for the Care and Use of Laboratory Animals, National Research Council, 2011.

References

- [1] Goff A, Mucker E, Raymond J, Fisher R, Bray M, Hensley L, et al. Infection of cynomolgus macaques with a recombinant monkeypox virus encoding green fluorescent protein. *Archives Virol* 2011;156(10):1877-81.
- [2] Herbert AS, Kuehne AI, Barth JF, Ortiz RA, Nichols DK, Zak SE, et al. Venezuelan equine encephalitis virus replicon particle vaccine protects nonhuman primates from intramuscular and aerosol challenge with ebolavirus. *J Virol* 2013;87(9):4952-64.

288 [3] Watanabe T, Kiso M, Fukuyama S, Nakajima N, Imai M, Yamada S, et al.
289 Characterization of H7N9 influenza A viruses isolated from humans. *Nature*
290 2013;501(7468):551-5.

291 [4] Wiseman RW, Karl JA, Bohn PS, Nimityongskul FA, Starrett GJ, O'Connor DH.
292 Haplessly hoping: macaque major histocompatibility complex made easy. *ILAR J.*
293 2013;54:196-210.

294 [5] Loffredo JT, Maxwell J, Qi Y, Glidden CE, Borchardt GJ, Soma T, et al. Mamu-B* 08-
295 positive macaques control simian immunodeficiency virus replication. *J Virol* 2007;
296 81(16):8827-32.

297 [6] Budde ML., Lhost JJ, Burwitz BJ, Becker EA, Burns CM, O'Connor SL, et al.
298 Transcriptionally abundant major histocompatibility complex class I alleles are
299 fundamental to non-human primate SIV-specific CD8+ T cell responses. *J Virol* 2011;
300 85(7):3250-61.

301 [7] Parham P, Ohta T. Population biology of antigen presentation by MHC class I
302 molecules. *Science* 1996;272(5258):67-74.

303 [8] Daza-Vamenta R, Glusman G, Rowen L, Guthrie B, Geraghty DE. Genetic
304 divergence of the rhesus macaque major histocompatibility complex. *Genome Research*
305 2004;14(8):1501-15.

306 [9] Otting N, Heijmans CM, Noort RC, de Groot NG, Doxiadis GG, van Rood JJ, et al.
307 Unparalleled complexity of the MHC class I region in rhesus macaques. *Proc Natl Acad*
308 *Sci USA* 2005;102(5):1626-31.

309 [10] Wiseman RW, Karl JA, Bimber BN, O'Leary CE, Lank SM, Tuscher JJ, et al. Major
310 histocompatibility complex genotyping with massively parallel pyrosequencing. *Nat Med*
311 2009;15(11):1322-6.

312 [11] Budde ML, Wiseman RW, Karl JA, Hanczaruk B, Simen BB, O'Connor DH.
313 Characterization of Mauritian cynomolgus macaque major histocompatibility complex
314 class I haplotypes by high-resolution pyrosequencing. *Immunogenetics* 2010;62:773-80.

315 [12] Karl JA, Bohn PS, Wiseman RW, Nimityongskul FA, Lank SM, Starrett GJ,
 316 O'Connor DH. Major histocompatibility complex class I haplotype diversity in Chinese
 317 rhesus macaques. *G3* 2013;3:1195-1201.

318 [13] Dudley DM, Karl JA, Creager HM, Bohn PS, Wiseman, RW O'Connor DH. Full-
 319 length novel MHC class I allele discovery by next-generation sequencing: two platforms
 320 are better than one. *Immunogenetics* 2014;66:15-24.

321 [14] de Groot NG, Otting N, Robinson J, Blancher A, Lafont BA, Marsh SGE, et al.
 322 Nomenclature report on the major histocompatibility complex genes and alleles of Great
 323 Ape, Old and New World monkey species. *Immunogenetics* 2012;64:615-31.

324 [15] Budde ML, Greene JM, Chin EN, Ericson AJ, Scarlotta M, Cain BT, et al. Specific
 325 CD8+ T Cell Responses Correlate with Control of SIV Replication in Mauritian
 326 *Cynomolgus* Macaques. *J Virol* 2012;86(14):7596-604

327 [16] Kløverpris HN, Harndahl M, Leslie AJ, Carlson JM, Ismail N, van der Stok M, et al.
 328 HIV control through a single nucleotide on the HLA-B locus. *J Virol* 2012;86(21):11493-
 329 500.

330 [17] Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C, et al. Association between
 331 presence of *HLA-B*5701*, *HLA-DR7*, and *HLA-DQ3* and hypersensitivity to HIV-1
 332 reverse-transcriptase inhibitor abacavir. *Lancet* 2002;359(9308):727-32.

333 [18] Karl JA, Wiseman RW, Campbell KJ, Blasky AJ, Hughes AL, Ferguson B, et al.
 334 Identification of MHC class I sequences in Chinese-origin rhesus macaques.
 335 *Immunogenetics* 2008;60(1):37-46.

336 [19] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing
 337 from single polymerase molecules. *Science* 2009;323:133–38.

338 [20] Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template
 339 format for circular consensus sequencing and SNP detection. *Nuc Acids Res*
 340 2010;38(15):e159.

341 [21] Smith DG, McDonough JW, George DA. Mitochondrial DNA variation within and
 342 among regional populations of longtail macaques (*Macaca fascicularis*) in relation to
 343 other species of the fascicularis group of macaques. *American Journal of Primatology*
 344 2007;69:182-98.

345 [22] Forsheew T, Murtaza M, Parkinson C, Gale D, Tsui DWY, Kaper F, et al.
 346 Noninvasive identification and monitoring of cancer mutations by targeted deep
 347 sequencing of plasma DNA. *Sci Transl Med* 2012;4:136ra68.

348 [23] Dodt, M, Roehr JT, Ahmed R, Dieterich C. FLEXBAR - flexible barcode and adaptor
 349 processing for next-generation sequencing platforms. *Biology* 2012;1:895-905.

350 [24] Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve
 351 genome assemblies. *Bioinformatics* 2011;27:2957-63.

352 [25] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient
 353 alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10:
 354 R25.1-10.

355 [26] Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD
 356 and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 2014;gku1161.

357 [27] Parham P, Norman PJ, Abi-Rached L, Guethlein LA. Human-specific evolution of
 358 killer cell immunoglobulin-like receptor recognition of major histocompatibility complex
 359 class I molecules. *Phil. Trans. R. Soc. B* 2012;367:800-11.

360 [28] Creager HM, Becker EA, Sandman KK, Karl JA, Lank SM, Bimber BN, Wiseman
 361 RW, Hughes AL, O'Connor SL, O'Connor DH. Characterization of full-length MHC class
 362 II sequences in Indonesian and Vietnamese cynomolgus macaques. *Immunogenetics*
 363 2011; 63(9): 611–618.

364 [29] Bimber BN, Moreland AJ, Wiseman RW, Hughes AL, O'Connor DH. Complete
 365 Characterization of Killer Ig-Like Receptor (KIR) Haplotypes in Mauritian Cynomolgus
 366 Macaques: Novel Insights into Nonhuman Primate KIR Gene Content and Organization.
 367 *J Immunol.* 2008; 181(9): 6301–6308.

368

369 Figure Captions

370 Figure 1. Relative expression levels determined with MiSeq versus PacBio cDNA
371 amplicons in MCM. Results are normalized as the percentage of total MHC class I
372 reads identified for each transcript in each individual. Dashed lines indicate a transcript
373 was not detected by PacBio sequencing.

374

375 Figure 2. Relative expression levels of ChCM transcripts determined by
376 pyrosequencing. Results are normalized as the percentage of total MHC class I reads
377 identified for each transcript in each individual. Alleles detected in four or more reads by
378 PacBio sequencing are highlighted with grey boxes. ^a The *Mafa-B*038* allele was
379 expected for this sample, but missed by pyrosequencing due to a known mismatch with
380 the genotyping primers.

381

FIGURE 1

Animal ID	cy0321		cy0326		cy0165		cy0350		cy0348	
MCM MHC Haplotype 1	M1		M1		M3		M2		M4	
MCM MHC Haplotype 2	M1		M3		M3		M2		M4	
Sequencing Platform	MiSeq	PacBio	MiSeq	PacBio	MiSeq	PacBio	MiSeq	PacBio	MiSeq	PacBio
Total Reads Identified	22,006	2,611	25,676	2,259	20,819	2,008	22,909	3,795	8,382	1,175
Mafa-A Major										
A1*063:01	30.4	3.9	15.6	6.9			37.1	5.8		
A1*063:02			15.6	6.9	32.2	11.7				
A1*031:01									33.1	5.7
Mafa-A Minor										
A2*05:01	0.5	-	0.3	-			0.6	0.2		
A2*05:04									1.4	0.6
A2*05:11			0.3	-	0.7	0.1				
A4*01:01	7.2	0.3	7.2	0.2	7.2	-	7.1	1.0		
A5*30:01									1.4	0.9
Mafa-B Major										
B*064:01	11.5	3.7	4.3	1.6						
B*104:01:01	18.8	85.0	12.1	43.6						
B*134:02	13.6	5.2	6.7	4.2						
B*144:02	11.3	1.0	6.1	0.8						
B*011:01			3.5	10.2	7.3	30.5				
B*075:01			19.3	24.3	41.3	55.8				
B*019:03							13.9	9.0		
B*148:01:01							17.8	21.6		
B*150:01:01							18.5	62.2		
B*027:02									11.7	4.9
B*147:01									43.5	71.3
I*03:01:01									6.4	13.2
Mafa-B Minor										
B*046:01:01	1.6	0.7	0.9	0.2						

[illegible]

FIGURE 2

Animal ID	CW006	CW008	CW014	CW016	CW021	CW025	CW044	CW056	CW066	CW085
<i>Mafa-A</i> Haplotype 1	A008	A043	A043	A022a	A043	A027	A026	A022a	A087	A007
<i>Mafa-A</i> Haplotype 2	A043	A067	A130	A070	A079	A056a	A065a	A065a	A092	A057
<i>Mafa-B</i> Haplotype 1	B001a	B039e	B028d	B056b	B017	B017	B028d	B013d	B073	B028a
<i>Mafa-B</i> Haplotype 2	B013a	B023	B013f	B154	B056a	B083b	B184	B075b	B036	B056c
Total Reads Identified	290	301	406	372	359	337	294	259	365	379

***Mafa-A* Major**

***Mafa-A* Minor**

A4*14:13

A4*14

4.0

6.2

1.9

1.8

0.8

A5*30:06

0.8

Mafa-B Major

B*001:01

12.8

B*007:01

13.6

5.8

5.6

B*007:06

6.6

B*007

8.9

B*013:09

10.8

B*013:10

11.6

B*013:12

9.7

B*017:02

16.2

11.6

B*018:01:01

9.5

7.1

B*021:02

4.4

8.2

B*021

9.0

B*023:03

15.0

B*028:02

20.0

28.9

B*028:04

18.7

B*030:01:02

3.4

B*030:03:01

1.0

B*030:03:02

3.2

2.4

B*030:05:01

5.3

10.7

B*030:15

14.5

B*031:01

4.4

B*034:04

4.7

B*036:04

16.7

B*038:01

a

B*039:01

18.6

B*041:02

3.0

B*041:03

7.3

5.0

9.0

B*044:01:01

5.2

B*044

9.4

B*056:01

17.3

B*056:02:02

17.9

B*056:04

8.1

<i>B*066:01</i>			8.3						
<i>B*068:02</i>									2.9
<i>B*068:04</i>			3.4				4.8		
<i>B*068:06</i>								5.8	
<i>B*068:07</i>				9.7					
<i>B*073:02</i>									5.2
<i>B*075:03</i>								13.5	
<i>B*083:02</i>						8.3			
<i>B*085:01</i>	7.2								
<i>B*124:01:01</i>									5.0
<i>B*124:01:03</i>			2.5				4.1		
<i>B*137:03</i>								10.8	
<i>B*144:05</i>								3.1	
<i>B*154:01</i>				14.5					
<i>B*155:01</i>				1.3					
<i>B*184:02</i>							9.2		
Mafa-B Minor									
<i>B*014:01</i>			1.7						
<i>B*050:06</i>								8.5	
<i>B*060</i>		7.3	1.7	3.2	5.8	1.8			
<i>B*061:03</i>					7.2	12.5			
<i>B*061:04</i>			8.6				6.5		
<i>B*070:04</i>	12.8								
<i>B*079:01</i>									1.1
<i>B*079:nov:03</i>							1.4		
<i>B*079</i>									5.5
<i>B*082:nov:04</i>									2.1
<i>B*082</i>	1.0								
<i>B*089</i>							2.7		
<i>B*115</i>	1.0		1.0				1.4		
<i>I*01</i>	1.0			1.3					1.4
<i>I*01:w:02</i>				1.1					