**Supplementary Information**

*De novo* genotyping of the MHC in an Australian dragon lizard aided by family group data

Hacking, J., Bradford, T., Pierce, K., Gardner, M.

*Introduction: Details on the AmpliSAS genotyping method*

During AmpliSAS clustering genotypes are resolved on an amplicon-by-amplicon basis; taking into account unequal amplification among samples and among alleles, rather than applying a global read depth frequency threshold (first and fourth issue, table 1). Variants are ordered by read depth and each variant either forms a cluster and becomes a dominant sequence, or is deemed an artefact sequence based on clustering parameters. Clustering parameters include user-defined substitution and indel error rates, sequence length parameters, and a dominance frequency threshold at which a sequence within a cluster is allowed to form its own cluster, becoming another dominant sequence. Dominant sequences that then pass user-defined filtering parameters, such as excluding very low frequency variants and chimeras, are considered true alleles. When a variant is classified as an artefact its read depth is added to that of the dominant sequence from which it likely originated. This helps to separate true alleles that have lower read depths and artefact sequences that occur at higher frequencies (second issue, table 1). Finally, the dominance frequency threshold applied within clusters allows the discrimination of true alleles that are relatively similar (third issue, table 1). In addition, the AmpliCHECK pre-genotyping tool, which is also part of the AmpliSAT suite ([Sebastian, Herdegen, Migalska, & Radwan, 2016](#_ENREF_7)), is extremely useful for *de novo* genotyping in non-model systems as it provides preliminary information on amplicon read depth, variant frequencies and estimated error rates.

*Materials and Methods: Illumina MiSeq library preparation*

A two-step PCR protocol was used to prepare a MiSeq sequencing library following [Pearson, Bradford, Ansari, Bull, & Gardner (2016)](#_ENREF_6). MHC-targeting primers were used in the first PCR, followed by a second PCR to add Illumina sequencing adapters and sample indexes. MHC-targeting primers were designed for α1 and α2 based on *C. decresii* transcriptome data ([table S2, Hacking, Bertozzi, Moussalli, Bradford, & Gardner, 2017](#_ENREF_2)) using Primer 3 ver. 2.3.4 ([Untergasser et al., 2012](#_ENREF_9)) in Geneious ver. 8.1.7 ([Kearse et al., 2012](#_ENREF_4)). The MHC-targeting primers and a MRT ([multiplex ready technology, Hayden, Nguyen, Waterman, & Chalmers, 2008](#_ENREF_3)) linker sequence constitute the inner primers. The outer primers consisted of indexes unique to each individual and Illumina-specific adapters with an MRT linker sequence. The inner and outer primers are connected via the MRT primer sequence. A summary of the whole library preparation process, primer design details and all inner and outer primer sequences are provided (figure S1, table S1).

MHC-targeting primer pairs MHC1\_A1\_G234\_F/MHC1\_A1\_G234\_R and MHC1\_A2\_G234\_F/MHC1\_A2\_G12\_R ([Hacking, et al., 2017](#_ENREF_2)) were used to amplify parts of the MHC I α1 (exon 2) and α2 (exon 3) domains in DNA samples for 337 individuals from Hawker, 40 individuals from Mount Remarkable, 27 individuals from the Barossa, 40 individuals from Morialta, and 41 individuals from Kangaroo Island (total of 485). Technical replicates (n = 12) amplified in independent PCR reactions were included in the MiSeq library. A small number of replicates were used as family group data were also available to investigate the accuracy of MHC genotypes. Amplification reactions for the first PCR totalled 12µl and contained 4mM each dNTP, 0.25 mg/ml BSA, 0.4µM forward and reverse primers, 0.5U Immolase enzyme and approximately 20ng of template DNA. Thermal cycling conditions consisted of an initial denaturisation step of 95°C for 10 minutes, 30 cycles of 94°C for 45 seconds, 57°C for 45 seconds and 72°C for 60 seconds, followed by a final extension of 72°C for 6 minutes. After PCR clean-up using MultiScreenHTS 384-well filter plates on a vacuum manifold (Merck Millipore), a second PCR was undertaken to attach the outer primers. Amplification reactions were 12µl and contained 4mM dNTP, 0.25 mg/ml BSA 0.4µM forward and reverse outer primer, 0.5U Immolase enzyme and approximately 20ng of template DNA. Thermal cycling conditions for the second PCR consisted of an initial denaturisation step of 95°C for 10 minutes, 10 cycles of 92°C for 15 seconds, 54°C for 60 seconds and 72°C for 1 minute and 30 seconds, followed by a final extension of 72°C for 10 minutes.

Gel electrophoresis (1.5% agarose) was performed for every amplicon following the second PCR and the resulting bands were scored for band strength (absent, weak, moderate, and strong). TapeStation analysis (Agilent 2200 TapeStation ©; Agilent Technologies) was undertaken on a subset of amplicons from each gel score category to determine the concentration, molarity and the amount of contaminating adapter dimer present at 125bp. Nine individuals did not have visible amplicons on the gel and were excluded from downstream analyses. Amplicons were combined into a weak pool, and a moderate/strong pool based on gel band score and TapeStation analysis. The weak pool contained a moderate amount of adapter dimer that was successfully removed using an Ampure size-selection clean-up ([Agencourt AMPure beads, Beckman Coulter, Inc., Li, Hofreiter, Straube, Corrigan, & Naylor, 2013](#_ENREF_5)) prior to sequencing. TapeStation analysis was undertaken on each cleaned pool to determine concentration and molarity. A portion of each of the cleaned pools was combined to make a final pool, with the volume added based on the number of amplicons within each pool, after equalising molarity. The completed MiSeq library was sent to the Australian Genome Research Facility (AGRF) for 300bp paired-end sequencing on the Illumina MiSeq platform with 10% PhiX (a viral genome) added to create additional complexity within the library for more efficient base detection.

*Materials and Methods: Pre-genotyping bioinformatics*

Several bioinformatics steps were undertaken to prepare data for genotyping (figure S2). MiSeq paired-end data was received from the AGRF de-multiplexed by sample index, with two separate files per sample (read 1; R1 and read 2; R2). The overall quality of the data was determined by performing FastQC ver. 0.11.2 ([Andrews, 2010](#_ENREF_1)) analyses on all R1 and all R2 files (concatenated across individuals). Paired-end reads for each individual were assembled using PEAR ver. 0.9.5 ([Zhang, Kobert, Flouri, & Stamatakis, 2014](#_ENREF_10)) with a 50bp minimum overlap of reads and a Q-value cutoff of 20. Improvement in the quality of the reads (henceforth, assembled paired-end reads are referred to as ‘reads’) was assessed by FastQC analysis of 15 randomly selected samples. Target region sequence lengths were investigated for a subset of samples (n=15) in Geneious ver. 8.1.7 ([Kearse, et al., 2012](#_ENREF_4)) using jMHC ver. 1.6.1624 ([Stuglik, Radwan, & Babik, 2011](#_ENREF_8)) output, in which all identical reads are collapsed for each individual for a given locus-specific primer pair. Pseudogenes were identified by the presence of premature stop codons. Only lengths that were not pseudogenes or a subset of a longer length were included in downstream analyses. At the time of analysis, the programs used for genotyping (AmpliCHECK and AmpliSAS) required reads to be indexed and concatenated into a single file. Therefore, a python script (<https://github.com/kellyp2738/genotyping_methods_hacking_etal>) was written to re-index the reads, which were then concatenated into a single compressed file. From this file, target lengths were extracted using Geneious to create smaller, more manageable files suitable for AmpliCHECK and AmpliSAS analysis.

*Results: Illumina MiSeq and pre-genotyping bioinformatics*

The complete MiSeq run produced 6.5 million paired-end reads, with 3.1 million of those assigned to the amplicons used in this study (one α1 primer pair and one α2 primer pair; table 2). Initial FastQC analysis of raw, un-paired reads indicated high per-base and per-sequence Phred quality scores and negligible per-base N content. Adapter dimers contributed to a small percentage of the reads produced. Post paired-end read assembly FastQC analysis of a subset of samples revealed that the per-base and per-sequence Phred quality scores were slightly improved (per-sequence quality score 39-40) and contaminating adapter had been removed. On average, there were 17,174 reads per sample (including both α1 and α2 amplicons) and 99.7% of raw reads were paired. Investigation of variant lengths revealed that the α1 primers amplified a 206bp and 225bp fragment. All variants of 225bp in length contained premature stop codons and were assumed to be a pseudogene and consequently excluded from further analysis. Variants of 217bp and 214bp in length were produced for α2, with the size difference due to a 3bp insertion or deletion, with both size variants translating. All target lengths were present in all populations and the 214bp and 217bp α2 variants were analysed together.

Tables and figures

**Table S1.** List of inner and outer primer sequences and indexes used to amplify MHC class I in *C. decresii* and identify individuals.

|  |  |  |
| --- | --- | --- |
| Primer/index name | Description | Sequence (5’ – 3’) |
| MHC1\_A1\_G234\_F | MHC class I α1 forward primer. Part of ‘inner’ primer. | GGCTCCTCCTCGCACTCCCTG |
| MHC1\_A1\_G234\_R | MHC class I α1 reverse primer. Part of ‘inner’ primer. | ACATTCTGCAGGCTCACTCTGAAC |
| MHC1\_A2\_G234\_F | MHC class I α2 forward primer. Part of ‘inner’ primer. | TGCAGTTGATGTACGGCTGTGAGC |
| MHC1\_A2\_G12\_R | MHC class I α2 reverse primer. Part of ‘inner’ primer. | CCTCAGTAGGCTCTCCCTCCCG |
| MRT forward tag | Link inner and out primers  | ACGACGTTGTAAAA |
| MRT reverse tag | Link inner and out primers  | CATTAAGTTCCCATTA |
| P5 | Part of ‘outer’ primer. Forward illumina-specific adapter. | AATGATACGGCGACCACCGAGATCTACAC |
| P7 | Part of ‘outer’ primer. Reverse illumina-specific adapter. | CAAGCAGAAGACGGCATACGAGAT |
| PE Read 1 Sequencing Primer | Part of ‘outer’ primer. Forward illumina-specific adapter. | ACACTCTTTCCCTACACGACGCTCTTCCGATCT |
| Multiplexing Read 2 Sequencing Primer | Part of ‘outer’ primer. Reverse illumina-specific adapter. | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| Index\_F\_1 | i5 forward barcode | TCTCTGTG |
| Index\_F\_2 | i5 forward barcode | ACTCACTG |
| Index\_F\_3 | i5 forward barcode | TCTACTCG |
| Index\_F\_4 | i5 forward barcode | TAGTAGCG |
| Index\_F\_5 | i5 forward barcode | AGACGACG |
| Index\_F\_6 | i5 forward barcode | ACTCGTAG |
| Index\_F\_8 | i5 forward barcode | TGACGCAG |
| Index\_F\_10 | i5 forward barcode | ACGCTATC |
| Index\_F\_11 | i5 forward barcode | AGTGCTGC |
| Index\_F\_12 | i5 forward barcode | TACTACGC |
| Index\_F\_14 | i5 forward barcode | TCAGCTAC |
| Index\_F\_15 | i5 forward barcode | AGAGCGAC |
| Index\_F\_16 | i5 forward barcode | ATGCTCAC |
| Index\_F\_17 | i5 forward barcode | TAGCACAC |
| Index\_F\_18 | i5 forward barcode | TGTACGTG |
| Index\_F\_19 | i5 forward barcode | AGAGTATG |
| Index\_F\_20 | i5 forward barcode | ATGACTCG |
| Index\_F\_21 | i5 forward barcode | AGATAGCG |
| Index\_F\_22 | i5 forward barcode | TGTAGACG |
| Index\_F\_23 | i5 forward barcode | TGCAGTAG |
| Index\_F\_24 | i5 forward barcode | AGCTGATG |
| Index\_F\_25 | i5 forward barcode | ATAGAGAG |
| Index\_F\_26 | i5 forward barcode | TGTCACGC |
| Index\_F\_27 | i5 forward barcode | ATACTGCG |
| Index\_F\_28 | i5 forward barcode | ACTGTGTC |
| Index\_F\_29 | i5 forward barcode | TACACAGC |
| Index\_F\_30 | i5 forward barcode | TCGCTACG |
| Index\_F\_31 | i5 forward barcode | ACGTACTC |
| Index\_F\_32 | i5 forward barcode | TCGATGAC |
| Index\_F\_33 | i5 forward barcode | TGCGATGC |
| Index\_F\_34 | i5 forward barcode | AGCATCAC |
| Index\_F\_35 | i5 forward barcode | TATCGATG |
| Index\_F\_36 | i5 forward barcode | TATCAGAG |
| Index\_F\_37 | i5 forward barcode | AGTCTAGC |
| Index\_F\_38 | i5 forward barcode | TATATGCG |
| Index\_F\_39 | i5 forward barcode | TACATGTC |
| Index\_F\_40 | i5 forward barcode | ATACGTAC |
| Index\_R\_1 | i7 reverse barcode | ATCGTCTG |
| Index\_R\_2 | i7 reverse barcode | TGATCTAG |
| Index\_R\_3 | i7 reverse barcode | ATGCATGC |
| Index\_R\_4 | i7 reverse barcode | AGATGCAC |
| Index\_R\_5 | i7 reverse barcode | ATGCGATG |
| Index\_R\_6 | i7 reverse barcode | ACGCAGAG |
| Index\_R\_7 | i7 reverse barcode | ATGTGAGC |
| Index\_R\_8 | i7 reverse barcode | TGCTCGCG |
| Index\_R\_9 | i7 reverse barcode | ATGACGTC |
| Index\_R\_10 | i7 reverse barcode | AGTAGTAC |
| Index\_R\_11 | i7 reverse barcode | TCTGACTC |
| Index\_R\_12 | i7 reverse barcode | AGAGTCGC |
| Index\_R\_13 | i7 reverse barcode | TCAGCACG |
| Index\_R\_14 | i7 reverse barcode | TGCGTACG |
| Index\_R\_15 | i7 reverse barcode | TCATGTCG |
| Index\_R\_16 | i7 reverse barcode | TGACGATC |
| Index\_R\_17 | i7 reverse barcode | TCACAGCG |

**

**Figure S1.** Steps taken during the two-step PCR Illumina MiSeq amplicon library preparation for *Ctenophorus decresii* MHC class I. \*Primer concept as per ([Pearson, et al., 2016](#_ENREF_6)); ^Agilent 2200 TapeStation © (Agilent Technologies); +Agencourt AMPure © beads (Beckman Coulter, Inc.).

**Figure S2.** Pre-genotyping bioinformatics steps for *Ctenophorus decresii* class I MHC data. \*FastQC ver. 0.11.2 ([Andrews, 2010](#_ENREF_1)), ^PEAR ver. 0.9.5 ([Zhang, et al., 2014](#_ENREF_10)), €jMHC ver. 1.6.1624 ([Stuglik, et al., 2011](#_ENREF_8)), £Geneious ver. 8.1.7 ([Kearse, et al., 2012](#_ENREF_4)), ʎAmpliSAT ([Sebastian, et al., 2016](#_ENREF_7)).





**Figure S3.** Steps taken during *Ctenophorus decresii* MHC class I genotyping. \*AmpliSAT ([Sebastian, et al., 2016](#_ENREF_7)). ^LAEA: low amplification efficiency alleles.



**Figure S4.** Bayesian trees including all α1 (A) and α2 (B) variants prior to post-genotyping filtering. Low amplification efficiency alleles are indicated with red triangles, purple triangles represent putative artefact sequences with accumulated errors (chimera and mismatch) and pseudogenes are indicated with blue triangles.

References

Andrews, S. (2010). FastQC a Quality Control Tool for High Throughput Sequence Data [Online]. Retrieved.

Hacking, J. D., Bertozzi, T., Moussalli, A., Bradford, T. M., & Gardner, M. G. (2017). Characterization of major histocompatibility complex class I transcripts in an Australian dragon lizard. *Manuscript submitted for publication*

Hayden, M., Nguyen, T., Waterman, A., & Chalmers, K. (2008). Multiplex-Ready PCR: A new method for multiplexed SSR and SNP genotyping. *BMC Genomics, 9*, pp. 1-12.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., . . . Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28*(12), pp. 1647-1649. Retrieved from <http://www.geneious.com>

Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. (2013). Capturing protein-coding genes across highly divergent species. *BioTechniques, 54*, pp. 321-326.

Pearson, S., Bradford, T., Ansari, T., Bull, M., & Gardner, M. (2016). MHC genotyping from next-generation sequencing: detailed methodology for the gidgee skink, *Egernia stokesii*. *Transactions of the Royal Society of South Australia, 140*(2), pp. 244-262.

Sebastian, A., Herdegen, M., Migalska, M., & Radwan, J. (2016). AMPLISAS: a web server for multilocus genotyping using next-generation amplicon sequencing data. *Molecular Ecology, 16*(2), pp. 498–510.

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

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research, 40*(15), pp. e115-e115. doi:10.1093/nar/gks596 Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3424584/>

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics 30*(5), pp. 614-620.