Infection, Genetics and Evolution 88 (2021) 104700

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Using PacBio SMRT data for identification of class I MHC alleles in a wildlife species, *Zalophus californianus* (California sea lion)

Ellen E.M. Duckworth [a,b], Kaitlyn R. Romoser [a], Jeannine A. Ott [a], Thaddeus C. Deiss [a], Frances M.D. Gulland [c], Michael F. Criscitiello [a,d,*]

[a] Comparative Immunogenetics Laboratory, Department of Veterinary Pathobiology, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA
[b] Department of Large Animal Clinical Sciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA
[c] The Marine Mammal Center, Fort Cronkhite, Sausalito, CA 94965, USA
[d] Department of Molecular Pathogenesis and Immunology, College of Medicine, Texas A&M Health Science Center, Bryan, TX 77807, USA

ARTICLE INFO

Keywords:
MHC
Next generation sequencing
MHC class I
Wildlife
Marine mammals

ABSTRACT

High allelic polymorphism and association with disease susceptibility has made the genes encoding major histocompatibility complex (MHC) antigen presentation molecules in humans, domesticated animals, and wildlife species of wide interest to ecologists, evolutionary biologists, and health specialists. The often multifaceted polygenism and extreme polymorphism of this immunogenetic system have made it especially difficult to characterize in non-model species. Here we compare and contrast the workflows of traditional Sanger sequencing of plasmid-cloned amplicons to Pacific Biosciences SMRT circular consensus sequencing (CCS) in their ability to capture alleles of MHC class I in a wildlife species where characterization of these genes was absent. We assessed two California sea lions (*Zalophus californianus*), a species suffering from a high prevalence of an aggressive cancer associated with a sexually transmitted gamma herpesvirus. In this pilot study, SMRT CCS proved superior in identifying more alleles from each animal than the more laborious plasmid cloning/Sanger workflow (12:7, 10:7), and no alleles were identified with the cloning/Sanger approach that were not identified by SMRT CCS. We discuss the advantages and disadvantages of each approach including cost, allele rarefaction, and sequence fidelity.

1. Introduction

The major histocompatibility complex, or MHC, includes a family of genes that encodes a set of transmembrane proteins that are critical to the adaptive immune system in vertebrates (Flajnik and Kasahara, 2010). MHC molecules bind peptides of both phagolysosomal and cytosolic proteins and present them to T-cells (Criscitiello et al., 2013). MHC proteins vary greatly because they must be able to present a multitude of antigenic peptides, making them the most polymorphic mammalian genes (Parham and Ohta, 1996). MHC class II proteins present peptides on antigen presenting cells, such as dendritic cells and macrophages, to $CD4^+$ helper T-cells. MHC class I proteins, however, are present on almost every cell in the body and present intracellular proteins to $CD8^+$ cytotoxic T-cells, which can lyse virally infected or transformed neoplastic cells (Neefjes et al., 2011).

Study of MHC genes is popular for wildlife conservation studies because it gives insight into both genetic diversity and disease resistance of animal populations (Fleming-Canepa et al., 2016; Lane et al., 2012; Savage and Zamudio, 2011; Sommer, 2005), including the California sea lion *Zalophus californianus* (Bowen et al., 2005). Diversity of MHC genes is maintained primarily by pathogen-mediated selection (heterozygote advantage, rare-allele advantage, and fluctuating selection) (Sommer, 2005; Spurgin and Richardson, 2010), and loss of MHC diversity has been associated with higher parasite load (Belasen et al., 2019), increased disease susceptibility (Morris et al., 2013), and greater disease severity (Savage et al., 2019).

Marine mammals, moreover, are sentinel species, or indicators of the health of aquatic ecosystems (Bossart, 2011; Reddy et al., 2001). California sea lions have an unusually high prevalence of an aggressive cancer that affects up to 26% of the adult animals that die after stranding

* Corresponding author at: Texas A&M University, Mailstop 4467, College Station, TX 77843.
E-mail addresses: eeduckworth@gmail.com (E.E.M. Duckworth), Kaitlyn.romoser@tamu.edu (K.R. Romoser), jott@cvm.tamu.edu (J.A. Ott), tdeiss@cvm.tamu.edu (T.C. Deiss), fmdgulland@ucdavis.edu (F.M.D. Gulland), mcriscitiello@cvm.tamu.edu (M.F. Criscitiello).

https://doi.org/10.1016/j.meegid.2020.104700
Received 18 July 2020; Received in revised form 24 December 2020; Accepted 28 December 2020
Available online 31 December 2020

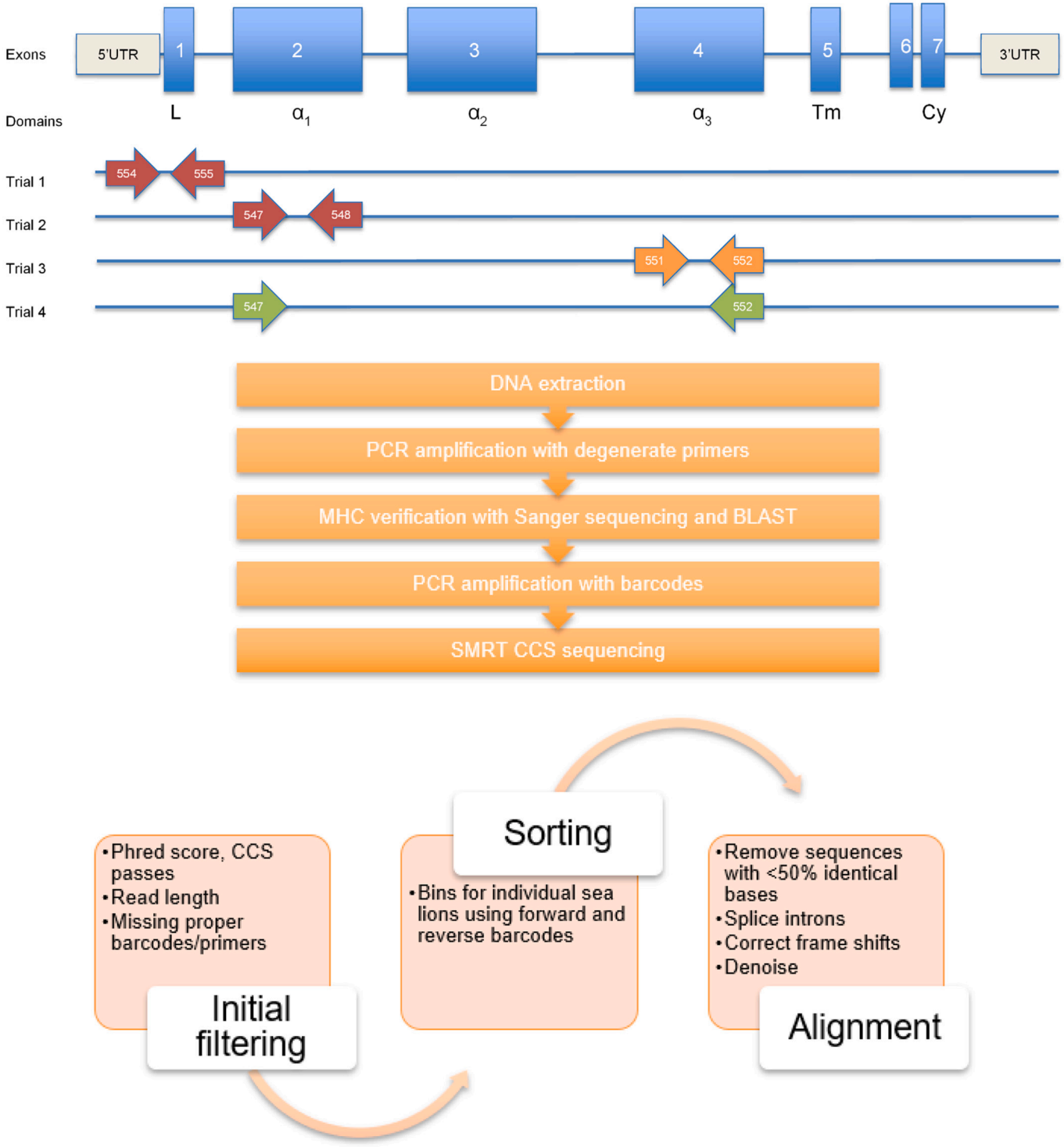

Fig. 1. A. Location of PCR primers used for amplification of MHC class I. Three primer pair combinations were attempted before successful generation of 1.6 kb band. Primers are represented by arrow symbols. Trial 1. Did not amplify MHC. Trial 2. Inconsistent amplification of α1 domain. Trial 3. Consistent amplification of α3 domain, but no peptide binding regions. Trial 4. Consistent amplification of 1.6 kb band containing α 1, 2, and 3 domains. B. Sequence generation and filtering. Pathway illustrating steps taken to generate sequences from skin samples and to ensure quality of reads obtained by CCS sequencing.

on the central California coast (Gulland et al., 1996), (Deming et al., 2018). Since a gamma herpesvirus has been implicated in the multi-factorial etiology (King et al., 2002; Lipscomb et al., 2000), we chose to study MHC class I in this species, the locus most commonly associated with immunogenetic susceptibility to neoplastic transformation and viral infection (Bashirova et al., 2011; Cheent and Khakoo, 2009; Orange and Ballas, 2006), which has not been previously characterized in this species. In humans, MHC class I polymorphisms affect susceptibility or resistance to viral infections and their associated neoplasias including HIV (Goulder and Walker, 2012), human papillomavirus (Davidson et al., 2003), and Epstein-Barr virus (Huang et al., 2012); and specific MHC class I alleles are associated with a higher risk of development of herpesvirus-associated Kaposi's sarcoma (Cornejo Castro et al., 2019).

Though the function of MHC genes and their linkage to genes

involved in antigen presentation pathways appear to be conserved across jawed vertebrates (Criscitiello et al., 2012; Ohta et al., 2002), significant variations in the genetic architecture have been described (de Sa et al., 2019). Large differences in the number and function of MHC loci in non-model species have been detected (Bentkowski and Radwan, 2019; Kelley et al., 2005). This, in combination with the large number of alleles, illustrates the need for an assessment of methods employing next generation sequencing when characterizing MHC genes in non-model species.

Traditional techniques using plasmid cloning and Sanger-based sequencing are labor intensive and generate low numbers of reads of moderate length and high accuracy, and therefore have been a mainstay of antigen receptor repertoire and MHC studies (Mashoof et al., 2014; Roberts et al., 2013). Next generation sequencing techniques offer the ability to generate thousands of sequences concurrently. However, popular next generation sequencing technologies such as Illumina, 454, and Ion Torrent have significant limitations including short read lengths and amplification biases (Quail et al., 2012; Roberts et al., 2013). Single-molecule real-time (SMRT) sequencing technology with the Pacific Biosciences (PacBio) RS II platform including circular consensus sequencing produces considerably longer DNA sequences with an average length of 13.5 kb, making it possible to obtain full-length coding regions from MHC class I from gDNA (Albrecht et al., 2017; Turner et al., 2018; Westbrook et al., 2015), which eliminates the need for downstream assembly of similar fragments. Though individual reads contain a relatively large number of errors, consensus accuracy is high due to the errors' random nature and iteration of reads possible with circular consensus sequencing (CCS) (Ono et al., 2013; Ross et al., 2013). Though this technique has recently been applied to several model species including humans (Albrecht et al., 2017; Turner et al., 2018), non-human primates (Hans et al., 2017; Maibach et al., 2017; Westbrook et al., 2015), and horses (Viluma et al., 2017), there is little work using CCS to characterize MHC in non-model species.

The goal of our study is to investigate the utility of PacBio SMRT CCS sequencing as an alternative to traditional approaches for characterizing MHC class I allelic polymorphism in the California sea lion *Zalophus californianus*.

2. Methods

2.1. DNA extraction

We selected our sample population from California sea lions that stranded on the central California coast and were taken to The Marine Mammal Center in Sausalito, CA, USA. The animals either died or were euthanized, and subsequently, extensive necropsies were performed. We selected two sea lions (#1 and #2) for use in the study, and full-thickness skin samples fixed in 90% EtOH were analyzed. Work was conducted under a Marine Mammal Protection Act permit (18786) issued to Dr. Frances Gulland, and results and data are freely shared through the NCBI database (accession numbers MT631846-MT631864). We extracted genomic DNA from the skin samples using Qiagen's Gentra PUREGENE Tissue Kit (Hilden, Germany). DNA was quantified using the NanoDrop® ND-1000 UV–Vis Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and diluted to a final concentration of 100 ng/μL.

2.2. PCR amplification, MHC verification and Sanger workflow

Selected segments of gDNA were amplified by polymerase chain reaction (PCR). PCR amplification was first performed by targeting β-actin, a highly-conserved mammalian housekeeping gene, as a positive control of gDNA quality. Primer sequences used to amplify the 275 bp fragment were FP-5′-GAG AAG CTG TGC TAC GTC GC- 3′ and RP-5′-CCA GAC AGC ACT GTYG TTG GC -3′ following previously reported PCR procedures (Browning et al., 2014).

We selected degenerate mammalian primer sequences previously used successfully on monk seal samples for PCR amplification of sea lion MHC class I gene fragments (Aldridge et al., 2006). Primer sequences FP-5'-GGC TCC CAC TCC CTG AGG T-3′ and RP-5'-CAG CCC CTC ATG CTG CAC-3′ targeted a long fragment containing the alpha 1, 2, and 3 domains (exons 2, 3, and 4, Fig. 1A). PCR reactions contained 100 ng gDNA, 10 μM of each primer, 25 mM $MgCl_2$, 10 μM dNTPs, 10× PCR Buffer B1 (Mango Biotechnology, Mountain view, CA, USA), and 5 U HOT FIREPol DNA polymerase (Mango Biotechnology) in a total volume of 50 μL. Amplifications were performed in a Bio-Rad MJ Mini Personal Thermal Cycler (Hercules, CA, USA) under the following conditions: 94 °C initial denaturation for 15 min; 35 cycles of 95 °C denaturation for 30 s, 56 °C annealing for 30 s, 72 °C extension for 1 min; and 72 °C final extension for 5 min. The PCR fragments were separated into bands according to size by agarose (0.8%) gel electrophoresis. We added Biotium's GelGreen fluorescent stain prior to separation to allow visualization of the bands by transillumination (Dark Reader Transilluminator; Clare Chemical Research, Dolores, CO, USA). Bands approximately 1.8 kb in size were extracted from gels using PureLink Quick Gel Extraction Kit (Life Technologies, Carlsbad, CA, USA) and quantified using the NanoDrop® ND-1000 UV–Vis Spectrophotometer. Three distinct amplification replicates were used to mitigate early round amplification bias that can skew final cloning frequency of the PCR of alleles.

To confirm that the PCR products were segments from MHC class I gene sequences, we cloned the isolated fragments from two different sea lions into a T/A vector (pCRII-TOPO; Life Technologies, Carlsbad, CA, USA). These constructs were transformed into *E. coli* (One Shot TOP10 Chemically Competent *E. coli*; Life Technologies, Carlsbad, CA, USA). After mini-prep culture, the *E. coli* cells that contained the vector were lysed (after blue/white selection on agar plates with carbenicillin and X-gal), and plasmid DNA was isolated using ZR Plasmid Miniprep – Classic Kit (Zymo Research, Irvine, CA, USA). The isolated DNA samples were sent to the Texas A&M DNA Technologies Core Lab for sequence determination using dideoxynucleotide chain termination (Sanger) technology in an automated sequencer (Model 3100; Applied Biosystems, Waltham, MA, USA). The sequences were then compared to known MHC class I sequences in other species using NCBI BLAST (Basic Local Alignment Search Tool). 24 plasmid clones were sequenced in the forward and reverse directions per individual, resulting in a total of 48 cloned MHC class I sequences.

2.3. PCR amplification and parallel workflow for SMRT CCS sequencing

After confirmation that the primers amplified the targeted MHC class I gene segments using Sanger sequencing, gDNA from the sea lions was used to undergo a second PCR reaction in preparation for using Single Molecule, Real-Time (SMRT) DNA sequencing (Pacific Biosciences, Menlo Park, CA, USA).

To differentiate samples during parallel sequencing, we tagged each sea lion sample with a specific barcode sequence (16 bp in length and custom-designed for the PacBio System). This PCR protocol differed slightly from above. PCR reactions contained 100 ng gDNA from gel slices of original amplification, 10 μM of each primer, 10 μM dNTPs, 5× Phusion HF Buffer (New England Biolabs, Ipswich, MA, USA), and 5 U Phusion High-Fidelity Taq polymerase (New England Biolabs) for a total volume of 50 μL. The Phusion polymerase was used for high fidelity and to create blunt-ended amplicons. Amplifications were performed in a Bio-Rad MJ Mini Personal Thermal Cycler under the following conditions: 98 °C initial denaturation for 30 s; 35 cycles of 98 °C denaturation for 10 s, 56 °C annealing for 20 s, 72 °C extension for 1 min; and 72 °C final extension for 5 min.

Approximately 120 ng of each barcoded sample was pooled (concentrations from ranged from 0.2–25 ng/μL), and approximately 1 mL of pooled sample was sent to the Duke Center for Genomic and Computational Biology Core for a total of 4 runs. Circular Consensus

Table 1
CCS filtering results. Initial quality filter was performed by Duke Center for Genomic and Computational Biology Core on a large pool of samples. Denoising was performed with Acacia software.

	Z. californianus 1	*Z. californianus* 2
% reads eliminated by Phred/CCS pass	85%	81%
# reads post quality filter	654	484
# reads eliminated by length	50	86
# reads eliminated by < 50% ID bases	1	4
# reads eliminated by denoising	8	9
# clusters	15	10
Alleles generated	12	10

Sequencing (CCS) was performed on a PacBio RSII instrument.

2.4. CCS quality control

Initial quality control and read filtering based on Phred quality scores and CCS passes was performed at Duke University Genome Center. CCSs were annotated within the Geneious Software Suite (Biomatters). Sequences that were substantially shorter or longer than the expected length (<1500 bp or >1700 bp) were excluded from further analysis. All remaining reads were sorted by forward and reverse barcodes representing individual sea lions, and the following steps were performed for each sea lion separately. Reads that did not contain proper barcode combinations or primer sequences were removed. Sequences were aligned using MUSCLE alignment with eight iterations. Reads were filtered when less than 50% of bases were identical to any of the other sequences. Predicted introns were trimmed to leave protein-coding regions for allelic analysis. Visual inspection was performed to check for a shift in the protein-reading frame. Nucleotides with a mean pair-wise identity of less than 30% over all sequences in the alignment that caused a shift in the open reading frame were manually removed. The resulting FASTA file was denoised using the Acacia software (Bragg et al., 2012) (Fig. 1 B). All remaining reads were used in the allele identification workflow.

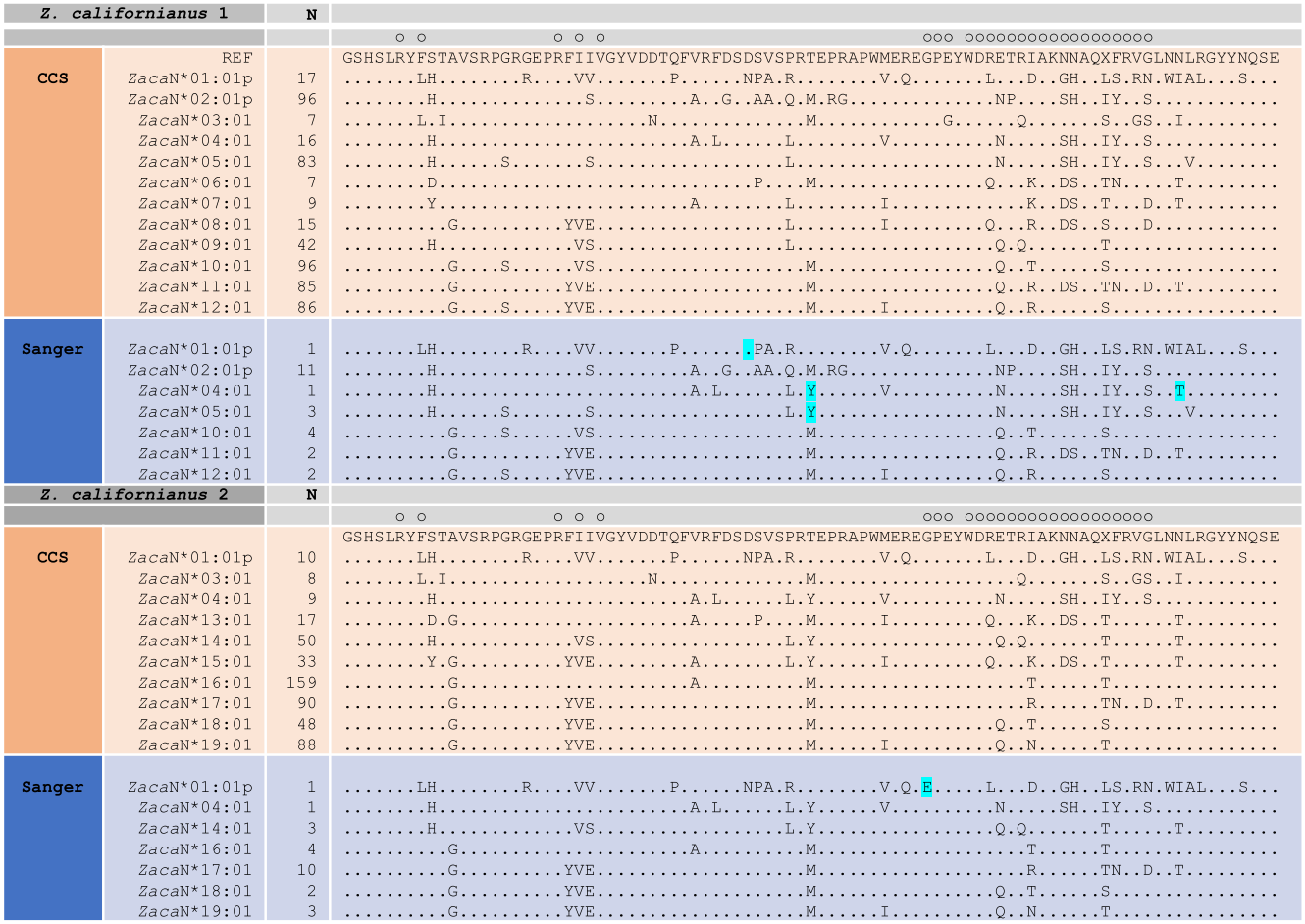

Fig. 2. A. Amino acid alignment of α1 domain. MHC class I alleles generated from two California sea lions is aligned to a representative allele. California sea lions have diverse MHC class I alleles. CCS sequencing detected more alleles that Sanger sequencing. N = number of reads representing each allele. Putative peptide binding sites marked by o (Pan et al., 2008, Bjorkman et al., 1987). Differences between CCS and Sanger sequences in the same allele are highlighted blue. B. Amino acid alignment of α2 domain. MHC class I alleles generated from two California sea lions is aligned to a representative allele. California sea lions have diverse MHC class I alleles. CCS sequencing detected more alleles that Sanger sequencing. N = number of reads representing each allele. Putative peptide binding sites marked by o (Pan et al., 2008){Bjorkman, 1987 #4885}. Differences between CCS and Sanger sequences in the same allele are highlighted blue. C. Amino acid alignment of α3 domain. MHC class I alleles generated from two California sea lions is aligned to a representative allele. California sea lions have diverse MHC class I alleles. CCS sequencing detected more alleles that Sanger sequencing. N = number of reads representing each allele. Putative peptide binding sites marked by o (Pan et al., 2008) {Bjorkman, 1987 #4885}. Differences between CCS and Sanger sequences in the same allele are highlighted blue. D. Venn diagram illustrating alleles captured by CCS sequencing and Sanger sequencing in two California sea lions. SMRT CCS discovered 12 MHC class I alleles in Z. californianus #1 1 and 10 in Z. californianus #2 2; the cloning/Sanger workflow identified seven alleles in each animal, all of which were also discovered by SMRT CCS. Two alleles (*Zaca*N*01:01p and *Zaca*N*04:01) were shared between the two animals.

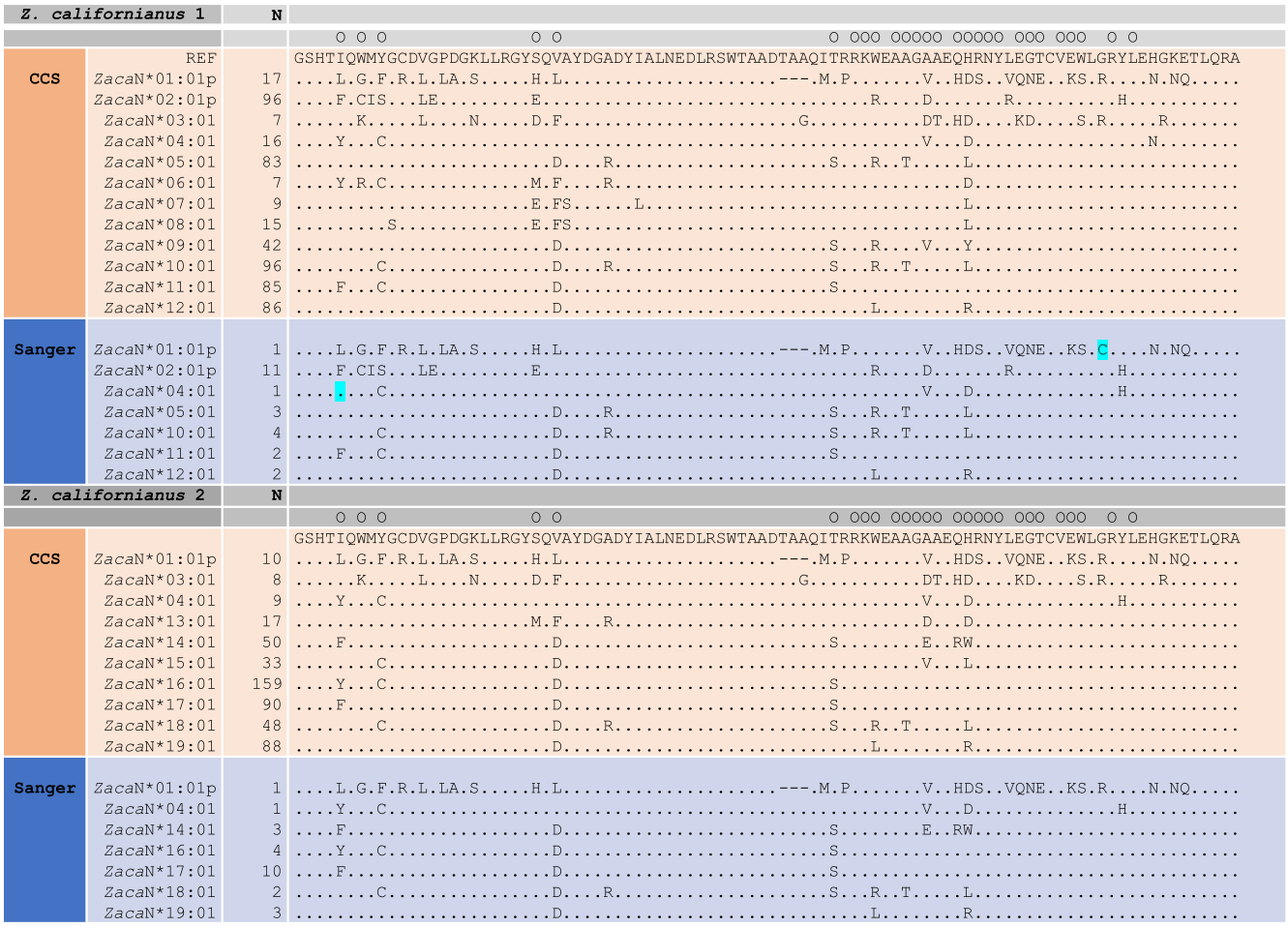

Fig. 2. (*continued*).

2.5. Allele identification and naming

Reads were clustered using the HKY UPGMA tree-builder function in Geneious. All reads not assigned to clusters were considered artifacts and excluded from further analysis. Clusters with >98.5% homology were considered distinct alleles (Hammond et al., 2012), though this conservative criteria due to a small sample set could prevent discovery of variants that differ by substitutions at one or two positions. Alleles were named following the nomenclature (Klein et al., 1990) for naming MHC alleles (*Mhc*Zaca-UA*xx), however, the prefix N was used to show that they do not yet have a designated locus (Hammond et al., 2012). Coding sequences with frameshift mutations or stop codons were annotated as pseudogenes; all other coding sequences were presumed to be functional.

2.6. Identification of MHC class I genes in the genome

The MHC class I gene and pseudogene loci were identified by aligning our nucleotide sequences against the sea lion genome assembly (https://www.ncbi.nlm.nih.gov/bioproject/561800) using NCBI BLAST. The base-pair locations of the matching sequences were then used to build a map of MHC class I genes and pseudogenes (Fig. 5 A). The sequences of the loci identified in the genome were then extracted and clustered with the alleles identified from *Zalophus californianus* #1 and #2 using HKY UPGMA tree-builder function (Fig. 5 B).

3. Results

3.1. SMRT CCS yields more alleles than traditional cloning

After multiple unsuccessful degenerate primer strategies (Fig. 1 A), an amplicon of 1.6 Kbp containing exons 2–4 of MHC class I encoding the α1 and α2 peptide binding domains as well as the immunoglobulin superfamily α3 domain was subjected to the workflow depicted in Fig. 1 B. Sanger sequencing resulted in 24 clones analyzed from each of two animals, whereas SMRT CCS resulted in 685 (*Z. californianus* #1) and 484 (*Z. californianus* #2) reads after quality control (Table 1).

The PacBio run for *Z. californianus* #1 initially contained 300,584 reads. 254,388 reads were eliminated based on Phred quality scores and CCS passes, leaving 46,196 reads. The polymerase read quality score increased from 0.137 to 0.849 after the filtering process. The batch for *Z. californianus* #2 initially contained 601,168 reads. 485,024 reads were eliminated, increasing the polymerase read quality score from 0.175 to 0.849. 654 and 484 reads were obtained for *Z. californianus* #1 and #2, respectively, after the read filtering process. A total of 595 (90.9%) reads of *Z. californianus* #1 and 385 (79.5%) reads of *Z. californianus* #2 passed the subsequent filtering process by showing the expected read length, base identity, and read quality through the Acacia denoising process. The Acacia program eliminated reads by filtering out any sequence greater than two standard deviations from the mean length or having an average read quality of less than 30.

Pairwise nucleotide identity of >98.5% was set as a threshold for calling distinct alleles. SMRT CCS discovered 12 MHC class I alleles in *Z. californianus* #1 and 10 in *Z. californianus* #2; the cloning/Sanger

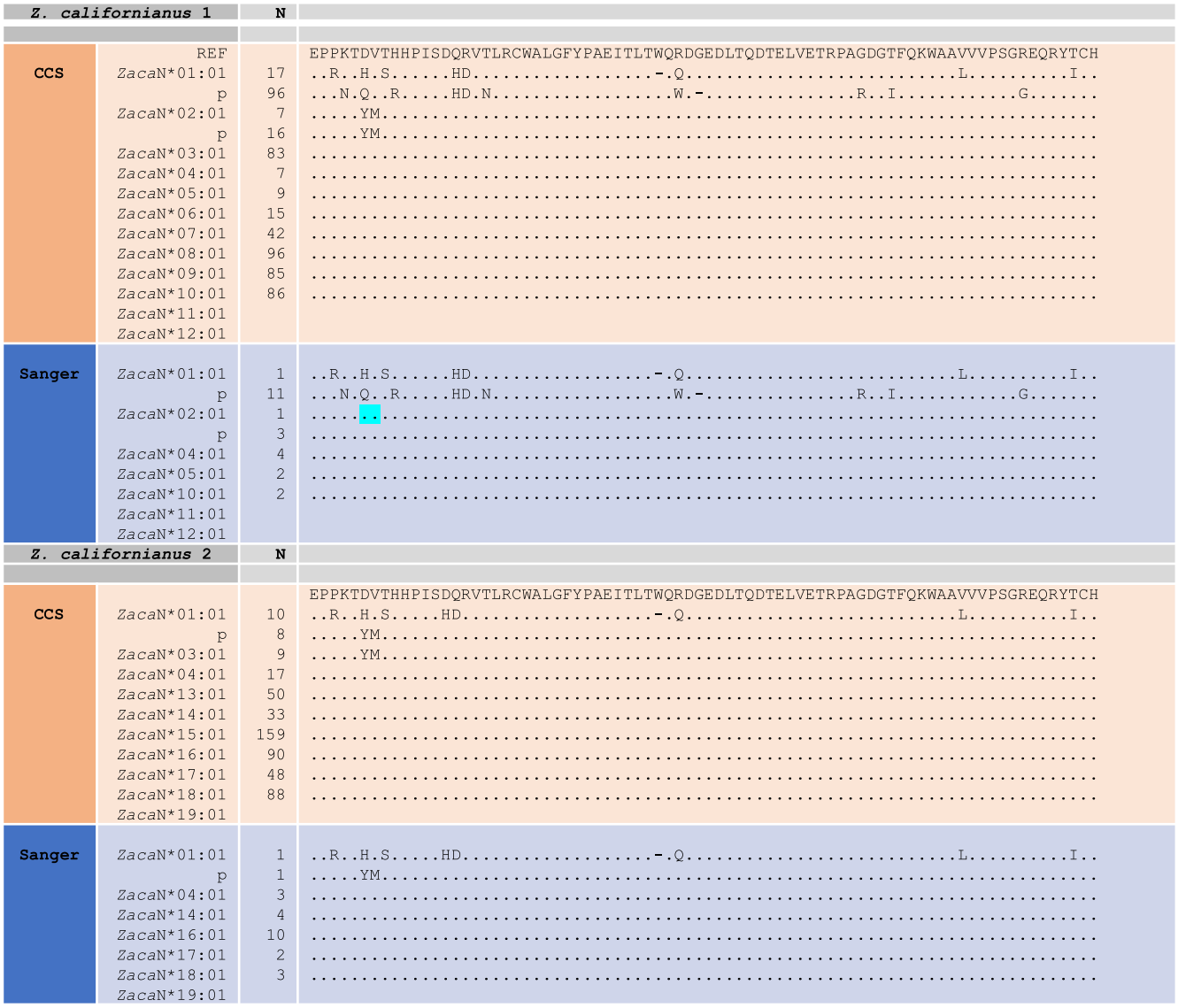

Fig. 2. (*continued*).

workflow identified seven alleles in each animal, all of which were also discovered by SMRT CCS (Fig. 2). Two alleles (*Zaca*N*01:01p and *Zaca*N*04:01) were shared between the two animals in this pilot study. Based on these results, it would take much deeper cloning to approach the rarefaction of new alleles identified by SMRT CCS.

3.2. *Zalophus californianus MHC class I sequences suggest polygenism and polymorphism*

The two allele characterization approaches identified a total of 19 unique alleles in the two sea lions. Individual alleles were confirmed by up to 11 clones in the Sanger workflow, but several singletons were not corroborated by this technique (Fig. 2). All alleles discovered in each animal were sequenced at least seven times in the SMRT CCS workflow, with many sequenced much more (e.g., *Zaca*N*16:01 sequenced 159 times from *Z. californianus* #2).

The large number of alleles discovered from a single animal (12 by CCS from *Z. californianus* #1) betray a complex locus with at least six MHC class I genes, assuming heterozygosity at each and no alleles shared between genes. Two differences in amino acid residues between SMRT CCS and Sanger sequences in the same allele were detected (highlighted in Fig. 2A and B). One difference was detected in the alpha 1 domain of *Zaca*N*01:01p in *Z. californianus* #1, and the other difference was detected in the alpha 2 domain of *Zaca*N*01:01p in *Z. californianus* #2. No differences were detected in the alpha 3 domain. Both of the differences occurred in alleles that were only verified by one Sanger clone.

*Zaca*N*01:01p contains an insertion near the beginning of the α2 domain and several deletions in the α2 and α3 domains which cause a frameshift in the amino acid translation for the majority of the coding regions for exon 3 and 4 (illustrated by Fig. 5 A and Fig. 5 B). These insertions and deletions were present in all sequences and conserved between the two sea lions. Similarly, *Zaca*N*02:01p contains two deletions in the α3 domain, creating a frameshift that is 30 amino acids in length (Fig. 4).

3.3. *Polygenism confirmed in Zalophus californianus genome*

Seven class I loci were identified on chromosome 6 of the recently published sea lion genome assembly mZalCal1 (Fig. 5 A). Two of those loci were closely related to *Zaca*N*01:01p and *Zaca*N*02:01p (Fig. 5 B). The twelve alleles discovered in *Zalophus californianus #1* are consistent

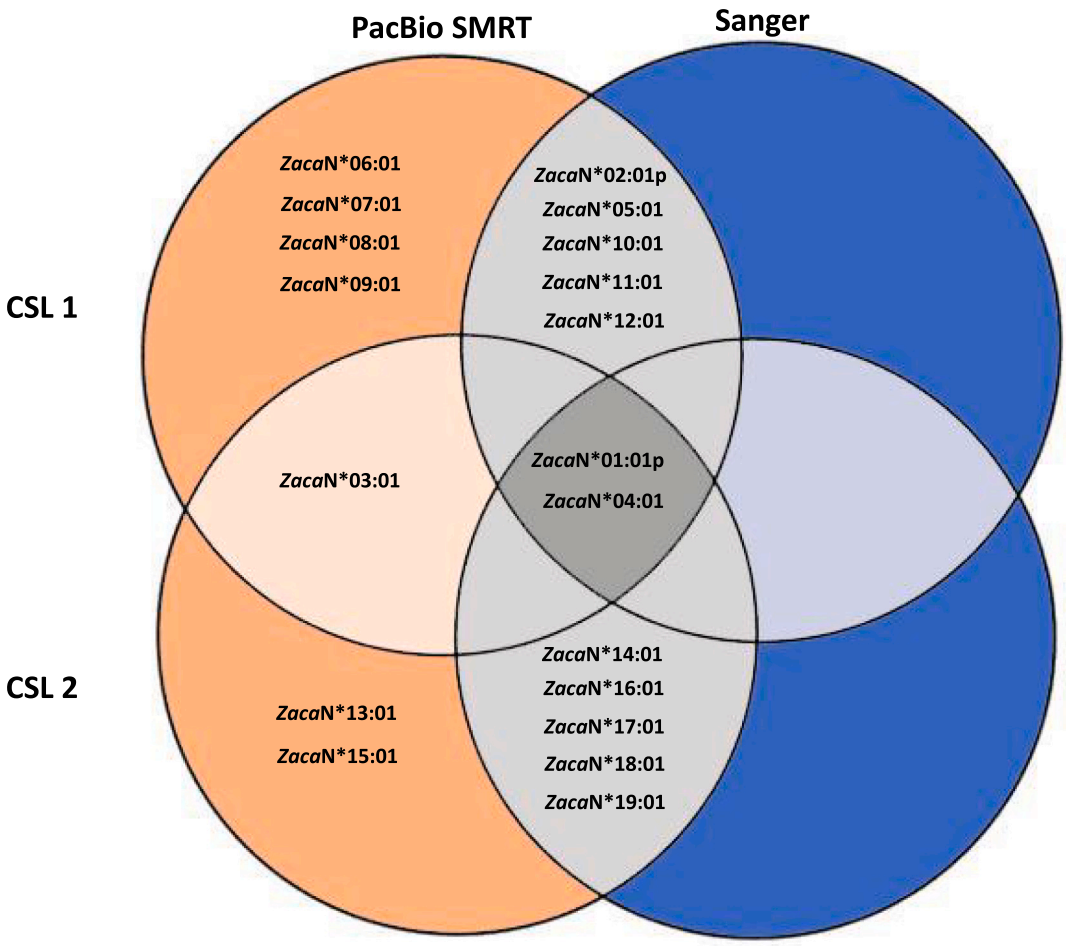

Fig. 2. (*continued*).

with five polymorphic class I loci (yielding ten alleles if heterozygous at each) and two pseudogene loci (*Zaca*N*01:01p and *Zaca*N*02:01p). *Zaca*N*01:01p but not *Zaca*N*02:01p was discovered in *Zalophus californianus* #2. The ten alleles discovered in *Zalophus californianus* #2 correspond with one pseudogene loci (*Zaca*N*01:01p) and five polymorphic class I loci, assuming that one locus was homozygous or an allele was not discovered. It is possible, however, that the sea lions were homozygous at more than one loci. Since there is no information currently available on MHC class I copy number variance in this species, Southern blotting and BAC clone analysis is necessary to definitively resolve the issue.

3.4. Sequence diversity concentrated in predicted peptide binding groove

The aligned amino acid translations of the α1 and α2 domains show the concentration of diversity in those residues predicted to interact with peptide antigen (Fig. 2 A and B). The α3 domain showed much less diversity amongst alleles, as expected for a portion of the molecule interacting with CD8 and not peptide or T-cell receptor (Fig. 2 C). Molecular modeling of one allele (*Zaca*N*04:01) that was shared between the two subjects conforms to the classical MHC tertiary and quaternary structure of alpha helices flanking a peptide binding groove floored by anti-parallel β-strands forming a sheet (Fig. 3). Putative peptide binding sites are used (Pan et al., 2008){Bjorkman, 1987 #4885}; crystallography or MHC immunoprecipitation and mass spectrometry are needed to definitively identify peptide binding sites.

The more complete SMRT CCS dataset provided the more robust data for diversity estimates (Table 2). *Z. californianus* #1 (from which the most alleles and sequences were sampled) yielded 37.6 polymorphic sites per sequence and a nucleotide diversity per site (π) of 0.04.

4. Discussion

4.1. MHC Class I in Z. californianus

A total of 19 unique alleles were discovered between the two sequencing techniques, including 12 from a single animal. Because of the small sample size, however, allele discovery is limited to preliminary findings. A future study using CCS sequencing from multiple PCR reactions that validates alleles by their discovery in at least three animals is indicated. Our results suggested polygenism, which was confirmed by the California sea lion reference genome that was recently made available, and polymorphism, though we did not amplify the entire MHC class I gene, which may have left some variation undetected. The number of MHC class II DRB genes has been shown to be predictive for both antigen recognition and disease severity with *Leptospira* infections in this species (Acevedo-Whitehouse et al., 2018).

The frameshifts in amino acid translation caused by insertions and deletions in two of the discovered alleles suggest pseudogenes or non-classical MHC function. Because the frameshifts were present in all sequences and, in the case of *Zaca*N*01:01p, discovered in two sea lions, the probability that the insertion and deletions were caused by PCR or sequencing error is low. ZacaN*02:01p was discovered in *Zalophus californianus* #1 but not #2. A broader study using more sea lions is indicated to better understand the prevalence of pseudogenes in the population.

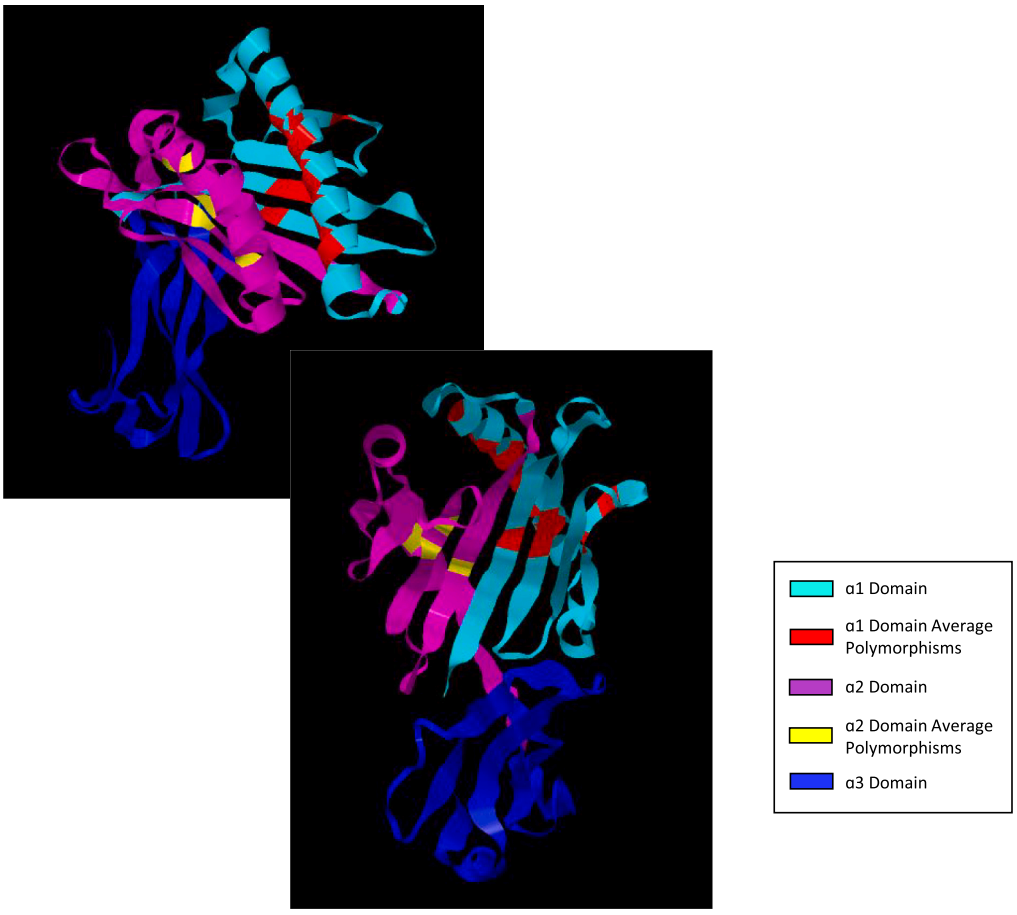

Fig. 3. 3-D ribbon structure of MHC class I molecule illustrating polymorphic sites. SWISS-MODEL Template 5vge.1 from GenBank record FLJ54183 was selected using the *Zaca*N*04:01 sequence. Once the model was created, the .PDB file was downloaded into Geneious software and depicted as a 3-D ribbon structure. Highlighted areas in the structure represent amino acids that displayed polymorphisms in half or more of the alleles. Putative MHC class I polymorphic sites are used (Pan et al., 2008){Bjorkman, 1987 #4885}.

4.2. Comparison of allele identification workflows

Cloning and Sanger sequencing is a well-established method for characterizing MHC in diverse non-model species (Fleming-Canepa et al., 2016; Lane et al., 2012; Ohta et al., 2002; Savage and Zamudio, 2011). However, this method is labor-intensive and expensive. Targeted CCS sequencing in non-model species is still relatively uncommon (Larsen et al., 2014), but may provide several benefits when characterizing clinically relevant genes in species where there is no reference genome. CCS sequencing quickly generates hundreds of more reads than traditional Sanger sequencing, potentially making allele discovery more complete as deeper read coverage should increase the rarefaction of poorer amplified alleles. In our study, CCS sequencing uncovered seven more alleles in two sea lions than traditional Sanger sequencing methods. Also, CCS sequencing generates long read lengths, which offers advantages when characterizing complex multi-gene families (Larsen et al., 2014).

Other recent studies have produced similar results. One study using CCS sequencing to target highly complex vomeronasal gene receptors in non-model grey mouse lemurs demonstrated that CCS sequencing identified 100% of V1R*J* sequences in the draft genome and 13 novel loci (Larsen et al., 2014). Recently, the genomic structure of MHC class II region in the horse has been resolved using PacBio SMRT sequencing (Viluma et al., 2017). In macaques, SMRT-CCS sequencing's ability to identify MHC class I sequences that had previously been characterized by Sanger sequencing was validated and then used to discover 60 novel full-length sequences (Westbrook et al., 2015). In human medicine, SMRT sequencing is now the gold standard for clinical sequencing of highly polymorphic HLA genes for stem cell transplants (Turner et al., 2018). Long-range next-generation sequencing techniques are becoming increasingly more affordable, and pooling of barcoded samples in runs makes this approach more cost-effective after accounting for the labor, time and expendables of traditional cloning.

Both Sanger and CCS sequencing contain PCR error and error within the sequencing process itself. However, the random nature of CCS error provides an advantage over other next-generation sequencing techniques (Larsen et al., 2014), as consensus accuracy is 99.997% which matches or exceeds the ability of short-read sequencing (Wenger et al., 2019), and there is no risk of recombined alleles. In our study, Sanger sequencing had more polymorphic sites per sequence, but discovered fewer variable sites overall when compared to CCS sequencing in both sea lions (Table 2). This is potentially due to a higher error rate in Sanger sequencing, but lower number of total discovered alleles in Sanger sequencing than CCS sequencing. Because of practical limits on the number of colonies that can be sequenced in the Sanger workflow, it is often difficult to identify and correct PCR and sequencing errors. One difference in amino acid residues was detected between CCS and Sanger sequences in *Zaca*N*01:01p in each sea lion. Because these differences were detected in alleles that were only verified by one Sanger clone, the difference is likely attributed to Sanger sequencing error. Because the sequences detected by both Sanger sequencing and CCS sequencing were identified by two independent PCR reactions, it is very unlikely that the shared alleles are artifacts. It is possible, however, that the alleles identified solely by the CCS platform contain PCR error since the alleles have not been verified by more than one PCR reaction.

Though the results from the comparison of CCS sequencing to Sanger sequencing are striking, the small sample size limits our ability to quantify the difference in allele discovery between the two techniques that would allow a reasonable prediction to be made for a future study. Despite this caveat, our work here demonstrates that CCS sequencing

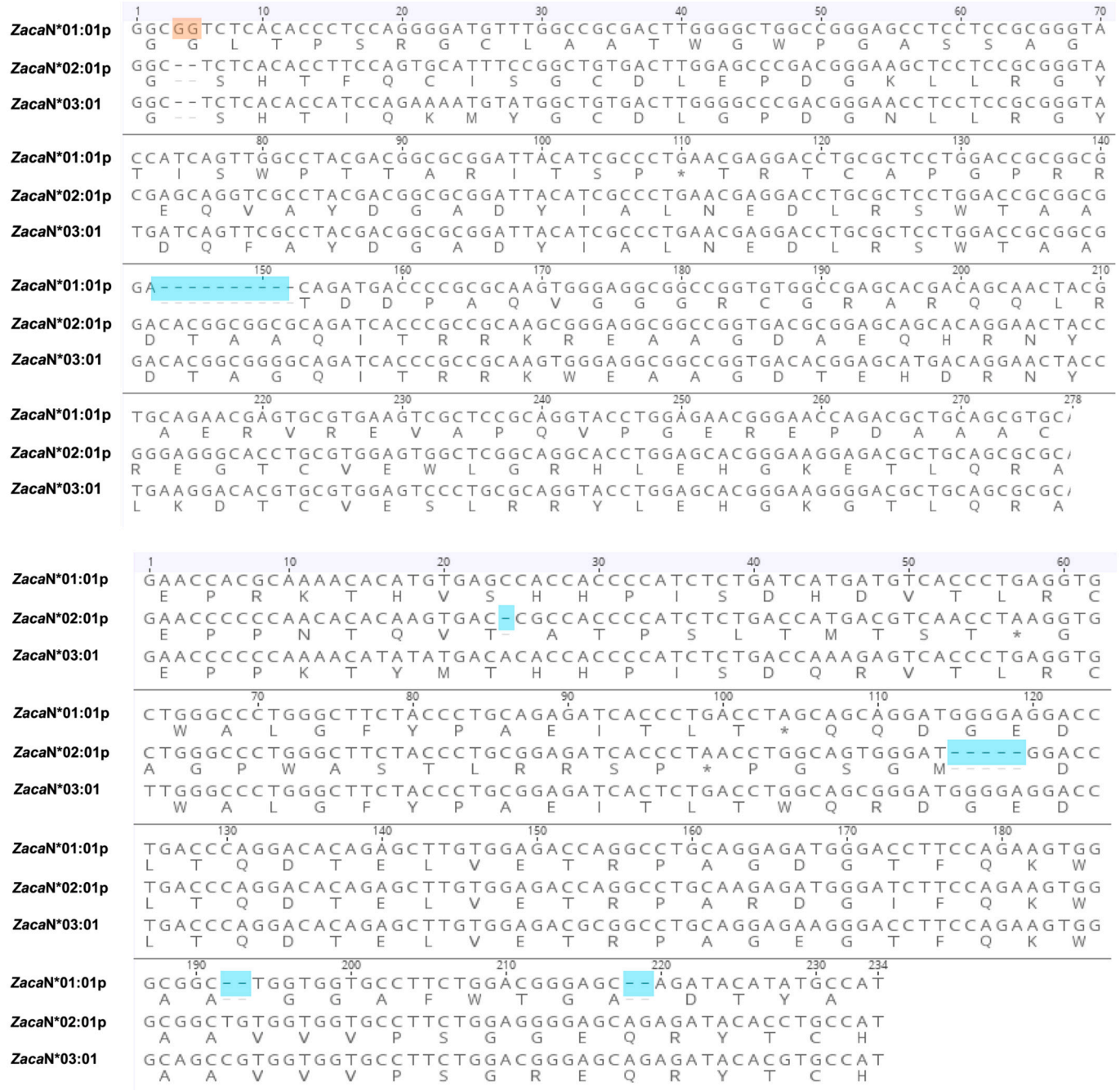

Fig. 4. A. Nucleotide alignment of α2 domain illustrating frameshift in amino acid translation caused by insertion and deletion in ZacaN*01:01p. Figure made using Geneious software. Insertion is highlighted orange. Deletions are highlighted blue. B. Nucleotide alignment of α3 domain illustrating frameshifts in amino acid translations caused by deletions in ZacaN*01:01p and ZacaN*02:01p. Figure made using Geneious software. Deletions are highlighted blue.

can be a useful tool for characterizing complex gene families in non-model organisms. The increasing recognition of viral and neoplastic diseases in wildlife populations places high priority on efficient and effective immunogenetic techniques in these animals. Based on the results of this pilot study, we will use the SMRT CCS techniques described in this paper to address the relationship between MHC class I alleles and cancer-status in sea lions, and also characterize MHC in other non-model wildlife species. The technique could also be applied to other immunogenetic systems where allelic polymorphism has been described (e.g. T-cell receptor constant domains of teleost fish (Criscitiello et al., 2004)). However, these advances in sequencing technologies do not replace the utility of Southern blotting and BAC clone analysis that may be necessary to resolve copy number variation.

5. Conclusions

In our study, CCS sequencing uncovered seven more alleles in two sea lions than traditional Sanger sequencing methods. The identified workflow for MHC characterization using CCS sequencing is a cost-effective, efficient technique that can be applied to other non-model wildlife species with conservation and disease implications. As the COVID-19 pandemic clearly demonstrates, better implementation of a One Health appreciation of the interconnectivity of the health of environment, wildlife and humans will benefit all three (Mwangi et al., 2016).

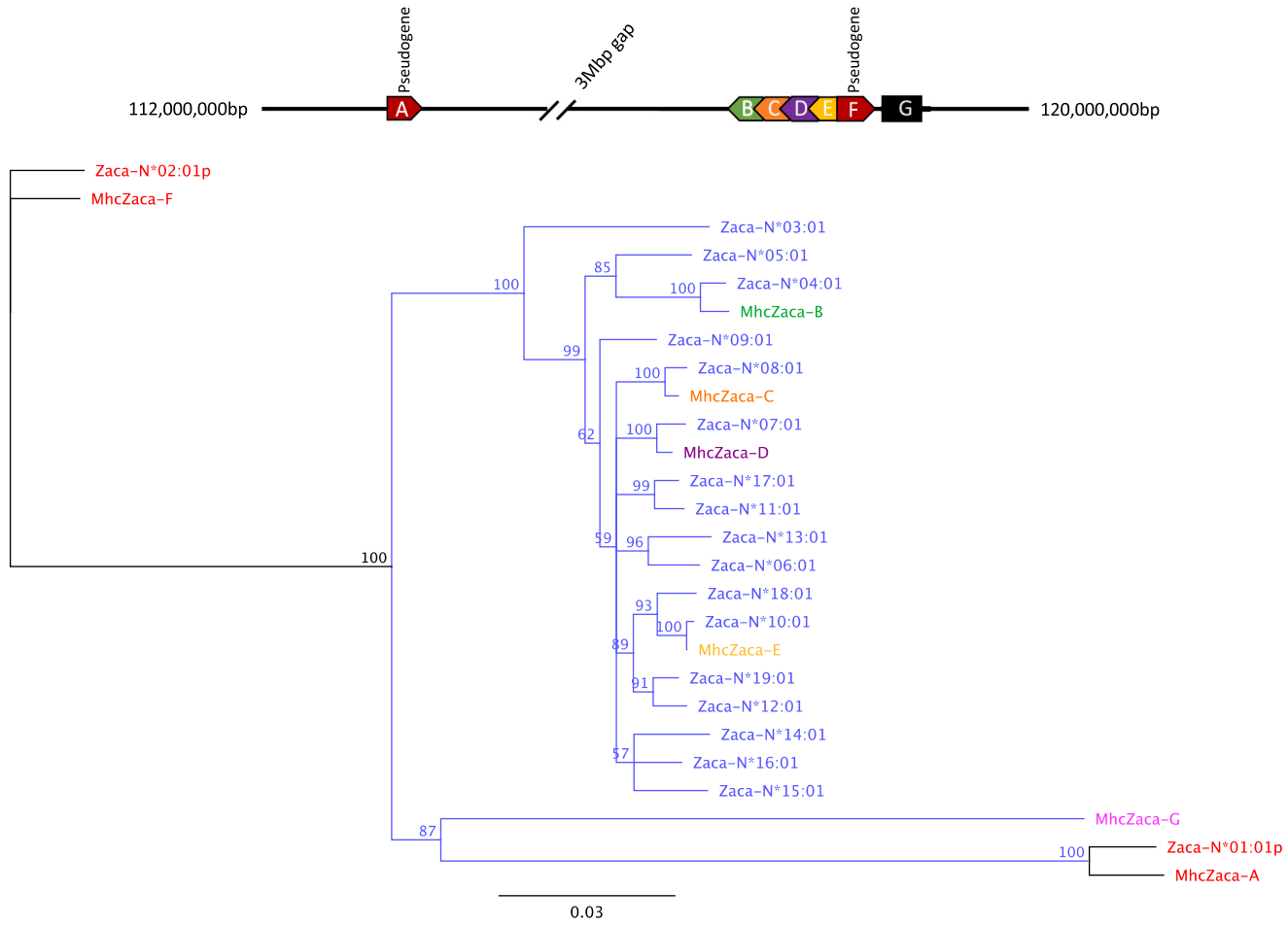

Fig. 5. A. Map of MHC class I gene and pseudogene loci. Sea lion MHC class I genes are located on a region of chromosome 6. By referencing the assembly from GenBank record mZalCal1.pri, we were able to align our alleles to loci along the chromosome. Each allele is organized into different loci, which can be seen in Fig. 5 B. The red blocks are pseudogenes, while the green, orange, purple, yellow, and pink blocks are different loci. The orientation of each block indicates transcriptional direction. B. Tree showing potential relationship of *Zalophus californianus* #1 and *Zalophus californianus* #2 alleles to MHC class I loci identified in the sea lion genome. Bootstrap support for nodes after 1000 iterations are shown. Colors of reference assembly genes correspond to schematic in 5 A.

Table 2
Summary of CCS and Sanger sequencing results and polymorphism data.

	#reads	#alleles	k	π	S	θ
Z. californianus 1						
Sanger	24	7	49.4	0.06	165	49.8
CCS	595	12	21.0	0.04	225	37.6
Z. californianus 2						
Sanger	24	7	28.1	0.04	164	48.7
CCS	385	10	18.2	0.03	217	40.0

Number of reads is the number left after filtering process was completed. k is average number of nucleotide differences between two sequences. π is nucleotide diversity per site between two sequences. S is total number of variable sites. θ is the number of polymorphic sites per sequence.

Credit author statement

EEMD and MFC wrote the paper.

EEMD, KRR, JOO, MFC and TCD performed experiments and analysis.

EEMD, FMDG and MFC planned the project.

Conflict of interest and authorship conformation form

All authors have participated in (a) conception and design, or analysis and interpretation of the data; a drafting the article or revising it critically for important intellectual content; and (c) approval of the final version. (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgements

The work was funded by a grant from the Department of Veterinary Large Animal Clinical Sciences at Texas A&M University to EED and NSF IOS 1656870 to MFC. The authors would like to acknowledge the Sea Lion Cancer Consortium (SLiCC) for access to samples and coordination of research on urogenital cancer in California Sea Lions. We thank Claire Christian for assistance with sequence submissions to Genbank.

References

Acevedo-Whitehouse, K., Gulland, F.M.D., Bowen, L., 2018. MHC class II DRB diversity predicts antigen recognition and is associated with disease severity in California sea lions naturally infected with Leptospira interrogans. Infect. Genet. Evol. 57, 158–165.

Albrecht, V., Zweiniger, C., Surendranath, V., Lang, K., Schofl, G., Dahl, A., Winkler, S., Lange, V., Bohme, I., Schmidt, A.H., 2017. Dual redundant sequencing strategy: full-length gene characterisation of 1056 novel and confirmatory HLA alleles. HLA 90, 79–87.

Aldridge, B.M., Bowen, L., Smith, B.R., Antonelis, G.A., Gulland, F., Stott, J.L., 2006. Paucity of class I MHC gene heterogeneity between individuals in the endangered Hawaiian monk seal population. Immunogenetics 58, 203–215.

Bashirova, A.A., Thomas, R., Carrington, M., 2011. HLA/KIR restraint of HIV: surviving the fittest. Annu. Rev. Immunol. 29, 295–317.

Belasen, A.M., Bletz, M.C., Leite, D.d.S., Toledo, L.F., James, T.Y., 2019. Long-term habitat fragmentation is associated with reduced MHC IIB diversity and increased infections in amphibian hosts. Front. Ecol. Evol. 6.

Bentkowski, P., Radwan, J., 2019. Evolution of major histocompatibility complex gene copy number. PLoS Comput. Biol. 15, e1007015.

Bjorkman J, P, Saper A, M, Samaraoui, B, Bennet S, W, Strominger L, J, Wiley C, D, 1987. The foreign antigen binding site and T cell recognition regions of 8 class I histocompatibility antigens. Nature 329 (6139), 512–518. https://doi.org/10.1038/329512a0.

Bossart, G.D., 2011. Marine mammals as sentinel species for oceans and human health. Vet. Pathol. 48, 676–690.

Bowen, L., Aldridge, B.M., Delong, R., Melin, S., Buckles, E.L., Gulland, F., Lowenstine, L. J., Stott, J.L., Johnson, M.L., 2005. An immunogenetic basis for the high prevalence of urogenital cancer in a free-ranging population of California sea lions (Zalophus californianus). Immunogenetics 56, 846–848.

Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P., Tyson, G.W., 2012. Fast, accurate error-correction of amplicon pyrosequences using Acacia. Nat. Methods 9, 425–426.

Browning, H.M., Acevedo-Whitehouse, K., Gulland, F.M., Hall, A.J., Finlayson, J., Dagleish, M.P., Billington, K.J., Colegrove, K., Hammond, J.A., 2014. Evidence for a genetic basis of urogenital carcinoma in the wild California sea lion. Proc. Biol. Sci. 281, 20140240.

Cheent, K., Khakoo, S.I., 2009. Natural killer cells: integrating diversity with function. Immunology 126, 449–457.

Cornejo Castro, E.M., Morrison, B.J., Marshall, V.A., Labo, N., Miley, W.J., Clements, N., Nelson, G., Ndom, P., Stolka, K., Hemingway-Foday, J.J., Abassora, M., Gao, X., Smith, J.S., Carrington, M., Whitby, D., 2019. Relationship between human leukocyte antigen alleles and risk of Kaposi's sarcoma in Cameroon. Genes Immun. 20, 684–689.

Criscitiello, M.F., Wermenstam, N.E., Pilstrom, L., McKinney, E.C., 2004. Allelic polymorphism of T-cell receptor constant domains is widespread in fishes. Immunogenetics. (55), 818–824.

Criscitiello, M.F., Ohta, Y., Graham, M.D., Eubanks, J.O., Chen, P.L., Flajnik, M.F., 2012. Shark class II invariant chain reveals ancient conserved relationships with cathepsins and MHC class II. Dev. Comp. Immunol. 36, 521–533.

Criscitiello, M.F., Dickman, M.B., Samuel, J.E., de Figueiredo, P., 2013. Tripping on acid: trans-kingdom perspectives on biological acids in immunity and pathogenesis. PLoS Pathog. 9, e1003402.

Davidson, E.J., Davidson, J.A., Sterling, J.C., Baldwin, P.J., Kitchener, H.C., Stern, P.L., 2003. Association between human leukocyte antigen polymorphism and human papillomavirus 16-positive vulval intraepithelial neoplasia in British women. Cancer Res. 63, 400–403.

de Sa, A.L.A., Breaux, B., Burlamaqui, T.C.T., Deiss, T.C., Sena, L., Criscitiello, M.F., Schneider, M.P.C., 2019. The marine mammal class II major histocompatibility complex organization. Front. Immunol. 10, 696.

Deming, A.C., Colegrove, K.M., Duignan, P.J., Hall, A.J., Wellehan, J.F.X., Gulland, F.M. D., 2018. Prevalence of urogenital carcinoma in stranded California Sea lions (Zalophus californianus) from 2005-15. J. Wildl. Dis. 54, 581–586.

Flajnik, M.F., Kasahara, M., 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat. Rev. Genet. 11, 47–59.

Fleming-Canepa, X., Jensen, S.M., Mesa, C.M., Diaz-Satizabal, L., Roth, A.J., Parks-Dely, J.A., Moon, D.A., Wong, J.P., Evseev, D., Gossen, D.A., Tetrault, D.G., Magor, K.E., 2016. Extensive allelic diversity of MHC class I in wild mallard ducks. J. Immunol. 197, 783–794.

Goulder, P.J., Walker, B.D., 2012. HIV and HLA class I: an evolving relationship. Immunity 37, 426–440.

Gulland, F.M.D., Trupkiewicz, J.G., Spraker, T.R., Lowenstine, L.J., 1996. Metastatic carcinoma of probable transitional cell origin in 66 free-living California sea lions (Zalophus californianus), 1979 to 1994. J. Wildl. Dis. 32, 250–258.

Hammond, J.A., Guethlein, L.A., Norman, P.J., Parham, P., 2012. Natural selection on marine carnivores elaborated a diverse family of classical MHC class I genes exhibiting haplotypic gene content variation and allelic polymorphism. Immunogenetics 64, 915–933.

Hans, J.B., Bergl, R.A., Vigilant, L., 2017. Gorilla MHC class I gene and sequence variation in a comparative context. Immunogenetics 69, 303–323.

Huang, X., Hepkema, B., Nolte, I., Kushekhar, K., Jongsma, T., Veenstra, R., Poppema, S., Gao, Z., Visser, L., Diepstra, A., van den Berg, A., 2012. HLA-A*02:07 is a protective allele for EBV negative and a susceptibility allele for EBV positive classical Hodgkin lymphoma in China. PLoS One 7, e31865.

Kelley, J., Walter, L., Trowsdale, J., 2005. Comparative genomics of natural killer cell receptor gene clusters. PLoS.Genet. 1, 129–139.

King, D.P., Hure, M.C., Goldstein, T., Aldridge, B.M., Gulland, F.M.D., 2002. Otarine herpesvirus-1: a novel gammaherpesvirus associated with urogenital carcinoma in California Sea lions (Zalophus californianus). Vet. Microbiol. 86, 131–137.

Klein, J., Bontrop, R.E., Dawkins, R.L., Erlich, H.A., Gyllensten, U.B., Heise, E.R., Jones, P.P., Parham, P., Wakeland, E.K., Watkins, D.I., 1990. Nomenclature for the major histocompatibility complexes of different species: a proposal. Immunogenetics 31, 217–219.

Lane, A., Cheng, Y., Wright, B., Hamede, R., Levan, L., Jones, M., Ujvari, B., Belov, K., 2012. New insights into the role of MHC diversity in devil facial tumour disease. PLoS One 7, e36955.

Larsen, P.A., Heilman, A.M., Yoder, A.D., 2014. The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. BMC Genomics 15, 720.

Lipscomb, T.P., Scott, D.P., Garber, R.L., Krafft, A.E., Tsai, M.M., Lichy, J.H., Taubenberger, J.F., Schulman, F.Y., Gulland, F.M.D., 2000. Common metastatic carcinoma of California Sea lions (Zalophus californianus): evidence of genital origin and association with novel gammaherpesvirus. Vet. Pathol. 37, 609–617.

Maibach, V., Hans, J.B., Hvilsom, C., Marques-Bonet, T., Vigilant, L., 2017. MHC class I diversity in chimpanzees and bonobos. Immunogenetics 69, 661–676.

Mashoof, S., Pohlenz, C., Chen, P.L., Deiss, T.C., Gatlin 3rd, D., Buentello, A., Criscitiello, M.F., 2014. Expressed IgH mu and tau transcripts share diversity segment in ranched Thunnus orientalis. Dev. Comp. Immunol. 43, 76–86.

Morris, K., Austin, J.J., Belov, K., 2013. Low major histocompatibility complex diversity in the Tasmanian devil predates European settlement and may explain susceptibility to disease epidemics. Biol. Lett. 9, 20120900.

Mwangi, W., de Figueiredo, P., Criscitiello, M.F., 2016. One health: addressing global challenges at the nexus of human, animal, and environmental health. PLoS Pathog. 12, e1005731.

Neefjes, J., Jongsma, M.L., Paul, P., Bakke, O., 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat. Rev. Immunol. 11, 823–836.

Ohta, Y., McKinney, E.C., Criscitiello, M.F., Flajnik, M.F., 2002. Proteasome, transporter associated with antigen processing, and class I genes in the nurse shark Ginglymostoma cirratum: evidence for a stable class I region and MHC haplotype lineages. J. Immunol. 168, 771–781.

Ono, Y., Asai, K., Hamada, M., 2013. PBSIM: PacBio reads simulator–toward accurate genome assembly. Bioinformatics 29, 119–121.

Orange, J.S., Ballas, Z.K., 2006. Natural killer cells in human health and disease. Clin. Immunol. 118, 1–10.

Pan, H.J., Wan, Q.H., Fang, S.G., 2008. Molecular characterization of major histocompatibility complex class I genes from the giant panda (Ailuropoda melanoleuca). Immunogenetics 60, 185–193.

Parham, P., Ohta, T., 1996. Population biology of antigen presentation by MHC class I molecules. Science 272, 67–74.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341.

Reddy, M., Dierauf, L., Gulland, F., 2001. Marine mammals as sentinels of ocean health. In: LA, D., FMD, G. (Eds.), Marine Mammals Medicine, 2nd Edition Ed. CRC Press, Boca Raton FL.

Roberts, R.J., Carneiro, M.O., Schatz, M.C., 2013. The advantages of SMRT sequencing. Genome Biol. 14, 405.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., Jaffe, D.B., 2013. Characterizing and measuring bias in sequence data. Genome Biol. 14, R51.

Savage, A.E., Zamudio, K.R., 2011. MHC genotypes associate with resistance to a frog-killing fungus. Proc. Natl. Acad. Sci. U. S. A. 108, 16705–16710.

Savage, A.E., Muletz-Wolz, C.R., Campbell Grant, E.H., Fleischer, R.C., Mulder, K.P., 2019. Functional variation at an expressed MHC class IIbeta locus associates with Ranavirus infection intensity in larval anuran populations. Immunogenetics 71, 335–346.

Sommer, S., 2005. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front. Zool. 2, 16.

Spurgin, L.G., Richardson, D.S., 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc. Biol. Sci. 277, 979–988.

Turner, T.R., Hayhurst, J.D., Hayward, D.R., Bultitude, W.P., Barker, D.J., Robinson, J., Madrigal, J.A., Mayor, N.P., Marsh, S.G.E., 2018. Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 International HLA and Immunogenetics Workshop cell lines. HLA 91, 88–101.

Viluma, A., Mikko, S., Hahn, D., Skow, L., Andersson, G., Bergstrom, T.F., 2017. Genomic structure of the horse major histocompatibility complex class II region resolved using PacBio long-read sequencing technology. Sci. Rep. 7, 45518.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Topfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162.

Westbrook, C.J., Karl, J.A., Wiseman, R.W., Mate, S., Koroleva, G., Garcia, K., Sanchez-Lockhart, M., O'Connor, D.H., Palacios, G., 2015. No assembly required: full-length MHC class I allele discovery by PacBio circular consensus sequencing. Hum. Immunol. 76, 891–896.