Preliminary characterization of little brown bats (*Myotis lucifugus*) immune MHC II DRB alleles using next-generation sequencing

Jon M. Palmer, Leah K. Berkman, Paula E Marquardt, Deahn M. Donner, Michelle A. Jusino, Daniel L. Lindner

Genes of the major histocompatibility complex (MHC) are of great interest to molecular ecologists due to their role in the immune response and the complex evolutionary mechanisms affecting their abundance and distribution. Next generation sequencing (NGS) offers the best potential for characterizing such genes by overcoming the low throughput constraints associated with traditional cloning and Sanger sequencing. However, available bioinformatics methods to effectively filter MHC data acquired through NGS are complex and have not been widely applied. Here we present a method to sequence MHC II DRB of the little brown bat (*Myotis lucifugus*) using semi-conductor sequencing on the Ion Torrent Personal Genome Machine (PGM) and subsequently analyze the data using a modified UPARSE clustering algorithm. As proof of concept, preliminary characterization of the MHC II DRB alleles of 15 little brown bats was accomplished through PGM sequencing. Analysis identified 24 MHC II DRB alleles overall and a high number of alleles per individual (average 6.8 ± 3.5, range 1 – 12), indicating little brown bats may have among the highest number of MHC DRB loci documented in mammals. Our method provides a framework for studying MHC II DRB diversity in little brown bats, which are experiencing significant population declines in North America due to the recent emergence of white-nose syndrome (WNS).
Preliminary characterization of little brown bats (Myotis lucifugus) immune MHC II DRB alleles using next-generation sequencing.

Jonathan M. Palmer¹, Leah K. Berkman², Paula E. Marquardt³, Deahn M. Donner³, Michelle A. Jusino¹, Daniel L. Lindner¹

¹ Center for Forest Mycology Research, US Forest Service, Madison, WI 53726, USA
² College of Environmental Science and Forestry, State University of New York, Syracuse, NY, 13210, USA
³ Institute for Applied Ecosystem Services, US Forest Service, Rhinelander, WI 54501, USA

Corresponding author:
Paula Marquardt³
5985 Highway K, Rhinelander, WI, 54501, USA
Email address: pmarquardt@fs.fed.us
ABSTRACT

Genes of the major histocompatibility complex (MHC) are of great interest to molecular ecologists due to their role in the immune response and the complex evolutionary mechanisms affecting their abundance and distribution. Next generation sequencing (NGS) offers the best potential for characterizing such genes by overcoming the low throughput constraints associated with traditional cloning and Sanger sequencing. However, available bioinformatics methods to effectively filter MHC data acquired through NGS are complex and have not been widely applied. Here we present a method to sequence MHC II DRB of the little brown bat (Myotis lucifugus) using semi-conductor sequencing on the Ion Torrent Personal Genome Machine (PGM) and subsequently analyze the data using a modified UPARSE clustering algorithm. As proof of concept, preliminary characterization of the MHC II DRB alleles of 15 little brown bats was accomplished through PGM sequencing. Analysis identified 24 MHC II DRB alleles overall and a high number of alleles per individual (average 6.8 ± 3.5, range 1 – 12), indicating little brown bats may have among the highest number of MHC DRB loci documented in mammals. Our method provides a framework for studying MHC II DRB diversity in little brown bats, which are experiencing significant population declines in North America due to the recent emergence of white-nose syndrome (WNS).

INTRODUCTION

The multi-gene family of the major histocompatibility complex (MHC), called HLA in humans, is of interest in molecular ecology due to its role in the activation of the adaptive immune system (Hill 2001; Murphy et al. 2007) and the evolutionary mechanisms that generate population level MHC variation (Spurgin & Richardson 2010; Sutton et al. 2011). MHC variants
can be associated with susceptibility or resistance to pathogens, which makes diversity of these
genes important for population viability (Dionne et al. 2009; Savage & Zamudio 2011).
Consequently, many studies of threatened, bottlenecked, or rapidly declining populations have
targeted the MHC to ascertain the loss of genetic diversity and adaptive potential of those
populations (e.g. Haig 1998; Hedrick et al. 2001; Miller et al. 2008; O'Brien et al. 1985).
Genotyping the MHC at the population level presents a challenge due to gene
duplications observed in many taxa (Kelley et al. 2005). Next generation sequencing (NGS)
provides the means to simultaneously analyze multiple MHC loci in many individuals, which
makes genotyping more tractable (Lighten et al. 2014). However, errors associated with
polymerase chain reaction (PCR), which is a requisite procedure for next generation sequencing
of amplicons (PCR products), can be problematic (Sommer et al. 2013). Additionally, primer
bias can cause preferential binding and an uneven degree of amplification among loci and alleles.
Many researchers have developed bioinformatics “pipelines”, or a sequence of filters, to separate
PCR artifacts from real MHC alleles while attempting to avoid discarding low frequency alleles
(Babik et al. 2009; Galan et al. 2010; Pavey et al. 2013; Zagalska-Neubauer et al. 2010). These
pipelines may be effective for the taxon for which they were developed, but none have been
widely accepted or thoroughly tested (Lighten et al. 2014).
Studies of MHC diversity in bats are timely and of particular importance due to the recent
emergence of white-nose syndrome (WNS), a disease of bats caused by the fungus
Pseudogymnoascus destructans (Gargas et al. 2009; Lorch et al. 2011). This disease was first
observed in North American hibernating bats in 2006 (Blehert et al. 2009). The little brown bat
(Myotis lucifugus) is particularly susceptible and has suffered large-scale population decline as a
result of WNS (Blehert et al. 2009; Frank et al. 2014). However, very little is known about the
immunological response of little brown bats to invasion by *P. destructans*. One way in which bats and other mammals detect and respond to extracellular pathogens is through the variable MHC II DRB locus, which encodes for an antigen-binding site responsible for sensing pathogens and activating the adaptive immune system (Sommer 2005). Although the MHC Class II DRB locus has been well-studied in other mammals, little is known about the genes involved in disease resistance in bats, and the MHC Class II DRB is the only MHC gene that has been examined to date (Del Real-Monroy et al. 2014; Mayer & Brunner 2007; Richman et al. 2010; Schad et al. 2012). In this study we built a bioinformatics pipeline based on the Ion Torrent Personal Genome Machine (PGM) sequencing platform and the UPARSE clustering algorithm for studying the MHC II DRB locus in little brown bats (*Myotis lucifugus*). Our methodology uses publically available and widely used software packages (Python and USEARCH) more commonly used to construct *de novo* “taxa” (operational taxonomic units or “OTUs”) in NGS studies of microbial diversity (Caporaso et al. 2010; Edgar 2010; Edgar 2013; Schloss et al. 2009).

**MATERIALS AND METHODS**

Tissue samples were obtained from little brown bats (*M. lucifugus*) live-captured with mist nets from summer roosting sites during June-September 2013 on the Chequamegon-Nicolet National Forest in Wisconsin (n = 8) and the Superior National Forest in Minnesota (n = 7) by US Forest Service employees as part of a larger habitat monitoring program. Because sampling occurred on federal lands by federal employees on a species not federally listed, sampling permits and Institutional Animal Care and Use Committee (IACUC) review were not required. Samples were collected in a humane and ethical manner outlined by the American Society of Mammologists for the safe handling and sampling of bats (Sikes et al. 2011). Mist nets were
operated by trained and experienced employees, at dusk into night, monitored continuously when bats were active, and handled carefully when removing bats to avoid injury to animals. Tissue was extracted from plagiopatagium membranes with 4 mm sterile biopsy punches (Wilmer and Barratt 1996). A biopsy was selected from a region of the wing with few or no visible blood vessels to reduce bleeding; one biopsy was collected from each wing. Tissue was stored in 2.0 mL tubes and shipped on ice packs to the US Forest Service Northern Research Station in Rhinelander, WI. To kill P. destructans or other potentially infectious agents, shipping containers and contents were decontaminated with 10% bleach solution (National WNS decontamination protocol v 06.25.2012; https://www.whitenosesyndrome.org/topics/decontamination). Paper and any material that could not withstand bleach treatment were UV irradiated for 10 minutes. Once received, samples were stored at -20° C in 100% ethanol and DNA was extracted using the DNeasy® Blood and Tissue Kit (Qiagen, Valencia, CA), following the manufacturer’s protocols.

**Next-generation sequencing on the Ion Torrent PGM platform**

Using a combination of BLAST searches and Ensemble tools, 16 putative MHC II DRB paralogs were identified from the little brown bat reference genome (Ensembl 73) (Flicek et al. 2013), and primers were designed in a conserved region of the genome surrounding exon 2 (Supplemental Figure S1). The forward primer was designed based on the EX2F primer used by Richman et al. (2010) for other Myotis spp. and extended into exon 2. Although the area was relatively conserved, 2 primers were designed that differed by 2 base pairs in an attempt to avoid amplification bias. The reverse primer sequences from Richman et al. (2010) were not conserved in little brown bats so a reverse primer was designed using a conserved block of sequence that extended slightly into the other end of exon 2 of MHC II DRB locus. To generate amplicons that
are compatible with the Ion Torrent PGM platform, fusion primers were synthesized that contained Ion P1 and Ion A adapter sequences; additionally, the reverse primer was barcoded (according to protocols distributed from the manufacturer) to allow for multiplex sequencing. The forward primers were synthesized as below (Ion P1 adapter is italicized): Mylu IntF_GC: 5’-CCTCTCTATGGGCAGTCGGTGAT|TGTCCCCGCRCGCGCATTTCTCTG-3’ and Mylu IntF_AA: 5’-CCTCTCTATGGGCAGTCGGTGAT|TGTCCCCGCAGCAAATTTCCTG-3’. The reverse primer contained the Ion A adapter (italicized), the key sequence, a 10 bp barcode sequence (bolded), followed by a single ‘A’ linker sequence: Mylu Int2R: 5’-CCATCTCATCCCTGCCTGTCTCCGAC|TCAG|CTAAGGTAAC|A|GGGTGCTCCTCACAGC CCTGTG-3’.

MHC II DRB loci were amplified in two separate reactions with the reverse primer (Mylu Int2R) and one of the forward primers (Mylu Int1F_GC or Mylu Int1F_AA). To minimize the risk of PCR error from the use of multiple primers at the same site (Kanagawa 2003), each reaction was run separately and the products were combined after PCR. All reactions contained 5U of Platinum Taq Polymerase (Invitrogen), 2 µL of 10X PCR Buffer, 0.4 µM each primer, 0.2 mM dNTPs, 2 mM MgCl₂ and 2-20 ng of template DNA in a total volume of 20 µL. Thermal cycling conditions were: 94⁰ C for 4 min followed by 28 cycles of 30 s at 94⁰ C, 45 s at 58⁰ C, and 90 s at 72⁰ C. PCR products were visualized on a 1.4% agarose gel stained with ethidium bromide and for each individual bat the two products were pooled equally according to band intensity on the gel. The pooled PCR products were then gel-purified by 0.8% CloneWell gels using the E-gel system and subsequently quantified with the Qubit High Sensitivity DNA assay (Thermo-Fisher). The purified PCR products were pooled in equimolar concentrations, attached to the Ion Sphere Particles (ISPs) using the Ion PGM Template OT2 400 kit, and
sequenced on a 314v2 chip using Ion PGM 400 bp Sequencing Kit according to manufacturer’s recommendations (Thermo-Fisher).

**Ion Torrent Data Processing**

Raw sequence data were obtained from the Ion Torrent Server using the BaseCaller option [--disable-all-filters] to keep the barcode sequences of the reads intact and disable 3’ quality filtering in order to prevent trimming of the reverse primer sequence. The unaligned BAM file from the Ion Torrent Server was converted to FASTQ using BedTools (Quinlan & Hall 2010). The reads were then processed using custom python scripts that are available online (https://github.com/nextgenusfs/mhc_cluster) (Figure 1). The first script, “mhc-process_reads.py”, functions to trim forward and reverse primer sequences from the read as well as re-label the header with the corresponding barcode name. Because our primer design resulted in sequencing from the opposite translation orientation of MHC II DRB exon 2, “mhc-process Reads.py” also reverse complements (optional) each read for downstream processing (Figure 1). The output of “mhc-process_reads.py” is a FASTQ file containing only those reads that are full length (i.e. valid barcode and both primer sequences – with barcodes and primers trimmed from the reads).

The second step of our pipeline is to run the “mhc-OTU_cluster.py script” (Figure 1). This script runs a modified UPARSE algorithm that does the following: (1) quality filters the data by removing reads that have an expected error > 1.0 (Edgar & Flyvbjerg 2015), (2) filters reads using a DNA Hidden Markov Model of DRB-exon2 using HMMER version 3.1b (hmm.org) (DNA HMM model for little brown bats was created using a MUSCLE (Edgar 2004) alignment of the exon 2 region of the 16 MHC II DRB paralogs identified from in silico
interrogation of the genome reference), (3) trims or pads (with N’s) each read to a set length, (4) runs the UPARSE algorithm (Edgar 2013) using a default clustering identity setting of 99%, and (5) optionally the script can translate the OTUs into amino acid sequences (Figure 1). The output of “mhc-OTU_cluster.py” is composed of several files including: multi-FASTA file containing OTUs, multi-FASTA file with translated OTUs, and an OTU table. In order to determine the functional variation of the DNA based OTUs, the “mhc-OTU_cluster.py script” can 3-frame translate each OTU to its corresponding amino acid sequence. The amino acid sequences are then aligned to the MHC_II_beta HMM model (pfam00969) using HMMer3, the region that produces a significant alignment is extracted for each OTU, and exact duplicates are concatenated to output unique protein sequences. These python scripts are customizable; for example, they can be used with primer sets for other MHC loci and can be easily adapted to other organisms by updating the HMM models used for PCR contaminant filtering. USEARCH version 8.0.1623 (Edgar 2010) was used for all the analysis in this manuscript and the reads were trim/padded to a set length of 320 base pairs.

RESULTS

A total of 875,516 raw reads were obtained from the Ion PGM (314v2 chip) sequencing run and after filtering sequences that did not have a valid barcode sequence or both forward/reverse primers, 62,009 sequences remained. Low quality sequences were removed by discarding reads with expected errors > 1.0 (Edgar & Flyvbjerg 2015), which resulted in the retention of 21,057 reads. Preliminary processing of the data indicated that some of the sequence data were derived from non-target contaminating sequences presumably due to non-specific amplification during the initial PCR reactions. Therefore, a DNA HMM model was used to quickly filter out contaminating sequences, and this resulted in removal of ~20% of the
sequences. The remaining 16,898 sequences were subsequently trimmed or padded to 320 base pairs and clustered using UPARSE (Edgar 2013) at 99% identity. These parameters would allow for sequences that have more than 4 differences across 320 base pairs to be clustered into discrete OTUs. The number of reads per individual bat that passed all 3 quality control criteria ranged from 1,165 – 15,556 (average ± SD: 4,133.9 ± 3,875.6). The UPARSE algorithm produced 24 OTUs from the 15 individual bats analyzed in this study. We subsequently used presence/absence of reads for each OTU to determine allelic diversity for each individual bat, which ranged from 1 – 12 OTUs per individual bat (average ± SD: 6.8 ± 3.5) (Figure 2A). Translation of these 24 OTUs to amino acid sequences indicated that 20 were unique across MHC II DRB exon 2, indicating that there are at least 20 different functional variants of exon 2 from these 15 individual bats (Figure 2B). Three of the unique translated OTUs had internal stop codons and thus represented either frame-shift mutations or non-functional alleles. Sequence data is available for download thru the NCBI SRA under the SRP067348 accession number.

Interestingly, no MHC II DRB alleles were present in all 15 individuals sampled; the allele with the highest abundance, MyLu-DRB_3, was found in 14 individuals. Moreover, 5 alleles with the lowest abundance were observed in single bats, which represents ~20% of the alleles discovered (MyLu-DRB_8, 14, 17, 22, and 24). Three of the low frequency alleles were found together in one individual bat (MN012; Figure 2A). Taken together with the range of MHC II DRB alleles detected per individual, these data suggest that a larger sampling effort is needed to adequately characterize the MHC II DRB alleles in little brown bats. However, despite the small sample size, there is considerable variation in the amino acid sequence of the MHC II DRB alleles discovered with our methodology (Figure 3).

**DISCUSSION**
We extended the use of an algorithm (UPARSE) designed to filter NGS data to the analysis of MHC genes in little brown bats. Previous NGS MHC pipelines have used complicated filtering parameters to differentiate between “artifact” sequences and “real” sequences; however, UPARSE is very efficient at removing low quality reads through a combination of error filtering and removal of low abundance reads (removing singletons) prior to clustering into OTUs (Edgar 2013). Our method offers a conservative estimate of MHC II DRB diversity, as sequences are clustered together based on a 99% identity threshold. This is notably different from traditional Sanger sequencing methodology and other NGS MHC pipelines, where sequences divergent by a single base pair have been referred to as an independent allele.

However, clustering highly similar MHC II DRB alleles may be more advantageous in downstream analyses because genetic differences among highly similar alleles may not translate into amino acid substitutions in the antigen-binding site, and thus these alleles would likely be functionality identical in terms of their interactions with pathogens. For example, we identified 24 alleles on a DNA level using a clustering threshold of 99%, while 20 were unique on a protein level.

There are some caveats to our technique, as amplicon sequencing on the Ion Torrent PGM is currently limited to regions less than 400 base pairs and thus only allows for characterization of small regions of the MHC II DRB locus (exon 2). As a result, functional frame-shift mutations are difficult to properly characterize with this technique because the sequences in the surrounding regions remain unknown. Moreover, in this study we have only used a single priming site for amplification of MHC II DRB alleles and it has previously been demonstrated that primers situated in conserved regions can be highly biased to a specific locus or allele and can fail to effectively co-amplify all multigene copies (Burri et al. 2014). Given that
NGS approaches can sequence each PCR amplicon thousands of times, allelic dropout, or the failure to detect an allele due to low throughput, is not as problematic with NGS protocols compared to traditional Sanger sequencing approaches (Lighten et al. 2014). While we were successful in this study with only a single priming site in amplifying MHC II DRB loci from little brown bats, careful evaluation of additional conserved primers is warranted for future studies.

The immune systems of bats remain poorly understood (Baker et al. 2013) and there are few studies of bat MHC diversity (Del Real-Monroy et al. 2014; Mayer & Brunner 2007; Richman et al. 2010; Schad et al. 2012). We have provided the first description of MHC II DRB diversity among little brown bats. Little brown bats have a relatively high number of MHC II DRB loci (up to 12 identified herein) compared to other mammals. Humans and other old world primates have among the highest number of characterized MHC II DRB genes with 9, 4 of which are functional (Klein et al. 2007). The number of DRB loci in bats varies considerably among species (Schad et al. 2012). Richman et al. (2010) recovered only 2 – 4 unique sequences (1 – 2 loci) per bat in *M. velifer* and *M. vivesi*, while Schad et al. (2012) found 10 functional DRB loci in an individual greater sac-winged bat (*Saccopteryx bilineata*). Due to our limited sample size (n=15), a definitive characterization of MHC II DRB copy number is not attainable, but the potential exists for little brown bats to have among the highest number of MHC DRB loci among mammals and considerable MHC DRB diversity.

Because MHC is involved in the activation of the adaptive immune system, expansion of MHC genes may have evolved to recognize and respond to a large number of pathogens (Hughes and Nei 1989; Hedrick 1991). Little brown bats also have highly diverse immunoglobulin genes (Bratsch et al. 2011), which indicates that an increased immune gene repertoire may be
characteristic of this species. Although such adaptations to efficiently recognize pathogens may
be highly developed in some species of bats (Schad et al. 2012), recent research suggests bats
may have features that suppress their immune response to infection (Baker et al. 2013; Wynne &
Wang 2013; Zhang et al. 2013). The precipitous decline of North American bats due to WNS has
prompted numerous investigations into how bats respond to this fungal pathogen (e.g., Meteyer
et al. 2011; Moore et al. 2011). One question that is particularly interesting is why North
American bats experience mass mortality when exposed to the fungus causing WNS, while
European bats do not experience mass mortality in response to this disease (Zukal et al. 2014).
Here we identified 20 unique MHC II DRB alleles at the amino acid level, which may be
indicative of a varied immune response to pathogens (perhaps WNS) among individual bats.
Low MHC diversity may be related to infection susceptibility and survival of bottlenecked
populations (Radwan et al. 2010); therefore, future characterization of MHC II DRB diversity in
North American little brown bat populations may offer insight into the susceptibility of these
bats to WNS.

ACKNOWLEDGEMENTS

We are grateful to Timothy Catton, David Grandmaison, Brian Heeringa, Kari Kirschbaum, and
Nicole Shutt (USFS) for assistance in obtaining samples. Anita Foss, Ben Prom, and Mark Banik
(USFS) provided laboratory support; Sue Lietz (USFS) delivered GIS support, while Jacqueline
Frair, ESF Roosevelt Wild Life Station (SUNY), contributed programmatic support.

LITERATURE CITED

genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology*


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

10.1146/annurev.genet.41.110306.130137


10.1038/nature10590


10.1038/sj.hdy.6800989


10.1371/journal.pone.0027430


1

Sequencing and clustering approach to MHC genotyping

Diagram of the methodology used to characterize MHC II DRB alleles from little brown bats using the Ion Torrent PGM semi-conductor sequencing platform. Barcoded Ion Torrent PGM compatible fusion primers are used to amplify a pool of PCR products from each individual bat (1 barcode per individual). The barcoded amplicons are then pooled in equal concentrations and sequenced on the Ion Torrent PGM. The raw sequences are obtained from the Torrent Server and processed with 2 python scripts. The first script, “mhc-process_reads.py”, functions to find only those reads that contain a valid barcode sequence and contain both forward/ reverse primers. The second script, “mhc-OUT_cluster.py”, quality trims the reads, filters the reads for contamination, and then runs the UPARSE clustering algorithm at a default setting of 99% identity. The result of the pipeline is multi-FASTA sequences for OTUs (DNA and amino acid) and an OTU table that can be used to genotype each individual using presence/absence metrics.
MHC II DRB Alleles

PCR with NGS Fusion Primers

Ion Torrent Compatible Amplicons

Sequence on Ion Torrent PGM

Pool Amplicons (one barcode each individual)

**mhc-process_reads.py**

1. Find barcode sequence, relabel header, discard reads w/o barcode sequence
2. Find/Strip forward primer and discard read if not found
3. Find/Strip reverse primer and discard read if not found

**mhc-OTU_cluster.py**

1. Quality Filter: remove reads with expected errors > 1.0 (USEARCH)
2. Remove Contaminants: filter each read against MyLu DRB HMM model (HMMer3)
3. Trim/Pad reads to the same length
4. Run UPARSE Pipeline: (USEARCH)
   - De-replication
   - Sort by Size (frequency)
   - Cluster at 99% identity: create OTUs
   - Map reads back to OTUs
   - Construct OTU table
5. Translate OTUs to protein space and filter with DRB HMM model (optional) (HMMer3)
DNA based Operational Taxonomic Unit Table

Figure 2: (A) Graphical representation of an OTU table using presence/absence metrics illustrates the distribution of DNA-based MHC II DRB alleles among the 15 little brown bats in this study (presence indicated by blue square). Twenty-four MHC II DRB OTUs were discovered with our method with the most abundant allele (MyLu-DRB_3) being found in 14 of 15 individuals from WI and MN. The range of alleles found per individual varied from 1 – 12. (B) Translation of the MHC II DRB OTUs to amino acid sequence resulted in 20 unique protein sequences, which have a similar distribution pattern per individual as the DNA-based OTUs. Sequences marked with an asterisks (*) were truncated at the amino acid level in exon 2 and thus representing either frame-shift mutations or are non-functional alleles.
Protein sequence alignments

Figure 3. Multiple sequence alignment of protein sequences from the 20 MHC II DRB alleles illustrate the functional variability seen in exon 2 from little brown bats. Sequences were aligned with MUSCLE and visualized using Geneious 7.1.8 (BioMatters). The Sequence Logo displays the level of conservation for each residue and completely conserved residues are colored gray.