MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 103–116 doi: 10.1111/1755-0998.12027

Comparison of 454 pyrosequencing methods for characterizing the major histocompatibility complex of nonmodel species and the advantages of ultra deep coverage

REBEKAH A. OOMEN,* ROXANNE M. GILLETT† and CHRISTOPHER J. KYLE†
*Department of Biology, Dalhousie University, Halifax, Nova Scotia B3H 4J1, Canada, †Forensic Science Department, Trent University, Peterborough, Ontario K9J 7B8, Canada

Abstract

Characterization and population genetic analysis of multilocus genes, such as those found in the major histocompatibility complex (MHC) is challenging in nonmodel vertebrates. The traditional method of extensive cloning and Sanger sequencing is costly and time-intensive and indirect methods of assessment often underestimate total variation. Here, we explored the suitability of 454 pyrosequencing for characterizing multilocus genes for use in population genetic studies. We compared two sample tagging protocols and two bioinformatic procedures for 454 sequencing through characterization of a 185-bp fragment of MHC DRB exon 2 in wolverines (*Gulo gulo*) and further compared the results with those from cloning and Sanger sequencing. We found 10 putative DRB alleles in the 88 individuals screened with between two and four alleles per individual, suggesting amplification of a duplicated DRB gene. In addition to the putative alleles, all individuals possessed an easily identifiable pseudogene. In our system, sequence variants with a frequency below 6% in an individual sample were usually artefacts. However, we found that sample preparation and data processing procedures can greatly affect variant frequencies in addition to the complexity of the multilocus system. Therefore, we recommend determining a per-amplicon-variant frequency threshold for each unique system. The extremely deep coverage obtained in our study (approximately 5000×) coupled with the semi-quantitative nature of pyrosequencing enabled us to assign all putative alleles to the two DRB loci, which is generally not possible using traditional methods. Our method of obtaining locus-specific MHC genotypes will enhance population genetic analyses and studies on disease susceptibility in nonmodel wildlife species.

Keywords: 454 sequencing, *Gulo gulo*, major histocompatibility complex, multilocus genes, pyrosequencing, wolverine

Received 2 May 2012; revision received 7 September 2012; accepted 11 September 2012

Introduction

The major histocompatibilty complex (MHC) is the genetic basis for pathogen resistance in vertebrates (Klein 1986) and the most polymorphic region described in the vertebrate genome (Hughes & Nei 1989; Klein *et al.* 2007). Bernatchez & Landry (2003) compiled compelling evidence that the MHC represents the best system available to investigate how natural selection can promote local adaptation in vertebrates at the gene level. Additional support for this argument has more recently been obtained from a variety of taxa, including birds (Ekblom *et al.* 2007), fish (Eizaguirre & Lenz 2010) and mammals (Vassilakos *et al.* 2009). Evidence of local adaptation to specific diseases (e.g. raccoon rabies virus; Srithayakumar *et al.* 2011), parasites (Kloch *et al.* 2010) and habitats (Cammen *et al.* 2011; McCairns *et al.* 2011) implicate these factors as major drivers of spatial patterns of MHC variation. High levels of variation in MHC genes have been interpreted as adaptive by allowing populations to respond to new and rapidly evolving pathogens and parasites via immunological responses and disease resistance (Yuhki & O'Brien 1990; Langefors *et al.* 2001). Conversely, low levels of genetic variation at the MHC have been linked with increased disease susceptibility and parasite loads (Paterson *et al.* 1998; Lenz *et al.* 2009).

There is a shift towards using functional genetic markers as tools to evaluate levels of intraspecific differentiation and define units for conservation (Vasemägi & Primmer 2005; Hansen 2010). Tracking the distribution and expression of functional genes in the environment will probably be more effective in determining how species adapt to changes in local (e.g. disease outbreaks,

Correspondence: Rebekah A. Oomen, Fax: +1-902-494-3736; E-mail: rebekahoomen@gmail.com

habitat alterations) and global (e.g. climate change) selective pressures (Zeisset & Beebee 2010).

Wolverines (*Gulo gulo*) are vagile, elusive carnivores of conservation concern (Lofroth & Krebs 2007). The current North American distribution of wolverines includes populations in the north-western United States that have been petitioned for endangered species status (US Fish & Wildlife Service) and, in Canada, both a western population of special concern (extending from the west coast of Canada to Ontario and the far North) and a functionally extirpated eastern population with an endangered status (in Quebec and Labrador; COSEWIC 2003; Fortin *et al.* 2005). Due to their extensive range, wolverines occupy a variety of habitat types, from mountains to tundra to boreal forest. Threats to wolverines also vary across their distribution, and include exploitation from the fur trade, predator control programs, and severe habitat loss and fragmentation resulting from anthropogenic activities and climate change (Schreiber *et al.* 1989; Wilson *et al.* 2000). Varying selective pressures across the range may promote adaptive variation in genes such as those in the MHC.

Although examining genetic variability in functional markers is important for assessing local adaptation in wildlife species, and MHC genes provide a means of doing so, MHC characterization in nonmodel species is rare because of the technical difficulties associated with profiling this region. Standardized profiling methods exist for some model organisms (Bunce *et al.* 1995); however, accurate genotyping for the majority of species remains challenging. For most wildlife systems, the traditional approach to detecting functional gene variants is to clone and Sanger sequence multiple PCR amplicons of the target gene. Multiple clones per sample must be sequenced to evaluate PCR artefacts that can confound interpretations and to ensure that all genetic variants present within an individual are identified. MHC genes are particularly difficult to characterize as they are multiplicated in many species [e.g. beluga whales (*Delphinapterus leucas*), Murray & White 1998; bank voles (*Clethrionomys glareolus*), Axtner & Sommer 2007; collared flycatchers (*Ficedula albicollis*), Zagalska-Neubauer *et al.* 2010] and the number of copies can even vary among populations within species [e.g. wild boars (*Sus scrofa*), Barbisan *et al.* 2009; threespine sticklebacks (*Gasterosteus aculeatus*), Eizaguirre *et al.* 2011]. Variable copy numbers of MHC genes is likely due to relatively frequent gene duplications and deletions (Nei *et al.* 1997), possibly caused by the trade-off between having a greater number of alleles (and consequently being able to defend against a greater number of pathogens) and the costs associated with them, such as a reduced T-cell repertoire when individual allelic variation is high (see the optimality hypothesis, Nowak *et al.* 1992; Milinski 2006). Multiple loci result in a greater maximum number of alleles per individual and a consequent increase in the number of clones that need to be sequenced to capture all of the allelic variation within an individual. For example, raccoons (*Procyon lotor*) have a duplicated MHC Class II DRB exon 2, requiring the sequencing of 16 clones per individual to have 96% confidence that all four possible alleles were detected (Castillo *et al.* 2010; Srithayakumar *et al.* 2011). Overall, cloning and Sanger sequencing of MHC genes is time-consuming and costly, making it difficult or practically impossible to obtain the high sample sizes required for many studies, particularly for those elucidating spatial patterns of genetic variation. Further, it is generally difficult to assign alleles to their respective loci to improve the power and resolution of genetic analyses.

To avoid the cost of cloning large numbers of individuals, several profiling techniques based on conformational screening of genetic variation have been developed. These include single-strand conformation polymorphism (Bryja *et al.* 2005; McCairns *et al.* 2011), denaturing gel gradient electrophoresis (Langefors *et al.* 2000) and reference strand-mediated conformational polymorphism (Kennedy *et al.* 2005). These systems are not only time-consuming, but are technically problematic, have limited resolution to distinguish between alleles and are often unreliable when genetic variability is high (Babik *et al.* 2009). More recently, the use of the massively parallel 454 sequencing method (Margulies *et al.* 2005) for MHC genotyping of nonmodel vertebrates has been explored in the literature (Babik *et al.* 2009; Galan *et al.* 2010; Zagalska-Neubauer *et al.* 2010). As sequences are derived from a single DNA molecule, this method is equivalent to sequencing clonally amplified products. Further, by using individually coded primers for PCR, amplifications of multiple individuals can be sequenced simultaneously and individual-based data separated after sequencing is complete (Binladen *et al.* 2007; Meyer *et al.* 2008; Babik *et al.* 2009). The high redundancy of data relative to cloning allows for the purging of sequence artefacts (Moore *et al.* 2006; Brockman *et al.* 2008; Galan *et al.* 2010). Multiple bioinformatics pipelines are being developed to increase the ease and efficiency of obtaining genotypes from data sets containing hundreds of thousands of sequences (Meglécz *et al.* 2010; Stuglik *et al.* 2011). Multilocus genotypes obtained *via* 454 sequencing have been verified by limited cloning (Kloch *et al.* 2010; Promerová *et al.* 2012) and replicate genotyping (Galan *et al.* 2010; Kloch *et al.* 2010; Promerová *et al.* 2012). Therefore, the use of 454 sequencing technology as a method to profile MHC in large-scale studies is promising and would benefit from further validation and a comparison of various technical and analytical methods being developed.

Here, we used DNA samples obtained from across the western Canadian range of wolverines to sequence a 185-bp fragment of MHC DRB exon 2 using 454 sequencing technology accompanied by traditional cloning and Sanger sequencing. Our main objectives were to: (i) evaluate sample preparation and data processing procedures for 454 sequencing of MHC compared with traditional cloning and Sanger sequencing methods and (ii) characterize the MHC DRB gene in wolverines for use in future studies of adaptive variation. We aim to evaluate the viability of the 454 sequencing approach for future profiling efforts of wolverines and other nonmodel species of interest to facilitate large-scale wildlife studies in evolutionary ecology and conservation.

Methods

Sample collection

This study combines samples collected and extracted by Kyle & Strobeck (2001, 2002) and Zigouris *et al.* (2012) from across the western Canadian range of wolverines. A total of 88 individual samples were used in this study, originating from British Columbia (n = 20), Yukon (n = 5), Northwest Territories (n = 18), Manitoba (n = 18) and Ontario (n = 27). All samples were collected between 1962 and 2009 and were either (i) tissue samples collected opportunistically from incidental mortalities, (ii) pelt samples obtained from trapper harvests, or (iii) hair samples collected using noninvasive hair snares. Samples were extracted using a DNEasy Blood & Tissue kit (QIAGEN) according to the manufacturer's instructions and microsatellite markers were used to confirm that each sample represented a different individual (see Zigouris *et al.* 2012 for additional details).

454 Sequencing

A 185-bp fragment of MHC class II DRB exon 2 was amplified using the primer pair DRB-5c (TCAATGGGACGGAGCGGGTGC; Gillett 2009) and DRB-3c (CCGCTGCACAGTGAAACTCTC; Murray & White 1998). There was no prior knowledge on DRB variability and duplication for wolverines. The amplified fragment was prepared for 454 Titanium sequencing in two ways using blunt-ended libraries. The first method (A) was used on a subset of 10 individuals, whereas the second method (B) was used on the remaining 78 samples (Table S1, Supporting information).

Method A. Amplicon libraries were produced using the modified primers A-DRB-5c and B-DRB-3c, which consisted of a 19-mer adaptor required for emulsion PCR and 454 sequencing added to the 5′ end of the DRB primers (GCCTCCCTCGCGCCATCAG for A-DRB-5c and GCCTTGCCAGCCCGCTCAG for B-DRB-3c; Table S2, Supporting information for fusion primer sequences). The key sequence TCAG at the 3′ ends of adaptors A and B was used as a quality control measure to validate the reads during the BaseCall step. Amplification consisted of a 50 μL reaction containing 1× Q-solution (QIAGEN), 1× PCR Buffer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, 0.55 μM of each HPLC purified primer, 0.1 U/μL *Taq* DNA Polymerase (QIAGEN) and 20 ng of DNA with the following cycling conditions: 94 °C for 5 min, 35 cycles of 94 °C for 30 s, 61 °C for 1 min, and 72 °C for 1 min, followed by a final extension of 60 °C for 45 min. PCR product was purified and concentrated using a QIAGEN MinElute PCR Purification Kit following the manufacturer's instructions and quantified using PicoGreenTM fluorescence enhancement (Molecular Probes). Quantified product was mailed on ice to the Génome Québec Innovation Center at McGill University. Prior to sequencing, the amplicons were re-amplified using primers composed of (i) Roche/454 Titanium sequencing primers; a distinct 10 bp Multiplex Identifier (MID) adaptor developed by Roche Diagnostics for use in the 454 GS FLX Titanium Chemistry (added to the 5′ end of the A primer only), and (ii) the appropriate FLX adaptor sequence. The resulting bar-coded, Titanium-suitable amplicons were quantified using PicoGreenTM, pooled in equimolar concentrations for emulsion PCR, and sequenced in the forward strand orientation from DRB-5c to DRB-3c on 1/8th of a run using a 454 GS FLX system (Roche Diagnostics).

Method B. Amplicon libraries were produced using modified reverse primer B-DRB3c MID, consisting of a 6-mer (CTATGC) at the 5′ end, a 19-mer required for use in emulsion PCR and 454 sequencing (GCCTTGCCAGCCCGCTCAG), and the DRB-3c primer (Table S2, Supporting information). Each sample in a run was amplified using a unique forward primer consisting of a 6-mer (CGTATC) at the 5′ end, a 19-mer (GCCTCCCTCGCGCCATCAG), a unique 10-bp MID adaptor (MID1-MID8, MID10-MID11, MID13-MID16; Roche Diagnostics), and the DRB-5c primer. Amplification consisted of a 50 μL reaction containing 1× Q-solution (QIAGEN), 1× PCR Buffer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, 0.45–0.55 μM of each HPLC purified primer, 0.1 U/μL *Taq* DNA Polymerase (QIAGEN) and 40 ng of DNA with the following cycling conditions: 94 °C for 5 min, 35 cycles of 94 °C for 30 s, 61 or 63 °C for 1 min, and 72 °C for 1 min, followed by a final extension of 60 °C for 45 min. PCR product was purified using a QIAGEN MinElute PCR Purification Kit following the manufacturer's instructions, quantified using PicoGreenTM fluorescence enhancement and standardized to 30 ng/μL. A portion

of the standardized product was mailed on ice to the Génome Québec Innovation Center at McGill University where it was amplified by emulsion PCR and sequenced in the forward strand orientation from DRB-5c to DRB-3c on 4/8th's of a run (12–14 samples per 1/8th of a plate) using a 454 GS FLX system (Roche Diagnostics). Remaining samples were run in-house (Natural Resources DNA Profiling and Forensics Centre) using a Roche 454 GS Junior System. A maximum of only 14 MID tags were used per run and tag sequences differed by at least 6/10-bp, making misassignment of reads to amplicons due to sequencing errors unlikely.

Data processing using a four-step procedure. Data generated from 454 sequencing was analysed in two ways. First, a subset of 20 samples were processed following a four-part stepwise procedure based on Galan *et al.* (2010) to detect and discard the majority of reads that exhibited sequencing errors or represented nontarget genes. Briefly, the steps are as follows:

Step 1: Bar-coded amplicons were initially separated into samples using 454 software. Reads with incomplete primers or barcodes, or containing indels that were not multiples of 3 bp were removed from the data set. The remaining reads were filtered using Sequencer (an early version of the jMHC software; Stuglik *et al.* 2011). Reads that were not an exact match to the forward primer or < 200 bp in length were removed from the data set. Variants that occurred only once in an individual amplicon were then removed from the data set to facilitate bioinformatics by improving resolution when examining variants according to their frequency in a particular amplicon.

Step 2: The T_1 threshold developed by Galan *et al.* (2010) represents the minimum number of sequences per amplicon necessary for reliable genotyping, or the number of sequences necessary to achieve a 99.9% probability of amplifying, at least three times, all variants of a gene in a particular sample. This threshold depends on the number of copies of the gene of interest, not previously known for DRB in wolverines. However, due to the extreme depth of sequences obtained in our study, the T_1 thresholds were surpassed for the number of gene copies expected based on characterized DRB genes in other mammals. For example, T_1=46 for a duplicated gene such as DRB in sea otters (*Enhydra lutris*; Bowen *et al.* 2006) and raccoons (*Procyon lotor*; Castillo *et al.* 2010; Srithayakumar *et al.* 2011) and the lowest number of sequences obtained for a single amplicon in our study was 1605 after initial screening processes in step 1.

Step 3: The T_2 threshold was designed to eliminate artefactual variants arising from substitution errors based on the frequency of a variant within a sample (F_{ij}) and the assumption that artefactual variants should occur at lower frequencies than true variants. This threshold is expected to vary depending on the complexity of the system (e.g. the number of gene copies). Galan *et al.* (2010) recommended a T_2 of 4% on a sample-by-sample basis for their rodent system based on a plot of the distribution of F_{ij}, which resulted in 95% reproducible genotyping. To assess the appropriateness of this threshold value for our system and to evaluate the complexity of our system we plotted the distribution of the variant frequencies per individual sample (F_{ij}) and compared it with the expected distribution of a single copy (maximum number of alleles [m] = 2), duplicated (m = 4), and triplicated (m = 6) gene in a diploid species. All variants with $F_{ij} < T_2$ in each sample were considered artefactual within that sample and were removed from that sample.

Step 4: All remaining variants were aligned together as well as separately for each sample and visually inspected in MEGA v. 4.1 (Kumar *et al.* 2008) as both nucleotide and amino acid sequences. We identified and eliminated pseudogenes based on the presence of indels or stop codons and eliminated PCR chimeras based on their distinction of always co-occurring with both parental sequences of higher frequencies in the same sample. Remaining variants were considered to be putative alleles.

Data processing using SESAME. The second method of data processing was undertaken for 78 samples using SESAME (Sequence Sorter & Amplicon Explorer), developed by Meglécz *et al.* (2010). Unlike the previous approach, which utilizes sequences that have already been separated based on their sample-specific MID tags, SESAME requires the complete raw data file for a run and then assigns reads to loci and individuals. We used a wolverine MHC DRB exon 2 sequence (*Gugu-DRB*04*) confirmed via cloning as a marker reference sequence. Sequences are assigned to samples based only on perfect tag matches, although primer mismatches are allowed. Variants are aligned for each sample individually using MUSCLE (Edgar 2004). SESAME removes the primer and tag sequences and provides statistics about each variant to aid in allele validation, such as the frequency of the variant in that particular sample (F_{ij}), the number of sequences of that variant in the sample and in the run, and the number of samples containing that variant in the run. SESAME automatically incorporates the T_1 threshold described earlier and red flags samples with inadequate numbers of sequences.

SESAME then allows the user to call alleles based on their own criteria. First, we used a conservative 8% T_2

threshold and called all variants with $F_{ij} > T_2$ as putative alleles. Due to our deep sequencing and high proportion of low-frequency variants in our data set, we scrutinized the few variants with a frequency of 4–8% and found they could be easily distinguished into two categories. First, there were variants that were only found in one sample and were identical to another putative allele in the same sample except for errors associated with homopolymer runs (a well-established issue with 454 Titanium technology; Moore *et al.* 2006; Brockman *et al.* 2008; Babik *et al.* 2009). These variants, as well as those with frequencies of less than 4%, were classified as artefactual variants. Second, variants with nucleotide differences from other alleles in the same sample outside of a poly-G region and were validated in other samples based on $F_{ij} > 8\%$ were called putative alleles.

Similar to Step 4 mentioned earlier, all putative alleles were aligned together in MEGA v. 4.1 (Kumar *et al.* 2008), as well as separately for each sample automatically in SESAME, to facilitate manual identification and elimination of pseudogenes and PCR chimeras. We performed a protein–protein BLAST search of the National Center for Biotechnology Information database (Altschul *et al.* 1990, http://blast.ncbi.nlm.nih.gov/Blast.cgi) for each translated allele to determine amino acid sequence identity to other mammalian DRB genes.

Cloning

Using the same primer pair as mentioned-above (DRB-5c and DRB-3c), a fragment of DRB exon 2 was amplified for cloning and Sanger sequencing in the same 20 individuals that were analysed via the stepwise procedure (Table S1, Supporting information). PCR amplification consisted of a 15 μL reaction containing 1× Q-solution (QIAGEN), 1× PCR Buffer (200 mM Tris–HCl [pH 8.4], 500 mM KCl), 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, 0.45 μM of each HPLC purified primer, 0.05 U/μL *Taq* DNA Polymerase (QIAGEN) and 20 ng of DNA with the following cycling conditions: 94 °C for 5 min, 35 cycles of 94 °C for 30 s, 59 °C for 1 min, and 72 °C for 1 min, followed by a final extension of 60 °C for 45 min. PCR product was cloned using a TOPO TA Cloning Kit (Invitrogen) according to the manufacturer's instructions with the following modifications: 0.8 μL of vector and DNA ligation was 30 min at room temperature. After an overnight incubation at 37 °C, colonies were picked from plates and added to 50 μL of 0.1× TE. Samples were boiled (100 °C for 10 min) and amplified using primers M13F and M13R (Invitrogen) to confirm that the PCR product was inserted. Amplification consisted of a 10 μL reaction (1× Q-solution, 1× PCR Buffer, 0.04 mM of each dNTP, 1.5 mM $MgCl_2$, 0.20 μM of each primer, 0.05 U/μL *Taq* DNA Polymerase and 2 μL of cloned product) with the following cycling conditions: 95 °C for 5 min, 30 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s, followed by a final extension of 72 °C for 2 min. Amplified product was visualized on a 1.5% agarose gel stained with ethidium bromide. PCR products containing inserts of the correct size were purified using ExoSap-IT (New England Biolabs) following the manufacturer's instructions, and sequenced using a BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). PCR product from the first amplification was also sequenced directly to confirm the sequence variation observed in the clones. A total of 326 clones were sequenced with between 7 and 35 clones sequenced per sample (mean = 16). PCR product was electrophoresed and visualized on an ABI 3730 DNA Analyzer (Applied Biosystems).

Sequences were edited and aligned using MEGA v. 4.1 (Kumar *et al.* 2008). Sequences were accepted as putative alleles if they were identified from multiple clones in more than one individual or multiple clones in one individual from two separate PCRs and therefore unlikely to be the result of PCR error. After pseudogenes and PCR chimeras were eliminated, putative alleles were compared to those determined using 454 sequencing.

Characterization

Genotyping and locus assignment. Between two and four alleles were detected per individual, suggesting that two loci were amplified. Due to the semi-quantitative nature of pyrosequencing, we were able to distinguish whether individuals were homozygous or heterozygous at a given locus based on the frequency of allelic variants in each sample (F_{ij}). We observed that in samples with four alleles, all alleles were present at about equal frequencies (e.g. 1×). In samples with three alleles, one allele was present at approximately twice the frequency (2×) of the remaining two alleles (1×). In samples with two alleles, both alleles had a frequency of 2×. This pattern was observed in all samples and the sum of the frequencies of all putative alleles in a sample equalled 4×. This evidence is consistent with alleles at a locus for which an individual is homozygous producing a frequency of 2× and alleles at a locus for which an individual is heterozygous producing a frequency of 1×. There was no evidence to suggest alleles may be shared among loci (e.g. alleles with frequencies > 2×); however, further genotyping is needed to validate this assumption. Based on individuals who were presumed homozygous at a locus, we were able to assign all alleles to the two loci, thereby creating complete diploid genotypes for all individuals. We used Arlequin v. 3.11 (Excoffier *et al.* 2005) to perform a Monte Carlo approximation of Fisher's exact test for linkage disequilibrium between the two loci ($\alpha = 0.05$).

Tests for selection. Allelic nomenclature was based on rules set by Klein *et al.* (1990). We identified sites putatively associated with the peptide binding region (PBR) based on the human MHC II molecular structure (Brown *et al.* 1993; Stern *et al.* 1994) and tested for selection on each locus separately. We computed the following statistics in MEGA for the entire fragment, the PBR only, and the non-PBR only: (i) Average pairwise nucleotide distance using the Kimura 2 parameter model (K2P); (ii) Poisson-corrected amino acid distance; (iii) Average rate of synonymous (d_S) and nonsynonymous (d_N) substitutions per site (modified Nei–Gojobori method with the Jukes–Cantor correction for multiple substitutions; Nei & Gojobori 1986). The 1000 bootstrap replicates were used to obtain standard errors for each statistic. We performed a one-tailed *Z*-test in MEGA using rates of d_N and d_S calculated under models of neutrality and positive selection to test methods of selection acting on DRB exon 2 in wolverines.

Phylogenetic analysis. We determined the best model of nucleotide substitution to be F81 + G by evaluating Akaike's Information Criterion (Akaike 1974) in jModelTest v. 0.1.1 (Posada 2008). The Markov chain Monte Carlo analysis in MrBayes v. 3.1 (Ronquist & Huelsenbeck 2003) was used for Bayesian inference of phylogeny. Analyses were run using 10 million generations and a sample frequency of 5000, with 25% discarded as burn-in. We used SplitsTree4 (Huson & Bryant 2006) to visualize the resulting unrooted phylogenetic tree. European mink (*Mustela lutreola*; accession number EU263556.1; Becker *et al.* 2009), sea otter (*Enhydra lutris*; accession number EU121855.1; Aguilar *et al.* unpublished data), and raccoon (*Procyon lotor*; accession number GU388377.1; Castillo *et al.* 2010) DRB exon 2 sequences were included based on high sequence identity to the wolverine DRB alleles to assess the relationship among these mammalian DRB genes. These species have either one (European mink; Becker *et al.* 2009) or two (sea otter Bowen *et al.* 2006 and raccoon Castillo *et al.* 2010) DRB loci.

Results

454 sequencing

A 185-bp region of MHC DRB exon 2 was amplified and sequenced using a 454 Titanium sequencer for 88 wolverines. A total of 183 655 sequences were obtained for 20 individuals that were analysed using a four-part stepwise procedure based on Galan *et al.* (2010). 135 824 sequences (74%) were retained following the initial screening processes in step 1, of which 31 853 (17% of all reads) had a per-amplicon frequency of 1 and were subsequently removed. Following step 1, the average coverage was 5199 ± 1586 reads per amplicon (range = 3512–8219). Due to this extremely high coverage, all samples were assumed to surpass the T_1 threshold. Therefore, no samples were removed during step 2. The distribution of variant frequencies per sample (Fig. 1) showed the highest proportion of variants having a frequency of 0–2%, with a lack of variants in the frequency range of 30–100%. This distribution is consistent with a high heterogeneity of alleles at a locus that is multiplicated. The abundance of variants found at extremely low frequencies represents artefactual variants due to PCR or sequencing errors. Based on the variant frequency category with the fewest number of sequences, we determined the T_2 threshold separating artefactual variants from true alleles in our data set to be about 4–6%. Considering only those variants with either $F_{ij} > 6\%$ or $F_{ij} < 6\%$ and confirmed *via* cloning (of which there were 11; Table 1), nine putative alleles were identified. Upon alignment, one variant that was present in all individuals was identified as a pseudoallele based on the fact that it contained a single base pair deletion immediately

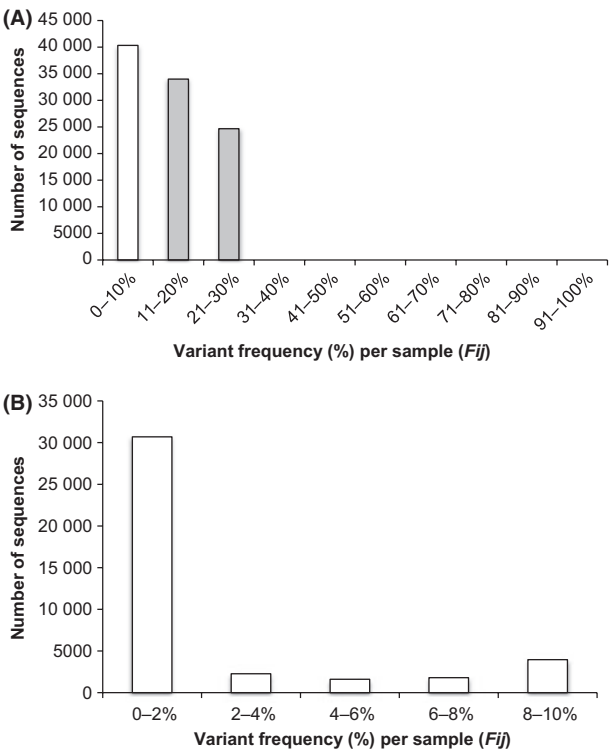

Fig. 1 The distribution of the frequency of each variant *j* in each sample *i* (F_{ij}) for frequencies of (A) 0–100% and (B) 0–10% obtained from 454 sequencing of wolverine (*Gulo gulo*) MHC DRB exon 2 alleles and analysed following methods by Galan *et al.* (2010). Variants with a per-amplicon frequency of one were not included.

Table 1 Per-amplicon variant frequencies (F_{ij}) of major histocompatibility complex DRB exon 2 alleles and a pseudoallele obtained *via* 454 sequencing and stepwise analysis with tagging method A or B for 20 wolverines (*Gulo gulo*)

Individual	*Gugu-DRB1*02*	*Gugu-DRB1*04*	*Gugu-DRB2*01*	*Gugu-DRB2*05*	*Gugu-DRB2*06*	*Gugu-DRB2*08*	*Gugu-DRB2*09*	*Gugu-DRB2*11*	*Gugu-DRB*03*
Method A									
ON-298		18.1 (3)			11.4 (4)				9.4 (2)
ON-36734		16.8 (3)			4.1 (3)*		4.9 (1)*		9.7 (3)
ON-36735	5.4 (6)*	7.6 (0)		6.6 (5)	4.0 (6)*				8.3 (5)
ON-36736	16.1 (1)		19.3 (3)						7.5 (1)
ON-36739	6.4 (4)	7.7 (3)	9.6 (3)				5.0 (2)*		8.0 (1)
ON-36740	17.5 (3)		17.5 (7)						8.1 (6)
ON-806	14.7 (1)		9.6 (4)			5.6 (3)*			7.0 (4)
ON-811	14.9 (10)		10.6 (3)	5.1 (6)*					9.3 (4)
ON-860		15.8 (13)			2.9 (3)**		3.7 (7)**		7.4 (3)
ON-875	6.2 (3)	8.3 (6)		5.6 (1)*		4.0 (3)*			11.1 (11)
Method B									
NWT-2215	11.2 (2)	10.4 (0)	12.6 (5)				11.3 (0)		11.3 (4)
NWT-2216	23.9 (2)			25.0 (1)					11.9 (1)
NWT-2217	13.4 (0)	10.3 (0)	13.5 (1)			13.0 (2)			6.5 (2)
NWT-2218	11.8 (2)	12.6 (1)		10.7 (2)				9.2 (0)	11.5 (4)
NWT-2219	13.1 (3)	10.2 (0)	14.4 (2)				13.9 (1)		6.2 (8)
BC-236	24.6 (0)		14.2 (3)	11.9 (1)					8.2 (1)
BC-237	24.4 (5)			11.4 (0)		12.0 (2)			7.99 (2)
BC-246	21.6 (5)					22.0 (3)			13.1 (5)
BC-247	10.5 (3)	9.7 (1)				23.9 (6)			12.3 (6)
BC-202	12.4 (1)	11.5 (1)				23.3 (4)			11.8 (2)

The number of clones of the same sequence obtained *via* Sanger sequencing are in parentheses. Variants with low sample frequencies are indicated by asterisks (* $\leq$ 6%, ** $\leq$ 4%).

following the primer sequence that resulted in a frameshift mutation and two premature stop codons. There were no indels or premature stop codons detected in the remaining variants and no alleles displayed the pattern characteristic of PCR chimeras. Therefore, eight putative alleles were retained in step 4.

The 510 651 reads from 454 sequencing of 78 samples were inputted in SESAME. Following assignment of reads to the marker reference sequence and sample MID tags, a total of 428 324 sequences were aligned and trimmed. The average coverage per sample was 5491 (± 1704) and ranged from 2241 to 10 330; therefore, all samples surpassed the T_1 threshold. Variants with $F_{ij} > 8\%$ were considered putative alleles. Of the eight variants that had a sample frequency of 4–8%, three satisfied the criteria of (i) containing nucleotide differences from other alleles in the same sample outside of a poly-G region and (ii) being validated in other samples based on $F_{ij} > 8\%$. These three variants were considered putative alleles at frequencies of 4%, 6%, and 8%. The five variants that did not meet these criteria had frequencies of 4–5% and were eliminated along with all variants with $F_{ij} < 4\%$. Again, we found and eliminated a pseudoallele that was detected in all samples (identical to that detected by the stepwise procedure above). No other putative alleles were eliminated, resulting in ten putative alleles, eight of which were identical to those detected in the 20-sample subset by the stepwise procedure.

Of the two methods of sample tagging used for 454 sequencing, sample variant frequencies for putative alleles tended to be higher using Method B (Table 1). For example, there were 11 incidences of putative alleles with $F_{ij} < 6\%$ in the ten samples prepared using Method A and zero incidences in the ten samples prepared using Method B.

Cloning

Cloning and Sanger sequencing of 20 individuals failed to detect nine alleles observed using pyrosequencing, leading to incorrect genotypes for 7/20 individuals (Table 1). In one individual (ON-860), two alleles observed using cloning were also detected using 454 sequencing, but at sample frequencies lower than 4%. Nine alleles observed using cloning were observed at per sample variant frequencies of 4–6%. All 11 incidences of $F_{ij} < T_2$ for cloned alleles occurred in samples prepared using Method A.

Characterization

Ten unique alleles were found at the two DRB loci amplified in this study as well as one pseudoallele that was

Table 2 Amino acid sequences of major histocompatibility complex DRB exon 2 alleles in wolverines (*Gulo gulo*)

Allele			*		*		*					*	*										*															*
*Gugu-DRB1*02*	F	L	D	R	Y	F	Y	N	R	E	E	Y	V	R	F	D	S	D	V	G	E	Y	R	P	V	T	E	L	G	R	P	D	A	Q	Y	W	N	S	Q	K
*Gugu-DRB1*04*	.	.	E	.	H	V	.	.	G	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.
*Gugu-DRB2*01*	.	.	E	.	H	V	.	.	G	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
*Gugu-DRB2*05*	.	.	E	.	H	V	.	.	G	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
*Gugu-DRB2*06*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.
*Gugu-DRB2*07*	L	.	V	.	W	I	.	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.
*Gugu-DRB2*08*	L	.	V	.	W	I	.	.	.	.	.	.	L	.	.	.	.	.	.	.	.	F	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.
*Gugu-DRB2*09*	L	.	V	.	W	I	.	.	.	.	.	.	L	.	.	.	.	.	.	.	.	F	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.
*Gugu-DRB2*10*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.
*Gugu-DRB2*11*	.	.	E	.	H	V	.	.	G	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.

Allele				*		*			*		*					*								*								*							*
*Gugu-DRB1*02*		D		I		L		E		Q		K		R		A		N		V		D		T		Y		C		R		H		N		Y		G		V		G
*Gugu-DRB1*04*		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.
*Gugu-DRB2*01*		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.
*Gugu-DRB2*05*		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		V
*Gugu-DRB2*06*		.		.		.		.		.		R		.		.		E		.		.		.		V		.		.		.		.		.		.		V
*Gugu-DRB2*07*		.		.		.		.		.		R		.		.		E		.		.		.		V		.		.		.		.		.		.		V
*Gugu-DRB2*08*		.		.		.		.		.		R		.		.		E		.		.		.		.		.		.		.		.		.		.		.
*Gugu-DRB2*09*		.		.		.		.		.		R		.		.		.		.		.		.		.		.		.		.		.		.		.		.
*Gugu-DRB2*10*		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.
*Gugu-DRB2*11*		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.		.

Dots indicate sequence identity to the top allele. Putative PBR sites (*) are based on Brown *et al.* (1993) and Stern *et al.* (1994).

Table 3 Average nucleotide and amino acid distances among wolverine major histocompatibility complex DRB exon 2 alleles in percentages per site, average rates of nonsynonymous substitutions per synonymous site (d_N) and synonymous substitutions per synonymous site (d_S) in percentages, and *Z*-tests of positive selection on all sites, the peptide binding region (PBR) only, and the non-PBR only

Sites	K2P nucleotide distance	Poisson-corrected amino acid distance	d_N	d_S	Z	P
DRB1						
All	5.3 (1.7)	9.7 (3.8)	5.2 (2.2)	5.5 (3.4)	−0.087	1.000
PBR	9.3 (3.9)	19.1 (9.6)	10.7 (5.6)	5.5 (5.7)	0.761	0.224
Non-PBR	3.2 (1.6)	4.9 (3.7)	2.3 (1.5)	5.6 (4.3)	−0.730	1.000
DRB2						
All	6.9 (1.4)	12.1 (3.3)	7.6 (2.1)	5.3 (2.1)	0.900	0.185
PBR	15.2 (3.6)	28.3 (8.6)	18.8 (6.3)	6.7 (4.0)	2.045	0.022
Non-PBR	2.8 (1.1)	4.5 (2.6)	2.1 (1.2)	4.6 (2.5)	−0.846	1.000

The K2P model and Poisson distribution were used to correct for multiple substitutions when calculating the nucleotide and amino acid distances, respectively. Standard errors (in parentheses) were obtained through 1000 bootstrap replicates. Putative PBR sites are based on Brown *et al*. (1993) and Stern *et al*. (1994).

detected in all individuals (Table 2; GenBank Accession numbers JX409655–JX409665). Individuals possessing four alleles were the most common (44/88; 50%), while 25/88 (28%) had three alleles and 19/88 (22%) had two alleles. No alleles appeared to be shared among loci. Two alleles were assigned to DRB1 and eight alleles were assigned to DRB2. An exact test for linkage disequilibrium suggests DRB1 and DRB2 may be physically linked (Exact $P = 0.000 \pm 0.000$).

Of the 185 nucleotides, 26 (14.1%) were variable, as were 13/61 (21.3%) amino acid sites. Sequence divergence among putative alleles ranged from 1 to 17 nucleotides and 0 to 12 amino acids. Correcting for the frameshift mutation, the pseudoallele (*Gugu-DRB*03*) was substantially diverged from all putative alleles by a minimum of 20 nucleotide substitutions. Putative alleles were presumed functional based on (i) a lack of indels or stop codons and (ii) extremely high amino acid sequence identity (87–95%) to functional mammalian DRB alleles; however, expression studies will be necessary to confirm functionality. We calculated average pairwise K2P nucleotide distances and Poisson-corrected amino acid distances for all sites, as well as for the PBR and non-PBR separately (Table 3). A *Z*-test for positive selection showed a significant excess of nonsynonymous substitutions in the PBR only for DRB2 ($Z = 2.05$, $P = 0.022$; Table 3). The test for positive selection was not significant for the PBR in DRB1 ($Z = 0.76$, $P = 0.224$); however, the number of nonsynonymous substitutions was nearly double that of synonymous substitutions and the power of the test was limited because there were only two alleles. A phylogeny constructed using Bayesian inference did not reveal distinct lineages and failed to confidently resolve the relationships between the putative alleles and the pseudoallele or alleles from other species (Fig. 2).

Discussion

This study represents the first exploration of MHC variation in wolverines, which will permit future investigations into understanding the effects of local selective pressures on adaptive variation in this species of conservation concern. We determined that the wolverine DRB exon 2 is duplicated, with two putatively linked polymorphic loci. Rates of nonsynonymous substitutions were 1.9 and 2.8 times greater than synonymous substitutions at the peptide binding region for DRB1 and DRB2, respectively (Table 3), a difference similar to that observed for raccoons when utilizing the same peptide binding sites (in which d_N was 2.6 times greater than d_S for the two loci combined; Castillo *et al*. 2010). The significant excess of nonsynonymous substitutions in the peptide binding region only is indicative of positive selection acting on the MHC in wolverines.

Multiple MHC class II loci are common [e.g. sea otters, Bowen *et al*. 2006; raccoons, Castillo *et al*. 2010; grey seals (*Halichoerus grypus*), Cammen *et al*. 2011]. These loci are often functional and allow for the detection of a greater number of pathogens by increasing the number of alleles an individual possesses (Hughes 1994). Despite the duplication of the DRB gene in wolverines, we found a relatively low number of DRB exon 2 alleles compared with other terrestrial mammals. Ten DRB alleles were found in 88 wolverines compared with 34 alleles in 50 red deer (Swarbrick *et al*. 1995) and 66 in 246 raccoons (Castillo *et al*. 2010). High levels of allelic diversity in DRB exon 2 are thought to be maintained by a

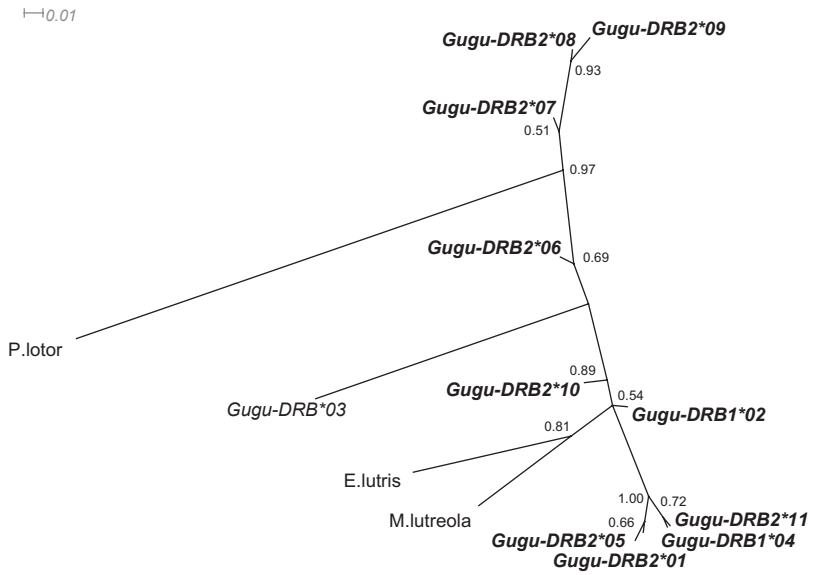

Fig. 2 Phylogenetic relationships among ten wolverine DRB exon 2 alleles (in bold) and one pseudoallele using Bayesian inference and the best-fit model of nucleotide substitution based on jModelTest (Posada 2008). European mink (*Mustela lutreola*; accession number EU263556.1; Becker *et al.* 2009), sea otter (*Enhydra lutris*; accession number EU121855.1; Aguilar *et al.* unpublished data), and raccoon (*Procyon lotor*; accession number GU388377.1; Castillo *et al.* 2010) DRB exon 2 sequences were included as outgroups. Bootstrap values < 0.5 are not shown.

form of balancing selection, such as overdominance (Hughes & Nei 1989) or frequency-dependent selection (Clarke & Kirby 1966). The low level of allelic variation we found in wolverines is similar to species that are thought to have experienced past population bottlenecks, such as moose (*Alces alces*; Ellegren *et al.* 1996) and Commander Arctic foxes (*Vulpes lagopus*; Ploshnitsa *et al.* 2011). While a bottleneck is not documented for Canadian wolverines, it remains a possibility. Alternatively, extremely low densities (approximately 1 wolverine per 200 km^2 in the range core; Quick 1953; Becker & Gardner 1992; Lee & Niptanatiak 1993) and a primarily northern range occupied by fewer pathogens may play a role in alleviating disease-related selective pressures. Further, the moderate level of sequence variation we observed in wolverine DRB alleles (Table 2) may permit recognition of a broader array of pathogens according to the divergent allele advantage hypothesis (Wakeland *et al.* 1990a; Lenz 2011), thus potentially decreasing the intensity of balancing selection on these loci. Lastly, the alleles we detected are limited by the primers we employed, which is an important limitation of any PCR-based approach to characterization.

The failure of Bayesian inference to resolve phylogenetic relationships among DRB alleles of wolverines and related carnivores is unsurprising given the high level of polymorphism exhibited in this relatively short sequence (Fig. 2). However, the phylogeny suggests that DRB1 and DRB2 may not represent distinct allelic lineages and that wolverine alleles do not represent distinct lineages from alleles of other mammals, an observation consistent with trans-species polymorphism (the maintenance of ancient lineages in natural populations; Wakeland *et al.* 1990b).

Recent advances in sequencing technology allow for large-scale genetic analyses of nonmodel species at a low cost and high efficiency compared with traditional methods (Wegner 2009). However, further investigation is needed to maximize the potential of next-generation sequencing technologies. Here, we evaluated two sample tagging protocols and two bioinformatic methods for 454 Titanium sequencing and compared the results with traditional cloning and Sanger sequencing for characterizing the wolverine MHC DRB exon 2 locus.

We obtained similar results from the stepwise procedure (Galan *et al.* 2010) and SESAME (Meglécz *et al.* 2010) with respect to the alleles that were validated and the depth of coverage per amplicon (5199 ± 1586 and 5491 ± 1704, respectively). A greater proportion of reads were eliminated following step 1 in this study (43%) compared to Galan *et al.* (2010) (33%). However, due to the extremely high number of reads per amplicon, we eliminated all variants with a per-amplicon frequency of one, whereas Galan *et al.* (2010) only eliminated singletons (i.e. variants that occurred only once in the entire data set) at this step. Although meaningful comparisons are difficult given slight variations in data processing and reporting, the average coverage per amplicon in our study (5199 ± 1586) was at least an order of magnitude higher compared with previous 454 sequencing of MHC in nonmodel species [e.g. 93.6 ± 44.4 (Babik *et al.* 2009), 541 ± 166 (Zagalska-Neubauer *et al.* 2010), 301.4 ± 89.2 (Promerová *et al.* 2012)]. Substantially varied coverage among amplicons is not unusual despite careful standardization of sample concentrations prior to sequencing (Brockman *et al.* 2008; Babik *et al.* 2009).

Although our deep coverage allowed us to surpass any conceivable T_1 threshold for our species, one must

be cognizant of the role of pseudogenes in determining this value. Pseudogenes are common in class II multigene families (Yuhki *et al.* 2003; Zagalska-Neubauer *et al.* 2010). The co-amplification of a pseudogene in our study reduced the proportion of reads representing true alleles. Therefore, the T_1 value for a maximum number of alleles (m) of six ($T_1 = 74$; Galan *et al.* 2010) seems more appropriate for our system than the T_1 threshold of 46 for a duplicated locus (where $m = 4$), essentially conservatively treating the pseudogene like an additional locus.

A substantial challenge with high-throughput sequencing data lies in distinguishing true alleles from artefactual variants. Babik *et al.* (2009) used an average per-individual frequency of 3% to distinguish true alleles from artefacts in bank voles, whereas Galan *et al.* (2010) used a 4% variant frequency threshold on a sample-by-sample basis to identify true alleles. Analysing each sample separately allows us to reliably validate rare variants, which are important to population genetic studies, and distinguish between the same variant occurring as an artefact in one sample and as a true allele in another (Galan *et al.* 2010). In our system, we found that variants with a frequency of 4–8% in a sample could be confidently called as alleles when they had nucleotide differences not associated with homopolymer runs and were validated with a $F_{ij} > 8\%$ in other samples. For efficiency in large-scale studies using this system, we recommend a T_2 threshold of 6%. This cut-off would have conservatively resulted in one true allele being discarded for the 78 samples prepared using tagging Method B in this study. The pseudoallele (*Gugu-DRB*03*) was an exception, which passed the stringent threshold criteria and was detected at high frequencies (approximately 6–14%) in all samples. However, its characteristic indel and otherwise highly diverged sequence makes it easily identifiable during the manual inspection step. A lower threshold would be more appropriate for samples prepared using Method A, given the generally lower variant frequencies; however, this would increase the chance of falsely identifying variants as true when they are not. The lower frequencies obtained using Method A may be due to the extra PCR step used to assemble the tags, which could have increased the proportion of PCR artefacts that were sequenced, thus decreasing the proportion of reads representing true alleles. Given the generally higher frequencies obtained for putative alleles using Method B, which involved a single PCR step, we recommend this method for 454 sequencing sample preparation.

It is also important to consider the impact of different data processing procedures on the accuracy of the T_2 threshold. For example, removing singletons in the stepwise procedure inflates the frequency of remaining variants. Therefore, we argue that the threshold not only varies according to the complexity of the multilocus system (Babik *et al.* 2009; Galan *et al.* 2010), but also the sample preparation and bioinformatic methods. We recommend evaluating the T_2 threshold for each new method-system combination using the graphical procedure described by Galan *et al.* (2010) combined with an assessment of those variants suspected to be near the threshold according to our criteria described previously.

Previous studies sequenced a high number of clones in a single individual and confirmed all alleles obtained from clones using 454 sequencing (Kloch *et al.* 2010; Promerová *et al.* 2012). In our study, where we cloned several individuals, all alleles determined using cloning were obtained *via* 454 sequencing as well, albeit at varying per sample variant frequencies, but the reverse was not true. The alleles that were successfully cloned but failed to reach the T_2 threshold when pyrosequenced may have been detected at a higher sample frequency if the sample had been tagged using Method B instead of Method A, as there were no occurrences of this nature when Method B was used. The failure of cloning to detect all alleles in our study is not surprising, given that the average 16 clones/individual that we sequenced represents the minimum number of clones required to be 96% confident of detecting all alleles at a duplicated locus, assuming no preferential amplification or cloning of alleles (Castillo *et al.* 2010). More extensive cloning is necessary in our system because 23% of clones represented a pseudogene that was co-amplified (Table 1).

Multiple loci resulting from recent duplications often share allelic lineages (Ellis *et al.* 1999; van Oosterhout *et al.* 2006), necessitating simultaneous amplification of more than one locus (Babik *et al.* 2009). Without prior knowledge, it is generally not possible to assign alleles to loci or determine zygosity. This information is extremely useful for population genetic estimates and a lack thereof can hinder the ability of researchers to detect patterns of genetic structure. For example, levels of genetic differentiation may be underestimated if monomorphic loci are unknowingly included in computations. Further, the locus-specific genotype of an individual can be more powerful than just a list of an individual's alleles when evaluating associations with disease susceptibility and resistance. We used a novel method of assigning alleles to multiplicated loci by taking advantage of the semi-quantitative nature of 454 Titanium sequencing that has been shown to be reliable when comparing read abundances within the same species (Amend *et al.* 2010). Using the ratios of variant frequencies in each sample, we inferred the zygosity of each individual at each locus and used individuals who were homozygous at a locus to assign alleles to the two loci amplified in this study. As a result, complete diploid genotypes were obtained for each locus for use in future population genetic studies. This approach may not be appropriate for species

with a much higher number of multiplications, because it becomes increasingly difficult to assign alleles to loci as the number of loci increases, or for those in which alleles are shared among loci. Replicate genotyping of wolverines included in this study and applying this approach to other species will be necessary to confirm our results and determine the universality of our approach.

There is a movement towards maximizing the number of samples that are sequenced on a single 454 run (Galan *et al.* 2010), which is limited by the number of unique identifiers available with which to label samples. This approach is more cost-effective by reducing the amount of machine time required (the primary expense of 454 sequencing; Wegner 2009), although the cost of primer synthesis increases as more unique sequence tags are needed. Another major cost associated with maximizing the capacity of a single 454 run is a decrease in coverage per amplicon. Reduced coverage decreases the likelihood of obtaining the minimum number of sequences per amplicon that are necessary for reliable genotyping (the T_1 threshold; Galan *et al.* 2010). Due to the extremely high coverage obtained in our study, no samples were eliminated due to insufficient coverage. We found no evidence to suggest that the frequency of artefactual variants is increased at extremely high levels of coverage. Therefore, our ability to distinguish between true alleles and artefacts was not impeded, as artefactual variants generally occurred at very low per-amplicon variant frequencies and exhibited other characteristics that were distinct from true alleles. Further, high coverage per sample enabled the assignment of alleles to loci through consistent per-amplicon variant frequency ratios reflecting zygosity of individuals due to semi-quantitative amplification. The use of per-amplicon variant frequencies for this purpose would likely be less reliable when sample coverage is low, because the ratios would be more strongly influenced by a sampling bias. However, the >5000× coverage in this study is likely unnecessary to maintain quantitative assessment of alleles for genotyping. Given this, and the fact that $T_1 = 74$ for our system (Galan *et al.* 2010), a coverage of 100–500× would likely be sufficient. Therefore, we have since moved from approximately 14 to 96 MID-tagged samples for subsequent runs on our Roche 454 GS Junior System for population genetic analysis of DRB exon 2 in wolverines. Further, we plan to sequence several MHC loci simultaneously to better characterize local adaptation in this species given the capacity of 454 sequencing relative to the coverage needed for semi-quantitative assessment of sequences for genotype assignment.

Funding

This research was supported by the Mountain Equipment Co-op, Ontario Ministry of Natural Resources and by the Natural Sciences and Engineering Research Council of Canada through a Discovery Grant to Dr. C.J. Kyle and a USRA to R.A. Oomen.

Acknowledgements

We thank K. Austen and L. Skitt from the Ontario Ministry of Natural Resources, Red Lake District, and D. Berezanski from Manitoba Conservation for providing wolverine samples from Ontario and Manitoba. V. Srithayakumar and M. Harnden provided essential laboratory and technical assistance. We also thank J. Morris-Pocock, J. Zigouris, D. Keith, P. Debes, N. Roney, S. Castillo and three anonymous reviewers for providing helpful discussion and comments.

References

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403–410.

Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology*, **19**, 5555–5565.

Axtner J, Sommer S (2007) Gene duplication, allelic diversity, selection processes and adaptive value of MHC class II DRB genes of the bank vole, *Clethrionomys glareolus*. *Immunogenetics*, **59**, 417–426.

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources*, **9**, 713–719.

Barbisan F, Savio C, Bertorelle G, Patarnello T, Congiu L (2009) Duplication polymorphism at MHC class II DRB1 locus in the wild boar (*Sus scrofa*). *Immunogenetics*, **61**, 145–151.

Becker EF, Gardner CL (1992) *Wolf and Wolverine Density Estimation Techniques*. State of Alaska, Dept. of Fish and Game, Division of Wildlife Conservation, Juneau, AK.

Becker L, Nieberg C, Jahreis K, Peters E (2009) MHC class II variation in the endangered European mink *Mustela lutreola* (L. 1761)-consequences for species conservation. *Immunogenetics*, **61**, 281–288.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Binladen J, Gilbert MT, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.

Bowen L, Aldridge BM, Miles AK, Stott JL (2006) Expressed MHC class II genes in sea otters (*Enhydra lutris*) from geographically disparate populations. *Tissue Antigens*, **67**, 402–408.

Brockman W, Alvarez P, Young S *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, **18**, 763–770.

Brown JH, Jardetzky TS, Gorga JC *et al.* (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature*, **364**, 33–39.

Bryja J, Galan M, Charbonnel N, Cosson J-F (2005) Analysis of major histocompatibility complex class II gene in water voles using capillary electrophoresis-single stranded conformation polymorphism. *Molecular Ecology Notes*, **5**, 173–176.

Bunce M, Fanning GC, Welsh KI (1995) Comprehensive, serologically equivalent DNA typing for HLA-B by PCR using sequence-specific primers (PCR-SSP). *Tissue Antigens*, **45**, 81–90.

Cammen K, Hoffman JI, Knapp LA, Harwood J, Amos W (2011) Geographic variation of the major histocompatibility complex in Eastern Atlantic grey seals (*Halichoerus grypus*). *Molecular Ecology*, **20**, 740–752.

Castillo S, Srithayakumar V, Meunier V, Kyle CJ (2010) Characterization of major histocompatibility complex (MHC) DRB exon 2 and DRA exon 3 fragments in a primary terrestrial rabies vector (*Procyon lotor*). *PLoS ONE*, **5**, e12066.

Clarke B, Kirby DR (1966) Maintenance of histocompatibility polymorphisms. *Nature*, **211**, 999–1000.

COSEWIC (2003) *COSEWIC Assessment and Update Status Report on the Wolverine Gulo gulo in Canada*. Committee on the Status of Endangered Wildlife in Canada, Ottawa, Ontario. vl + 44 pp.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Eizaguirre C, Lenz TL (2010) Major histocompatibility complex polymorphism: dynamics and consequences of parasite-mediated local adaptation in fishes. *Journal of Fish Biology*, **77**, 2023–2047.

Eizaguirre C, Lenz TL, Sommerfeld RD, Harrod C, Kalbe M, Milinski M (2011) Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three-spined stickleback ecotypes. *Evolutionary Ecology*, **25**, 605–622.

Ekblom R, Saether SA, Jacobsson P *et al*. (2007) Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Molecular Ecology*, **16**, 1439–1451.

Ellegren H, Mikko S, Wallin K, Andersson L (1996) Limited polymorphism at major histocompatibility complex (MHC) loci in the Swedish moose *A. alces*. *Molecular Ecology*, **5**, 3–9.

Ellis SA, Holmes EC, Staines KA *et al*. (1999) Variation in the number of expressed MHC genes in different cattle class I haplotypes. *Immunogenetics*, **50**, 5–6.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Fortin C, Banci V, Brazil J *et al*. (2005) *National Recovery Plan for the Wolverine (Gulo gulo) [Eastern Population]*. National Recovery Plan No. 26. Recovery of Nationally Endangered Wildlife (RENEW), Ottawa, Ontario.

Galan M, Guivier E, Caraux G, Charbonnel N, Cosson J-F (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, **11**, 296.

Gillett RM (2009) *MHC Genotypes and Reproductive Success in the North Atlantic Right Whale (Eubalaena glacialis)*. PhD thesis. Trent University, Peterborough, Ontario.

Hansen MM (2010) Expression of interest: transcriptomics and the designation of conservation units. *Molecular Ecology*, **19**, 1757–1759.

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, **256**, 119–124.

Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 958–962.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

Kennedy LJ, Quarmby S, Fretwell N *et al*. (2005) High-resolution characterization of the canine DLA-DRB1 locus using reference strand-mediated conformational analysis. *Journal of Heredity*, **96**, 836–842.

Klein J (1986) *Natural History of the Major Histocompatibility Complex*. Wiley, New York.

Klein J, Bontrop RE, Dawkins RL *et al*. (1990) Nomenclature for the major histocompatibility complexes of different species: a proposal. *Immunogenetics*, **31**, 217–219.

Klein J, Sato A, Nikolaidis N (2007) MHC, TSP, and the origin of species: from immunogenetics to evolutionary genetics. *Annual Review of Genetics*, **41**, 281–304.

Kloch A, Babik W, Bajer A, Siński E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole *Myodes glareolus*. *Molecular Ecology*, **19**, 255–265.

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, **9**, 299–306.

Kyle CJ, Strobeck C (2001) Genetic structure of North American wolverine (*Gulo gulo*) populations. *Molecular Ecology*, **10**, 337–347.

Kyle CJ, Strobeck C (2002) Connectivity of peripheral and core populations of North American wolverines. *Journal of Mammalogy*, **83**, 1141–1150.

Langefors Å, Lohm J, von Schantz T, Grahn M (2000) Screening of Mhc variation in Atlantic salmon (*Salmo salar*): a comparison of restriction fragment length polymorphism (RFLP), denaturing gradient gel electrophoresis (DGGE) and sequencing. *Molecular Ecology*, **9**, 215–219.

Langefors Å, Lohm J, Grahn M, Andersen Ø, von Schantz T (2001) Association between major histocompatibility complex class IIB alleles and resistance to *Aeromonas salmonicida* in Atlantic salmon. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, **268**, 479–485.

Lee J, Niptanatiak A (1993) *Ecology of the Wolverine on the Central Arctic Barrens: Progress Report, Spring 1993*. Department of Renewable Resources, Government of NWT, Yellowknife, NWT.

Lenz TL (2011) Computational prediction of Mhc Ii-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution*, **65**, 2380–2390.

Lenz TL, Wells K, Pfeiffer M, Sommer S (2009) Diverse MHC IIB allele repertoire increases parasite resistance and body condition in the Long-tailed giant rat (*Leopoldamys sabanus*). *BMC Evolutionary Biology*, **9**, 269.

Lofroth EC, Krebs J (2007) The abundance and distribution of wolverines in British Columbia, Canada. *The Journal of Wildlife Management*, **71**, 2159–2169.

Margulies M, Egholm M, Altman WE *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

McCairns RJ, Bourget S, Bernatchez L (2011) Putative causes and consequences of MHC variation within and between locally adapted stickleback demes. *Molecular Ecology*, **20**, 486–502.

Meglécz E, Piry S, Desmarais E *et al*. (2010) SESAME (SEquence Sorter & AMplicon Explorer): genotyping based on high-throughput multiplex amplicon sequencing. *Bioinformatics*, **27**, 277–278.

Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nature Protocols*, **3**, 267–278.

Milinski M (2006) The major histocompatibility complex, sexual selection, and mate choice. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 159–186.

Moore MJ, Dhingra A, Soltis PS *et al*. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, **6**, 17.

Murray BW, White BN (1998) Sequence variation at the major histocompatibility complex DRB loci in beluga (*Delphinapterus leucas*) and narwhal (*Monodon monoceros*). *Immunogenetics*, **48**, 242–252.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, **3**, 418–426.

Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 7799–7806.

Nowak MA, Tarczy-Hornoch K, Austyn JM (1992) The optimal number of major histocompatibility complex molecules in an individual. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10896–10899.

van Oosterhout C, Joyce DA, Cummings SM (2006) Evolution of MHC class IIB in the genome of wild and ornamental guppies, *Poecilia reticulata*. *Heredity*, **97**, 111–118.

Paterson S, Wilson K, Pemberton JM (1998) Major histocompatibility complex variation associated with juvenile survival and parasite resistance in a large unmanaged ungulate population (*Ovis aries* L.). *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3714–3719.

Ploshnitsa AI, Goltsman ME, Macdonald DW, Kennedy LJ, Sommer S (2011) Impact of historical founder effects and a recent bottleneck on MHC variability in Commander Arctic foxes (*Vulpes lagopus*). *Ecology and Evolution*, **2**, 165–180.

Posada D (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Promerová M, Babik W, Bryja J, Albrecht T, Stuglik M, Radwan J (2012) Evaluation of two approaches to genotyping major histocompatibility complex class I in a passerine-CE-SSCP and 454 pyrosequencing. *Molecular Ecology Resources*, **12**, 285–292.

Quick HF (1953) Wolverine, fisher, and marten studies in a wilderness region. *Transactions of the North American Wildlife Conference*, **18**, 513–532.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Schreiber A, Wirth R, Riffel M, Van Rompaey H (1989) *Weasels, Civets, Mongooses, and Their Relatives: An Action Plan for the Conservation of Mustelids and Viverrids*. IUCN/SSC Mustelid & Viverrid Specialist Group, IUCN, Gland, Switzerland.

Srithayakumar V, Castillo S, Rosatte RC, Kyle CJ (2011) MHC class II DRB diversity in raccoons (*Procyon lotor*) reveals associations with raccoon rabies virus (*Lyssavirus*). *Immunogenetics*, **63**, 103–113.

Stern LJ, Brown JH, Jardetzky TS *et al*. (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature*, **368**, 215–221.

Stuglik MT, Radwan J, Babik W (2011) jMHC: Software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Molecular Ecology Resources*, **11**, 739–742.

Swarbrick PA, Schwaiger FW, Epplen JT, Buchan GS, Griffin JF, Crawford AM (1995) Cloning and sequencing of expressed DRB genes of the red deer (*Cervus elaphus*) Mhc. *Immunogenetics*, **42**, 1–9.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Vassilakos D, Natoli A, Dahlheim M, Hoelzel AR (2009) Balancing and directional selection at exon-2 of the MHC DQB1 locus among populations of Odontocete cetaceans. *Molecular Biology and Evolution*, **26**, 681–689.

Wakeland EK, Boehme S, She JX *et al*. (1990a) Ancestral polymorphisms of MHC class II genes: divergent allele advantage. *Immunologic Research*, **9**, 115–122.

Wakeland EK, Boehme S, She JX (1990b) The generation and maintenance of MHC class II gene polymorphism in rodents. *Immunology Reviews*, **113**, 207–226.

Wegner KM (2009) Massive parallel MHC genotyping: titanium that shines. *Molecular Ecology*, **18**, 1818–1820.

Wilson GM, Van Den Bussche RA, Kennedy PK, Gunn A, Poole K (2000) Genetic variability of wolverines (*Gulo gulo*) from the Northwest Territories, Canada: conservation implications. *Journal of Mammalogy*, **81**, 186–196.

Yuhki N, O'Brien SJ (1990) DNA recombination and natural selection pressure sustain genetic sequence diversity of the feline MHC class I genes. *The Journal of Experimental Medicine*, **172**, 621–630.

Yuhki N, Beck T, Stephens RM, Nishigaki Y, Newmann K, O'Brien SJ (2003) Comparative genome organization of human, murine, and feline MHC class II region. *Genome Research*, **13**, 1169–1179.

Zagalska-Neubauer M, Babik W, Stuglik M, Gustafsson L, Cichoń M, Radwan J (2010) 454 sequencing reveals extreme complexity of the class II Major Histocompatibility Complex in the collared flycatcher. *BMC Evolutionary Biology*, **10**, 395.

Zeisset I, Beebee TJC (2010) Larval fitness, microsatellite diversity and MHC class II diversity in common frog (*Rana temporaria*) populations. *Heredity*, **104**, 423–430.

Zigouris J, Dawson NJ, Bowman J, Gillett RM, Schaefer JA, Kyle CJ (2012) Genetic isolation of wolverine (*Gulo gulo*) populations at the eastern periphery of their North American distribution. *Conservation Genetics*, DOI 10.1007/s10592-012-0399-x.

Conceived and funded the project: C.J.K. Designed the project: R.A.O., R.M.G. and C.J.K. Provided samples, reagents, and equipment: C.J.K. Conducted laboratory analyses: R.A.O. and R.M.G. Analysis of cloning data: R.M.G. Analysis of 454 sequencing data: R.A.O. Wrote the manuscript: R.A.O. (methods by R.A.O. and R.M.G.). Edited the manuscript: C.J.K.

Data accessibility

DNA sequences: GenBank Accessions JX409655–JX409665; DRYAD entry doi:10.5061/dryad.s5b40.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 List of genotyped samples, including whether they were cloned and which sample tagging method and bioinformatics pipeline(s) was/were used.

Table S2 Descriptions and sequences of primers used to amplify DRB exon 2.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.