Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird

Running title: Genotyping of extreme MHC variation.

Aleksandra Biedrzycka[1], Alvaro Sebastian[2], Magdalena Migalska[2], Helena Westerdahl[3], Jacek Radwan[2]

[1]Institute of Nature Conservation, Polish Academy of Sciences, Al. Mickiewicza 33,
31-120 Kraków, Poland
[2]Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, ul.
Umultowska 89, 61-614 Poznań, Poland
[3]Department of Biology, Lund University, Ecology Building, Sölvegatan 37, 223 62 Lund,
Sweden

Key words: passerine MHC, copy number variation (CNV), next generation sequencing, genotyping, bioinformatics

Corresponding author:
Jacek Radwan
Evolutionary Biology Group
Faculty of Biology
Adam Mickiewicz University
ul. Umultowska 89, 61-614 Poznań, Poland
Email: jradwan@amu.edu.pl
Tel.: 0048 61 829 5853
Fax.: 0048 61 829 5636

Abstract

Characterisation of highly duplicated genes, such as genes of the major histocompatibility complex (MHC), where multiple loci often co-amplify, has until recently been hindered by insufficient read depths per amplicon. Here we used ultra-deep Illumina sequencing to resolve genotypes at exon 3 of MHC class I genes in the sedge warbler (*Acrocephalus schoenobaenus).* We sequenced 24 individuals in two replicates and used this data, as well as a simulated dataset, to test the effect of amplicon coverage (range: 500-20 000 reads per amplicon) on the repeatability of genotyping using four different genotyping approaches. A third replicate employed unique barcoding to assess the extent of tag jumping, *i.e.* swapping of individual tag identifiers, which may confound genotyping. The reliability of MHC genotyping increased with coverage and approached or exceeded 90% within-method repeatability of allele calling at coverages of >5 000 reads per amplicon. We found generally high agreement between genotyping methods, especially at high coverages. High reliability of the tested genotyping approaches was further supported by our analysis of the simulated dataset, though the genotyping approach relying primarily on replication of variants in independent amplicons proved sensitive to repeatable errors. According to the most repeatable genotyping method, the number of co-amplifying variants per individual ranged from 19 to 42. Tag jumping was detectable, but at such low frequencies that it did not affect the reliability of genotyping. We thus demonstrate that gene families with many co-amplifying genes can be reliably genotyped using HTS, provided that there is sufficient per amplicon coverage.

Introduction

The Major Histocompatibility Complex (MHC) genes encode proteins of key importance in adaptive immunity (Janeway *et al.* 2004) and the MHC region holds the most polymorphic genes known in vertebrates (reviewed in Bernatchez & Landry 2003; Sommer 2005). This polymorphism is thought to be maintained by balancing selection caused by pathogens, and indeed correlations between MHC genotypes and resistance/susceptibility to infection has been demonstrated both in humans and in many studies of animal in natural populations (reviewed in Blackwell *et al.* 2009; Piertney & Oliver 2006; Spurgin & Richardson 2010). MHC polymorphism can further be promoted by MHC-based mate choice (Ejsmond *et al.* 2014; Strandh *et al.* 2012; Winternitz *et al.* 2013).

The number of MHC genes per individual differs greatly between (reviewed in Kelley *et al*. 2005), and even within, species (*e.g.* fish, Wegner *et al.* 2003; amphibians, Nadachowska-Brzyska *et al.* 2012; reptiles, Radwan *et al.* 2014; birds, Bonneaud *et al.* 2004; mammals, Kloch *et al.* 2010b). In birds, the variation ranges from just a few genes in the chicken *Gallus gallus* and other birds of the Galliformes, to a large number of highly polymorphic genes among passerines of the order Passeriformes (Karlsson & Westerdahl 2013; Kaufman *et al.* 1999; Radwan *et al.* 2012; Sepil *et al.* 2012; Westerdahl *et al.* 2004a,b). Due to gene conversion that occurs both within and between loci, passerine MHC alleles exhibit high similarity and it is usually impossible to determine the locus to which a given allele belongs (Hess & Edwards 2002; Westerdahl *et al.* 1999). The presence of multiple MHC loci coupled with extensive copy number variation (CNV; *e.g. Passer domesticus* , Bonneaud *et al.* 2004; *Parus major*, Sepil *et al.* 2012; *Ficedula albicollis* Radwan *et al.* 2012; *Acrocephalus* spp., O’Connor *et al*. 2016) makes passerines an ideal system for studying the evolution, expansion and contraction of MHC genes. However, this potential has been limited by the technical difficulties of genotyping alleles from a large number of loci (Babik 2010).

During recent years high-throughput sequencing (HTS) technologies have become a useful tool for assessing polymorphism in complex MHC systems (Babik *et al.* 2009; reviewed in Lighten *et al.* 2014a). Nevertheless, despite significant recent advances in genotyping methods, the characterization of MHC variation in non-model species still remains a challenging task due to the extremely high numbers of co-amplifying genes. However, new platforms offering ultra-high coverage, such as Illumina and Ion Torrent, which have recently been used in genotyping multi-locus MHC systems (Herdegen *et al.* 2014; Lighten *et al.* 2014b), offer the potential to overcome this limitation.

HTS technologies are characterised by a relatively high frequency of sequencing errors (Huse *et al.* 2007; Liu *et al.* 2012; McElroy *et al.* 2012), and the ability to discriminate between true alleles and artefacts created by HTS is reduced when the number of co-amplifying loci increase. This difficulty can be aggravated by CNV in MHC genes observed in many species (reviewed in Babik 2010; Lighten *et al.* 2014a; Sommer *et al.* 2013), and by the possibility of tag jumping, *i.e.* appearance during ligation or pooled PCRs of false tag combinations of individual barcodes that can cause incorrect assignment of sequences to samples (*e.g.* Schnell *et al*. 2015).

One method that has been used for discriminating true alleles from artefacts is to compare obtained frequencies of alleles within amplicons to expectations from genetic models (Ferrandiz-Rovira *et al.* 2015; CNV method in Lighten *et al.* 2014b). However, this method becomes infeasible when the number of loci is high and the copy number varies from a few to more than ten, as is the case in some passerine systems (*e.g.* Sepil *et al.* 2012; Westerdahl *et al.* 2004a; Zagalska-Neubauer *et al.* 2010). Other methods rely upon the assumption that true alleles will occur at considerably higher frequencies than artefacts (*e.g.* Babik *et al.* 2009; Kloch *et al.* 2010a; DOC method in Lighten *et al.* 2014b; Radwan *et al.* 2012; Sommer *et al.* 2013). However, with a large number of co-amplifying loci, the expected allele frequencies may decline to less than those typically reached by artefacts. For example, artefacts were reported to reach 1.5% per amplicon in flycatchers *Ficedula albicollis* (Radwan *et al.* 2012) and rose finches *Carpodacus erythrinus* (Promerová *et al.* 2012), and 2.5% in great tits *Parus major* (Sepil *et al.* 2012). Thus, for systems with an extremely high number of co-amplifying MHC genes, the effectiveness of discriminating artefacts from true alleles based upon frequency is questionable. The most recent genotyping approaches attempt to increase the frequencies of true variants versus artefacts by reassigning artefacts to their putative parental sequences and thereby increasing the read depths of true variants (Lighten *et al.* 2014b; Pavey *et al.* 2013; Sebastian *et al.* 2016; Stutz & Bolnick 2014). However, the performance of such approaches in genotyping families of many co-amplifying MHC loci has yet to be tested.

In this study we tested the potential of ultra-deep amplicon sequencing for characterising MHC diversity, using the sedge warbler *Acrocephalus schoenobaenus* as a test subject. MHC diversity has yet to be characterised for this species, but sedge warblers are closely related to three species with well-characterised, highly complex MHC systems (*Acrocephalus arundinaceus*, Westerdahl *et al*. 2004a; *Phylloscopus trochilus* and *P. collybita,* O'Connor *et al.* 2016). Using a set of replicated sedge warbler samples, we quantitatively evaluated four different approaches for removing artefacts from HTS amplicon data and genotyping: (1) Radwan *et al.* (2012); (2) Sommer *et al.* (2013); (3) Lighten *et al.* (2014b); and (4) Sebastian *et al.* (2016). We selected these four genotyping methods because, while they all share the assumption that artefacts will be rare compared to the true variants, each one is based upon different classifying, clustering or filtering principles, as described below. A more detailed description of each algorithm is given in the Methods section.

The first genotyping method (hereafter referred to as the 'allele validation threshold method', AVT) was developed by Radwan *et al.* (2012) and is based upon a combination of principles introduced by Kloch *et al.* (2010a) and Zagalska-Neubauer *et al.* (2010). This approach relies upon comparing rare variants to more common variants within the same amplicon to determine whether the rare variant can be explained as a sequencing artefact (*e.g.* for Illumina data, a variant with 1-2 bp mismatches to another more common variant in a given amplicon, Herdegen *et al.* 2014) or a PCR chimera of more common variants. Typically, below a certain threshold (*e.g.* <1.5% per amplicon in Radwan *et al.* 2012) the vast majority of variants can be explained as sequencing artefacts. By removing all sequences below this threshold, possible contaminations, which are typically very rare (Lighten *et al.* 2014b), are likely to be removed.

The second genotyping method was proposed by Sommer *et al.* (2013), and is based on using replicated samples to genotype (hereafter referred to as the 'replicated amplicon method', RA). In summary, for likely artefacts (*i.e.* variants which are 1-2 bp different to a more common variant within the same amplicon) this genotyping method checks if the variant is also present in a replicate amplicon. If not, it is defined as an artefact; if yes, it can still be classified as putative artefact if its frequency is lower than the frequency of any earlier-defined putative artefact. Separate decision trees deal with chimeras and sequences more than 2 bp different that do not occur in both replicates.

The third genotyping method evaluated is the 'degree of change' (DOC) method of Lighten *et al.* (2014b). In this approach, rare sequences are clustered with similar (1-3 bp different), more common variants within amplicons and their reads are added to the reads of the higher frequency variant. The threshold dividing true sequences form artefacts is then set within each amplicon separately by ordering variants by their frequency and finding where the drop in cumulative frequency of consecutive variants is the highest. Variants that fall below the threshold are also manually assessed to identify poorly amplifying true alleles.

The fourth genotyping method (hereafter referred to as the 'adjustable clustering method', AC) is based upon a clustering algorithm that uses the error rate of a particular HTS technique and the expected length/s of the amplified sequences to set classifying parameters (Sebastian *et al.* 2016). This genotyping method is based upon the idea of the stepwise threshold clustering (STC; Stutz & Bolnick 2014), but it is optimized to accommodate efficient analysis of larger datasets. In short, it starts clustering using the most common variant in the amplicon as a core of the first cluster, adds similar variants (based on selected error rates) and then starts the next cluster with the most common, and yet un-clustered variant. The algorithm allows flexibility in setting criteria for similarity and takes into account the relative abundance of a variant within a cluster, as originally proposed by Stutz and Bolnick (2014). Sebastian *et al.* (2016) have benchmarked their algorithm using HTS data from a well characterized HLA system for which true sequences were established independently, which has not been done for the other three approaches.

We investigated the repeatability of genotypes retrieved by these methods at different coverages (*i.e.* numbers of reads per amplicon) obtained by subsampling. We also checked whether tag jumping affected genotyping results. In addition to investigating the repeatability of genotypes within methods, we also assessed the consistency of these genotypes between the four methods. High agreement between the different protocols would give us extra confidence in the utility of ultra-deep amplicon sequencing characterising the highly complex, co-amplifying MHC genes. Furthermore, because researchers have historically used different protocols, checking consistency between them is important as it indicates whether genotypes and conclusions are comparable between studies. Finally, we used a simulated dataset to compare the ability of different genotyping methods to retrieve known genotypes.

Methods

Sample collection and DNA extraction

A total of 24 samples used in this study were a random subset of samples provided to us by Drs. T. Zając, W. Solarz and W. Bielański from the Institute of Nature Conservation PAS, Kraków. The samples were collected from a sedge warbler population in the Nida marshes (southern Poland; 20°28'–20°32' E, 50°33'–50°35' N) during the 2009 breeding season. The birds were mist-netted and blood samples were obtained from the brachial veins of adult birds. The blood samples were preserved in 95% ethanol. Genomic DNA was extracted with the Nucleospin Tissue Kit (Macherey and Nagel, Germany).

Illumina sequencing

Illumina sequencing was performed on amplicons obtained with primers on the 24 genomic DNA samples. Specific primers were designed based on cDNA and gDNA sequences (Biedrzycka et al. unpublished) obtained from warblers to amplify 235-241bp fragments of exon 3, encoding the α2 domain of the antigen-binding groove of the MHC class I molecule: forward (5'- GAGYGGGGTCTCCACAC-3'), reverse (5'-TGCGMTCCAGYTCCTTCTGCCC-3'). The primers were optimised to fit the complete range of alleles previously identified for exon 3 of MHC class I genes in sedge warblers (Biedrzycka et al. unpublished; not that in line with other publications on MHC diversity, we use the term 'allele' to describe variants amplified with the same primers, although we cannot assign alleles to particular loci). Illumina sequencing primers and unique 6bp tags were added to the 5' ends of both the forward and reverse primers for the first PCR. The tags were custom made and designed to differ in at least 2 bp positions. Amplification was performed with HotStar Master Mix (QIAGEN), and the reaction was run for 27 cycles at 95^{o}C for 30 s, 66^{o}C for 30 s, 72^{o}C for 1 min 30 s. The concentration of the PCR products was estimated by eye from agarose gels stained with Gel Red, and PCR products were pooled in approximately equimolar quantities. The resulting pools were purified using the MinElute PCR Purification

Kit (QIAGEN). Pooled amplicons were then amplified in the second PCR in order to add Illumina P5/P7 adaptor sequences (Syed et al. 2009). The pool was diluted by between 10x and 100x, depending on the density of the electrophoresis band, and PCR utilising primers carrying P5/P7 adaptors was performed for 12 cycles (according to Illumina MiSeq protocol). Second-stage products were again run on the gel and purified with the MinElute PCR Purification Kit (QIAGEN). Paired-end sequencing runs were performed on Illumina Miseq with the Miseq Reagent Kit v2 for 300 cycles (Illumina, Inc., San Diego, CA, USA).

Two replicates of the 24 samples, amplified in two independent PCRs and run in two separate sequencing experiments (henceforth ‘Replicates 1 and 2’) were amplified using three forward and eight unique reverse tags, such that each individual could be recognised by a unique combination. These two replicates were used to calculate repeatability of genotyping within methods and agreement between methods.

The same 24 samples were amplified a third time with 24 forward and 24 reverse tags, such that none of the tags was shared among samples. The aim of this third replication (henceforth Replicate 3) was to assess the magnitude of tag jumping between samples during the second PCR, where pooled amplicons are amplified in one PCR reaction, *i.e.* the potential forming of chimeras between sequences from different samples that would result in an exchange of 3’ tags.

Pre-processing of Illumina data

Paired-end reads within each run were merged with AmpliMERGE, a tool based on FLASH (Magoc & Salzberg 2011). AmpliMERGE optimized minimum and maximum overlapping length parameters (min_overlap=7, max_overlap=15) taking into account the lengths of: reads, amplified sequences, primers and barcodes. AmpliMERGE and other amplicon sequencing analysis tools referred in this section are available at: http://evobiolab.biol.amu.edu.pl/amplisat/.

To preliminarily explore the dataset, the three runs were analyzed with AmpliCHECK tool (Sebastian *et al.* 2016) using default Illumina parameters, minimum amplicon depth of 5 000 reads and minimum per amplicon frequency of 0.1%. AmpliCHECK revealed three major allele lengths among the most abundant variants: 235, 238 and 241 bp. A small number of high frequency variants were also found with lengths of 230 and 240 bp (2 variants of each), which were present in eight individuals, with one variant per individual. Aligning these sequences to the other variants showed deletions of one and 11 nucleotides respectively. We thus treated variant lengths of 230 and 240 bp as putative pseudogenes and excluded them from genotyping.

Assessment of genotyping repeatability within and between methods

We used the replicates 1 and 2 to compare genotyping results from four different published genotyping methodologies: (1) the allele validation threshold (AVT) method (Radwan *et al.* 2012, as applied to ultra-deep sequencing by Herdegen *et al.* 2014) (2) the replicated amplicons (RA) method (Sommer *et al.* 2013), (3) the degree of change (DOC) method (Lighten *et al.* 2014b), and (4) the adjustable clustering (AC) method using AmpliSAS tool (Sebastian *et al.* 2016), as described below. While our implementation of these methods captured their main rationale, we have adjusted some aspects of the post-processing recommended in their original descriptions, as detailed and justified below.

To estimate how the repeatability of genotyping is affected by coverage, we randomly sub-sampled amplicons from Replicates 1 and 2 to obtain subsets of 500, 1 000, 2 000, 5 000, 10 000, 15 000 and 20 000 reads. For the AVT, DOC and AC methods, we calculated the repeatability of genotyping as the percentage of alleles present in both replicates (1 and 2) with respect to the total number of alleles detected across both replicates. However, for the RA method, replicates 1 and 2 were both needed to calculate genotypes, so we do not report repeatability for this method.

To calculate agreement between methods, we first inferred consensus genotypes for AVT, DOC and AC using only the alleles that were recovered in both replicates (RA uses both replicates to yield one genotype). Based on these consensus genotypes, we calculated the agreement of genotypes between all tested methods as the percentage of alleles present in consensus genotypes obtained by two different methods.

The Allele Validation Threshold method (AVT) The AVT method establishes the frequency thresholds delimiting artefacts from true alleles based on examination of frequencies of putative artefacts (indels, 1-2 bp substitutions and chimeras). To determine these thresholds we used AmpliCHECK software (Sebastian *et al.* 2016) and then followed the modification of the protocol described in Herdegen *et al.* (2014), implemented within the AmpliLEGACY tool. To identify thresholds for facilitating the identification of true alleles, variants were ranked by their maximum per amplicon frequency (MPAF), by which each variant was represented across the entire dataset. Singletons were first removed. Then, starting from the MPAF of 0.1 (a value lower than in other studies because we expected much larger number of variants per amplicon), we checked whether the variant within each amplicon could be explained as a 1-2 bp substitution in a more common variant within the same amplicon or a chimera (a variant which can be explained by joining the forward and reverse sections of two more frequent variants, with ≥10 bp of length and ≥3 bp of difference attributable to each ‘parental’ sequence; see AmpliCHECK documentation for details). Such variants were considered putative artefacts. Appropriate annotation is performed automatically by AmpliCHECK. At the highest coverage (20 000 reads), 99% of variants within the 0.1% -

0.18% MPAF interval were classified as putative artefacts. Consequently, for each amplicon in the entire dataset, all variants below 0.18% per amplicon frequency (PAF) were considered artefacts. Similarly, all variants below 0.60, 0.30, 0.15, 0.16, 0.16, 0.16 and 0.18% per amplicon frequency (PAF) were considered artefacts for amplicon depth subsets 500, 1 000, 2 000, 5 000, 10 000, 15 000 and 20 000 respectively.

Unlike in earlier studies (*e.g.* Herdegen *et al.* 2014; Promerová *et al.* 2012; Radwan *et al.* 2012) we could not find a usable upper MPAF threshold, above which none of the variants would be classified as putative artefacts. This is because some high frequency variants, present in many individuals, differed by 1-2 bp from some more common variant within the same amplicon. Such variants constituted 18-22% of all variants left above the 0.18% PAF threshold, and their PAFs were often very high (up to 8.14 and 7.41% PAF in the first and second replicate at coverage = 20 000). Thus, our 'grey zone', in which alleles are called based on comparison to more common alleles within the same amplicon (Radwan *et al.* 2012) comprised practically all variants above the lower threshold. However, because high frequency variants occurring with slightly more common 1-2 bp different variants are likely to represent true alleles, we only considered them as artefacts if their frequency was low enough to be explained as sequencing error. Based on estimates showing that maximum substitution error frequency for Illumina may reach 6% for some specific motifs (McElroy *et al.* 2012), we conservatively considered variants as artefacts if their frequency was <10% of a more common variant within the same amplicon, from which they differed by 1-2 substitutions.

The Replicated Amplicon method (RA) The RA method, in addition to identifying putative artefacts, relies on agreement between two independent amplicons derived from the same sample. It was implemented using the automated online tool AmpliLEGACY (http://evobiolab.biol.amu.edu.pl/amplisat/) following, as closely as possible, the original amplicon-by-amplicon three-step workflow (Sommer *et al.* 2013). In the first step, singletons are discarded as artefacts, the most abundant variant is marked as putative allele, and the rest of the variants are classified into three groups. The first group (1-2 diff.) comprises of variants differing by 1 or 2 bp from another variant of higher frequency. The second group (>2 diff.) contains variants differing by more than 2 bp from other variants. The third group comprises variants that could be explained as the combination of another two higher frequency variants (*i.e.* likely PCR chimeras). Sequence comparisons are performed by pairwise global alignments as explained in Sebastian *et al.* (2016). To avoid classification of 1-2 bp differences as chimeras (*e.g.* in the 3' or 5' extreme of the sequence), only variants >2 bp different from parental sequence were classified as chimeras. In the second step, the presence of a variant in an amplicon replicate is checked. The following variants are treated as artefacts at this step: 1-2 bp diff. variants which did not occur in a replicate, >2 bp diff. variants which do not occur in a replicate or in other individuals, and chimeras which occur with parental sequences in at least one replicate. Remaining sequences are classified as putative alleles, putative artefacts or unclassified variants according to their frequencies (if it is lower than any putative artefact from step 2 or not) and their presence in other individuals. For more details see (Sommer *et al.* 2013).

RA workflow may classify the same variant differently in different replicates of the same individual (*e.g.* can be below artefact frequency threshold in one replicate, but not in the other, see File S1, Supporting information, for examples). While the authors of the RA method suggest manual curation of cases in which the two genotypes disagreed (Mazzoni, personal communication), we think this may involve a degree of arbitrariness unless strict decision rules are applied. Therefore, whenever intermediate classifications differed between replicates, we used a consensus (*i.e.* alleles classified as putative alleles in both replicates) to obtain final genotypes for comparisons with other methods. Our approach is conservative (*i.e.*

decreases the chance of reporting artefacts as true alleles), and can be easily replicated between labs. Nevertheless, in our dataset, manual curation yielded similar final genotypes, because the majority of discordant classifications were related to likely sequencing errors (*i.e.* 1-2 bp substitutions between the sequence and that of a higher frequency variant present within the same amplicon).

Degree Of Change method (DOC) The DOC method described in Lighten *et al.* (2014b) estimates the number of true alleles present in each amplicon independently based on the degree of change (DOC, the second derivative) in the cumulative sequencing depth between depth-ranked variants in an amplicon. The variant with the highest DOC value is considered the last true allele. We used the automatic DOC method available in AmpliLEGACY, following as closely as possible the original workflow of Lighten et al. (2014a). The DOC method first entails an error correction (*i.e.* clustering) step, where sequences that differ by 1-3 bp from a higher frequency variant (parental) are considered putative errors and their reads are added to the parent sequence. Because Lighten *et al.* (2014b) state that putative errors were “generally <2% depth”, we adapted this threshold, and pooled variants only if they had frequencies <2% compared to putative parental sequence. Following Lighten *et al.* (2014b), AmpliLEGACY performs error correction using pairwise global alignments of variants within each amplicon.

After error correction, the rate of change (first derivative) in cumulative sequencing depth is calculated for each variant resulting from the previous step, and then the percentage of DOC among variants. DOC value is calculated taking as many variants as the maximum number of expected alleles for a single individual set by researchers. Based on preliminary examination of the numbers of variants with AmpliCHECK, we set this maximum at 60.

Lighten *et al.* (2014b) recommended that amplicons which do not show “an obvious reduction in sequencing depths between putative alleles and artefacts, or an inflection point in a linear plot of cumulative sequencing depths at the point of greatest DOC” are removed. However, in our assessment, our data did not show any obvious distinctions between amplicons in this respect (compare *e.g.* Fig. S1, Supporting information), likely because our system included so many alleles as to which make identification of an inflection point problematic (Lighten *et al.* (2014b). We therefore included all amplicons in our analyses.

Lighten *et al.* (2014b) further recommended that poorly amplified putative alleles should be distinguished from contaminants on the basis that the latter should be identical to high copy number putative alleles in other amplicons. However, if all low frequency variants (which we defined as occurring below the greatest DOC, up to a maximum of that ranked 60th with respect to frequency) were used, it would lead to gross overestimation of the number of true variants, due to the presence of complex, but repeatable, errors (*e.g.* unfiltered artefacts >3 bp different from parental sequence) that would be classified as low amplifiers (see Supplementary material), so we omitted this step.

Adjustable Clustering with AmpliSAS (AC) AmpliSAS performs clustering of variants based on the parameters set by users depending on expected error rates (Sebastian *et al.* 2016). We used the AmpliSAS default parameters for Illumina data: a substitution error rate of 1%, and an indel error rate of 0.001%. Additionally we adjusted the ‘minimum dominant frequency’ clustering threshold to 10%, *i.e.* only sequences with frequency below this threshold were clustered with parental sequence (Sebastian *et al.* 2016), in order to retain true similar variants but eliminate high-frequency, motif-specific errors (McElroy *et al.* 2012), as explained above for the AVT method.

Following clustering, AmpliSAS allows the filtering of artefacts, *i.e.* removing low frequency clusters and clusters formed by chimeras of more common variants within an amplicon. We removed low frequency variants by setting the 'minimum amplicon frequency' filtering threshold to 0.4% PAF. This value is based upon the observation that when clusters within an amplicon were ordered by descending PAF, a distinct drop in frequency was observed at this value (Fig. S2, Supporting information). This drop was clearly visible at higher coverage subsets (above 5 000 reads), but, for consistency, a value of 0.4% PAF was used throughout the analyses. We also removed chimeras by enabling the appropriate AmpliSAS option. Chimeras represented a considerable proportion (on average 68.43%, maximum of 96%) of the low frequency variants which remained after clustering as, in contrast to sequencing errors, they cannot be incorporated to the parental sequences. However, they had very low PAFs (maximum 0.04%), so they would also have been removed by the minimum PAF criterion.

Variant amplification efficiencies and genotyping confidence

We estimated variant amplification efficiencies using R scripts provided by Sommer *et al*. (2013). Their method uses a maximum likelihood optimization approach to assign a single efficiency value to each allele, based on observed frequencies in a real dataset. The calculations were preformed independently for Replicate 1 and 2, based on coverages reported by the AC method for 20 000 reads subset. We chose the results of the AC method as it clusters errors to their parental sequences, rather than discarding them. As some sequence motifs were reported to be more prone to errors (McElroy *et al.* 2012), alleles bearing them might suffer more intense ‘coverage loss’, which could later obscure the genuine variability in amplification efficiency. We also adapted the calculations from Sommer et al. (2013) to estimate a required allele coverage that with 99.0% confidence would yield at least 10 reads of the lowest amplifying allele in an individual with the maximal number of alleles. For the details of the simulation see Sommer *et al*. (2013). The simulation was run 100 times, and the median is reported as a result. The provided scripts need a desired number of reads to be observed, rather than per amplicon frequency, therefore we cannot directly compare this result with minimum per amplicon frequency requirements used in earlier steps of our AC analysis. However, we calculated the minimal coverage that would yield at least 80 reads for each allele (0.4% PAF at coverage 20 000) with 99.0% confidence. In addition, we tested how the theoretical confidence of retrieving identical genotypes to those of AC (in Replicate 1, coverage 20 000) changed across different subsets of sequencing depth. We consider a genotype to be sufficiently covered if each allele reaches a minimal threshold of 0.4% PAF, that is 2 reads for 500 reads subset, 4 reads for 1 000 reads subset, etc. We simulated the number of reads corresponding to 90% of each coverage to account the loss of data due to removal of chimeras and unclustered errors with more than two substitutions (approximately 10% of reads).

Simulations

In order to better understand how well different methods retrieve known genotypes, we used a simulated dataset derived from consensus sequences of AC simulation (original genotypes, alleles and amplification efficiencies used as template in the simulation are shown in File S2, Supplementary information). The consensus genotypes of 24 individuals constituted ‘true genotypes’ in our simulations. We then simulated sequencing reads by (i) allowing uneven amplification efficiency of variants, (ii) simulating chimeras and (iii) errors (substitutions and indels). To simulate differences in amplification efficiency between variants, we used estimates obtained as described above using the pipeline of Sommer *et al*. (2013). Each variant was thus ascribed its specific amplification efficiency (calculated as a mean from estimates obtained in Replicates 1 and 2). To obtain a set of simulated reads for a given individual, its variants were drawn with a probability equal to its specific amplification efficiency. Chimeras were created by random mixture or two variants in 15% of the reads (a value estimated from our dataset, see Results, section *Evaluation of the extent of tag jumping in the second PCR*). Errors were then introduced at random nucleotide positions with a substitution probability of 0.1% and indel probability of 0.001% (Loman *et al.* 2012; Ross *et al.* 2013). Errors could also occur in primer and barcode sequences. Additionally, motif-specific errors were introduced. Specific Illumina v5 error-prone DNA motifs reported by McElroy *et al*. (2012) were located along the sequences and their specific errors inserted at the average frequencies reported by the authors (Table 3 in McElroy et al. 2012). Four sets of 50 000 simulated single end reads were generated with the tool AmpliSIM (http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplisim), each set being equivalent to one amplicon sequencing experiment. The first two sets of simulated reads were used to compare the results of AVT, RA and AC genotyping methods to the true genotypes, and the 3rd and 4th sets were treated as the replicate experiments of the 1st and 2nd for the RA genotyping method. For each amplicon we calculated false negative rate (FNR) and false discovery rate (FDR) in the following ways: FNR = FN/(TP+FN), FDR = FP/(TP+FP); where TP, number of true alleles in a known genotype; FP, number of false (artefactual) alleles called in a genotyping procedure; FN, number of false negatives, undetected alleles by a genotyping procedure.

Evaluation of the extent of tag jumping in the second PCR

Replicate 3 was used to evaluate the extent of tag jumping in the second PCR, in which Illumina adaptors were added to pooled amplicons (see Illumina Sequencing section above). Twenty two out of 24 samples run with unique sets of forward and reverse primers in

Replicate 3 were retrieved after de-multiplexing, of which three did not reach the 20 000 coverage required for genotyping. Replicate 3 was filtered using the same AmpliCLEAN parameters as used for processing replicates 1 and 2 (see above), and all the possible 462 tag jumping combinations (22 individuals x 21 not-self tags which were not originally present in any amplicon in this set of samples) were extracted and quantified. To assess if tag jumping during the second PCR affected genotyping results, we genotyped Replicate 3 using AC method as described above. We then compared the result with the consensus genotype of Replicates 1 and 2 obtained with the same method.

Results

MHC class I exon 3 sequences were genotyped successfully in 24 sedge warbler individuals using Illumina MiSeq. For the AC method (which proved the most replicable, see below) the distribution of number of alleles per individual varied between 19 and 42. In total we verified 403 alleles in 24 individuals, an impressively high MHC diversity in a relatively small dataset.

Numbers of variants called by different methods

The number of alleles per individual called across coverages of 500-20 000 reads per amplicon was similar for the three genotyping methods AVT, RA and AC (Fig. 1). The fourth genotyping method, DOC, called a similar average number of alleles to the other genotyping methods, but the range in the number of alleles was considerably greater, especially at low coverages (Fig. 1, File S1, Supporting information). This discrepancy resulted mostly from a few cases when DOC called a very low number of alleles compared to other methods and to most other DOC genotypes.

Repeatability of MHC genotyping within methods at different coverages

The repeatability of allele calling between two separate sequencing experiments (and independent PCR runs) within AVT, DOC and AC is shown in Figure 2. The repeatability increased steeply with coverage until 5 000 reads. Above this coverage, all three methods reached or exceeded a mean repeatability of 90% (AVT: 95%, DOC: 90%, AC: 98%). At 20 000 reads the repeatabilities within methods reached 97%, 91% and 99% for AVT, DOC and AC, respectively.

Agreement of MHC genotyping between methods

The agreements between consensus genotypes assigned by AVT, RA and AC were high and typically reached or exceeded 90% at a coverage of 2 000 reads (Fig. S3; because of high agreement between AVT and AC, only the latter is presented in Figure 3 for clarity). As can be seen in Figure 3, DOC agreed less well with other methods, especially at low coverages. At medium coverages (5 000 – 15 000 reads), the average agreement between DOC and other methods was 85-90%, and at the highest coverage (20 000 reads) DOC reached 88% of agreement with RA and 91% of agreement with AVT and AC genotypes (File S1, Supporting information). Above a coverage of 5 000 reads, 75% of the consensus genotypes were identical between DOC and AC (Fig. 3, File S1, Supporting information). The differences occurred when DOC assigned unusually low numbers of alleles in one or both replicates (*e.g.* amplicons 2 and 17; Fig. S1, File S1, Supporting information).

Variants' amplification efficiencies and genotyping confidence

Relative efficiencies of amplification calculated according to Sommer *et al.* (2013) ranged between 4.77 and 0.21 for Replicate 1, and between 5.03 and 0.24 for Replicate 2. The distribution of the amplification efficiencies is shown in Figure S4a (Supporting information).

The efficiencies for the same alleles were highly correlated between replicates (correlation coefficient: 0.94, Fig. S4b Supporting information).

The lowest amplification efficiency calculated was 0.2, and the highest number of alleles found in one individual was 42. According to the calculations adapted from Sommer *et al.* (2013) the coverage of 3 819 reads would be necessary to reach 99.0% genotyping confidence, with at least 10 reads of the lowest amplifying allele present in an individual possessing 42 alleles. To reach at least 80 reads per allele (0.4 PAF for 20 000 subset), with the same confidence level, minimal coverage of 20 902 reads would be needed.

Tests of confidence for retrieving genotypes identified with the AC method across different coverage showed a constant upward trend. With 500 reads the median probability of obtaining identical genotypes was only 68% (mean: 65%), but with 5 000 reads the median confidence in genotyping reached 99% (mean: 96%), and at 20 000 reads 21 out of 24 genotypes would be retrieved with a confidence of 99.9% or higher. For complete results see Table S2 (Supporting information).

Simulations

Genotyping of the simulated dataset showed that genotyping reliability increases with coverage for AVT, DOC and AC (Fig. 4). At 5 000 reads and above, all methods yielded almost no false negative results and all true alleles are recognized as such. Unlike for the real dataset, we observed premature highest DOCs yielding untypically low number of called alleles only below the coverage of 10 000 reads (File S3, Supporting information). At lower coverages, to varying extent, all methods falsely identify artefacts as alleles, but for AVT, DOC and AC the false discovery rates drop to zero at and above 5 000 reads. For RA, however, we observed high rates of false discoveries even above 5 000 reads, and the false discovery rate increased further with coverage (Fig. 4).

Evaluation of the extent of tag jumping in the second PCR

The search for all possible 462 tag jumping combinations (22 amplicons x 21 tag combinations which were not originally present in any amplicon in this set of samples) in Replicate 3 resulted in a total number of 239 436 reads, which constituted about 15% of 1 610 545 reads with correct tag combination. However, reads with wrong tag combinations were distributed among 95 917 unique variants, such that the per amplicon frequency for any variant with wrong tag combinations never exceeded 0.003%. Such low frequency variants would thus not affect genotyping in Replicates 1 and 2, in which non-unique tag jumping during the second PCR of amplicon pools would result in these variants being ascribed to the wrong individuals. Indeed, this was confirmed when we compared the consensus genotypes as determined by AC (Replicates 1 and 2) with those from a Replicate 3 (barcoded with a combination of unique tags). We obtained coverage > 20 000 reads from for 19 individuals in Replicate 3. The average agreement of their genotypes with the consensus of Replicates 1 and 2 was 98% (range 91-100%). The mean number of alleles for the 19 individuals in the three MiSeq runs was identical (35 alleles per individual). There was thus no indication that using non-unique combinations of forward or reverse tags resulted in additional sequences assigned to individuals due to tag jumping.

Discussion

Our results show that HTS provides a reliable tool for genotyping gene families with a large number of co-amplifying loci when the read depth is satisfactory (here, at least 5 000 reads per amplicon). Previous ultra-deep genotyping studies have examined systems with lower MHC complexity consisting of one to eight co-amplifying variants (Herdegen *et al.* 2014; Lighten *et al.* 2014b). In sedge warblers, the number of co-amplifying variants exceeded 40 in some individuals, making filtering true alleles from artefacts challenging. Despite this, we demonstrated that at or above 5 000 reads per amplicon, such complex, multi-locus genotypes can be retrieved in a replicable manner. The sufficiency of this coverage was independently confirmed using calculations of genotyping confidence according to Sommer *at al.* (2013), which for AC method, reached the median of 99.9% at the coverage of 5 000 reads. The accuracy of the genotyping is supported by both the high agreement between genotypes obtained from raw data by different genotyping methods, and by our analysis of a simulated dataset.

The four different genotyping methods are based on different principles: AVT (Radwan *et al.* 2012) sets a threshold based upon the observed frequencies of putative artefacts, RA (Sommer *et al.* 2013) relies upon the occurrence of the same variant in an amplicon replicate, DOC (Lighten *et al.* 2014b) relies upon a drop in sequencing depth between true alleles and artefacts, and AC (Sebastian *et al.* 2016) clusters variants in a fashion that is customized to platform-specific error profiles. High agreement across genotyping methods despite these differences adds confidence in the reliability of genotyping, which we see as the main benefit of the comparisons we have performed in the present paper. Additionally, our comparisons have revealed some strengths and weaknesses to the four genotyping methods tested.

The AVT method yielded very similar genotypes to AC, so we do not discuss it further in the between-protocol comparisons reported below. The genotypes obtained using RA and AC agreed well across coverages between 1 000-20 000 reads. However, even at high coverages agreement between some genotypes was below 90%, illustrating the difficulty in genotyping highly complex MHC genotypes. Agreement between the genotypes obtained using the DOC and AC methods was low at and below coverages of 5 000 reads, but above 5 000 these two methods reached the highest agreement, with very few disagreements resulting from outlier genotypes obtained using the DOC method, as discussed in detail below. Similar patterns of discrepancies were observed between the RA and DOC.

Low congruence of DOC with RA and AC at low coverages highlights the fact that DOC is sensitive to stochastic variation in allele frequencies inevitably worsened at low coverage. However, this should not be seen as a weakness, as this method was originally designed for ultra-deep sequencing (Lighten *et al.* 2014b). Lighten *et al.* (2014b) noted that coverages of thousands of reads may be necessary for the DOC method to reliably genotype individuals with dozens of co-amplifying alleles. Indeed, at high coverages, the repeatability of DOC between two independent PCRs and sequencing experiments in our study at 20 000 coverage reached 91% , which exceeded that reported by these authors for a much simpler guppy MHC system (84% repeatability). Genotyping of the simulated dataset showed high reliability of DOC at coverages > 5000 reads. In real data, however, even at the highest set of coverages (10 000 – 20 000) a few cases were found where DOC called exceptionally low or exceptionally high number of alleles compared to both RA and AC. The presence of the artefact cut-off threshold (*i.e.* the highest DOC) after unusually low numbers of alleles called by DOC in some amplicons might result from (1) preferential amplification of some alleles during PCR or (2) real differences in the number of copies of variants (true and artefactual alleles) present in a genome, which may cause the highest DOC to occur between variants which differ in the numbers of copies in the genome. Indeed we found the more than 20-fold differences in variant “amplification efficiency”, and the estimates were highly correlated between replicates, indicating that the differences were consistent. It cannot be known, however, whether these differences result from uneven amplification or from a disparate number of variant copies among genomes. While variation in “amplification efficiency” in our simulated dataset did not prevent reliable genotyping by DOC, one has to bear in mind that the efficiency was calculated as the average across genotypes. In real dataset, contextspecific differences in amplification efficiencies or in the numbers of copies of particular variants in a given genome might be higher, which might explain appearance of (few) cases of premature maximum DOC values in real data even at the highest coverages.

Our protocol did not involve screening of variants below the highest DOC for consistent low amplifiers, as suggested by Lighten *et al*. (2014b), because in our dataset it would have led to gross overestimation of the number of alleles across the dataset (see methods). In an attempt to implement screening for low amplifiers, we performed an additional analysis in which we excluded variants below 0.4% PAF (a minimum threshold used for AC, see methods for justification but note that Lighten *et al*. 2014a,b argue against using any fixed thresholds). However, this modified screening protocol was not very effective in dealing with cases in which apparently too few alleles were called, adding just a few variants to the final genotypes (File S4, Supporting information). Another difference between our analysis and that of Lighten *et al.* (2014b) was that we did not eliminate “low quality amplicons” based on the absence of a clear inflection point in a plot of cumulative distribution of variants. However, as Lighten *et al*. (2014b) note, in a system with many MHC loci the presence/absence of such inflection point is not easily discernable. Furthermore, the differences in the cumulative sequencing depth plots of ‘good’ amplicons versus amplicons with abnormally low or high numbers of alleles may not be obvious (see Fig. S1, Supporting information). Even if ‘low quality amplicons’ could be eliminated based on objective criteria, this would imply the loss of data which could otherwise be reliably retrieved by other methods.

Analysis of the simulated dataset revealed that the RA method overestimates the number of true variants even at ultra-high coverages, because at ultra-high converges there is a reasonably high chance of getting the same sequencing error in both replicates. Indeed for real data, RA yielded a higher number of alleles than other methods, so it appears that our conservative approach of using the consensus of intermediate genotype classifications only partly dealt with this problem. Thus, the application of RA to ultra-high coverage data may require additional criteria. One possibility is to introduce a minimum PAF. To test such a modification, we carried out an additional analysis using a minimum of 0.4% PAF criterion (as for AC) and indeed, we obtained high agreement with genotypes obtained by AC genotyping method (File S4, Supporting information). Likewise, application of 0.4% PAF threshold to the simulated dataset effectively eliminated all repeatable errors and yielded correct genotypes (File S5, Supporting information).

Thus, using a minimum PAF threshold (based on examination of the data, see methods) appears to be quite effective in eliminating artifacts. Lighten *et al.* (2014b) applied a fixed threshold of three sequences per amplicon to their data (as used by Zagalska-Neubauer *et al.* 2010) to filter artefacts, and concluded that such a threshold would lead to overestimation of the number of variants compared to DOC or copy number variation methods. However, recent studies using the AVT method (*e.g.* Herdegen *et al.* 2014; Promerová *et al.* 2012; Radwan *et al.* 2012; Sepil *et al.* 2012) used PAF thresholds (based on assessment of observed artefacts frequency), rather than fixed thresholds. Using the criterion of three or even ten reads (also tested by Lighten *et al.* 2014b) on a set of data with thousands of reads per amplicon inevitably leads to overestimation of the number of alleles (see File S6, Supporting information), but this would not occur if PAF thresholds were used. In fact, Herdegen *et al.* (2014) used a PAF threshold of 3% for AVT typing of guppies and obtained similar ranges of number of alleles in guppies to those reported by Lighten *et al.* (2014b). Also in the present study we found that AVT and DOC yielded similar results.

Overall, for genotyping of highly complex system, the AC method appears to be the protocol of choice, because of its high reliability (as shown by simulations and agreement with other methods) and repeatability. The AVT genotyping method gave very similar results to AC, but was somewhat inferior in terms of repeatability. RA also agreed reasonably well with AC across the range of coverage, suggesting that it is unnecessary to run replicates of all samples in order to obtain reliable genotypes. As mentioned above, RA appears to suffer from an increased frequency of false positives at ultra-high sequencing depth compared to other methods, and thus requires adjustment to account for repeatable sequencing errors. Finally, the DOC method yielded identical genotypes to AC for most amplicons, but produced a few outliers even at very high coverages.

Although we found evidence of tag jumping in the second PCR (*i.e.* tag combinations that could not be assigned to any amplicon), this phenomenon had little apparent effect on the final genotypes. Indeed Replicate 3 (unique tags) yielded very similar genotypes to Replicates 1 and 2 (unique tag combinations, but forward or reverse tags shared between individuals). Thus, despite the fact that the chimeric variants with swapped tags made up as many as 15% of reads, they were distributed over many variants (*i.e.* many parental pairs combination, with each pair giving rise to many possible chimeras), each with a negligibly low per amplicon frequency. Consequently, each variant was not frequent enough to make a difference to the final genotypes. Nevertheless, these chimeric sequences constituted a considerable loss of data. When barcoding is performed on a large number of individuals, it is not practical to use unique pairs of tags. In such cases, the loss of data from tag-jumping could be partially avoided by PCR-free library construction. However, tag jumping may also occur during ligation of sequencing primers (Schnell *et al.* 2015), but this risk can be avoided if, as in some systems, sequencing primers can be added in a single step, without the requirement for the second PCR of pooled amplicons (see *e.g.* Herdegen *et al.* 2014).

In the course of this study, we have documented the highest-to-date number of MHC class I alleles per individual in a passerine bird, belonging to a group characterized by extreme MHC diversity (*e.g.* O'Connor *et al*. 2016, Sepil *et al.* 2012; Zagalska-Neubauer *et* *al.* 2010). The number of MHC alleles per sedge warbler individual varied considerably, from 19 to 42 alleles according to AC method, suggesting substantial CNV. Our result show that systems with many co-amplifying loci and CNV can be reliably genotyped using reasonably high read coverages (at 5000 reads or higher for RA or AC in the present study). Knowledge of such threshold in coverage should allow researchers to efficiently and cost-effectively genotype the large datasets (hundreds of individuals) that can be required to address questions in ecology and evolutionary biology. Such studies would be also facilitated by our finding that tag jumping does not significantly confound genotyping results, which permits cost-effective tagging of large sets of individuals without the need to use a unique pair of primers for each individual.

Conclusions

Our study demonstrates that HTS enables accurate genotyping of large sets of co-amplifying loci if sufficient coverage per amplicon is reached. This will allow researchers to address important questions in molecular ecology and evolutionary biology, such as the evolution of MHC copy number and MHC associations with disease resistance and mate choice. Our analyses indicated that of the four protocols tested, AC was the most effective. Moreover, AC can easily be implemented with existing software (Sebastian *et al.* 2016). Tag jumping during the second PCR of pooled amplicons caused some data loss, but did not change the genotyping results.

Acknowledgements

We would like to thank Emily O’Connor and Karl Phillips for their comments during the manuscript’s preparation, and Jackie Lighten, Camille Mazzoni, Mark Gillingham and three anonymous reviewers for their comments during peer review. All helped to improve the quality of the final product. This work was supported by grants from the Polish Science

Foundation awarded to Aleksandra Biedrzycka (POMOST 2010-2/1) and National Centre for Science (NCN) UMO-2013/08/A/NZ8/00153 to Jacek Radwan.

Supporting information

Additional Supporting Information may be found in the online version of this article:

File S1. Excel file containing data on repeatability of the methods tested and agreement between them: (a) within method repeatability comparison for AVT, DOC & AC; (b) between method agreement for AVT, RA, DOC & RA; (c) genotyping results for RA (number of alleles in intermediate classifications for replicates 1 and 2; number of alleles after final classification).

File S2. Genotypes, alleles and their amplification efficiencies used in simulations

File S3. Excel file containing the number of alleles and false negative and false discovery rates for the genotyping methods tested.

File S4. Excel file containing data on repeatability of the RA and DOC methods and their agreement with other methods after introducing a minimum cutoff of 0.4% PAF and screening for 'low amplifiers' in DOC.

File S5. Genotypes, alleles and their amplification efficiencies used in simulations after introducing a minimum cutoff of 0.4% PAF and screening for 'low amplifiers' in DOC.

File S6. Calculation of probability of obtaining at least x identical reads in a sample of n reads.

Table S1. Numbers of reads after length and quality filtering.

Table S2. Estimates of coverage confidence.

Figure S1. Example of the cumulative sequencing depth plots of 'good' amplicons versus amplicons with abnormally low or high numbers of alleles called by the DOC method.

Figure S2. Variants ordered by descending per amplicon frequency.

Figure S3. Pairwise comparison of genotyping agreement between all methods.

Figure S4. Distribution of the relative amplification efficiencies calculated for Replicate 1 and 2 and the correlation between them.

Data accessibility

Merged and cleaned reads (see Materials and Methods section) are deposited in FASTQ format in the European Nucleotide Archive (ENA) with the study accession PRJEB11775.

Data is also available from http://evobiolab.biol.amu.edu.pl/amplisat/index.php?examples.

References

Babik W (2010) Methods for MHC genotyping in non-model vertebrates. *Molecular Ecology Resources* **10**, 237-251.

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.

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.

Blackwell JM, Jamieson SE, Burgner D (2009) HLA and infectious diseases. *Clinical Microbiology Reviews* **22**, 370-385.

Bonneaud C, Mazuc J, Chastel O, Westerdahl H, Sorci G (2004) Terminal investment induced by immune challenge and fitness traits associated with major histocompatibility complex in the house sparrow. *Evolution* **58**, 2823-2830.

Ejsmond MJ, Radwan J, Wilson AB (2014) Sexual selection and the evolutionary dynamics of the Major Histocompatibility Complex. *Proc. R. Soc. Lond. B* **281**, 20141662.

Ferrandiz-Rovira M, Bigot T, Allaine D, Callait-Cardinal MP, Cohas A (2015) Large-scale genotyping of highly polymorphic loci by next-generation sequencing: how to overcome the challenges to reliably genotype individuals? *Heredity* **114**, 485-493.

Herdegen M, Babik W, Radwan J (2014) Selective pressures on MHC class II genes in the guppy (*Poecilia reticulata*) as inferred by hierarchical analysis of population structure. *Journal of Evolutionary Biology* **27**, 2347-2359.

Hess CM, Edwards SV (2002) The evolution of the major histocompatibility complex in birds. *Bioscience* **52**, 423-431.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**, R143.

Janeway CA, Travers P, Walport D, Shlomchik MJ (2004) *Immunobiology: The Immune System in Health and Disease* Garland Publishing, New York.

Karlsson M, Westerdahl H (2013) Characteristics of MHC class I genes in house sparrows *Passer domesticus* as revealed by long cDNA yranscripts and amplicon sequencing. *Journal of Molecular Evolution* **77**, 8-21.

Kamiya, T., et al. (2014) A quantitative review of MHC-based mating preference: the role of diversity and dissimilarity. *Molecular ecology* **23**, 5151-5163.

Kaufman J, Milne S, Gobel TWF*, et al.* (1999) The chicken B locus is a minimal essential major histocompatibility complex. *Nature* **401**, 923-925.

Kelley J, Walter L, Trowsdale J (2005) Comparative genomics of major histocompatibility complexes. *Immunogenetics*, 56, 683–695.

Kloch A, Babik W, Bajer A, Sinski E, Radwan J (2010a) 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.

Kloch A, Babik W, Bajer A, Siński E, Radwan J (2010b) 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.

Lighten J, Van Oosterhout C, Bentzen P (2014a) Critical review of NGS analyses for de novo genotyping multigene families. *Molecular Ecology* **23**, 3957-3972.

Lighten J, Van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014b) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Molecular Ecology Resources*.

Liu L, Li Y, Li S*, et al.* (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 251364.

Loman NJ, Misra R V, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

Magoc T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963.

McElroy KE, Luciani F, Thomas T (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. *Bmc Genomics* **13**, 74.

Nadachowska-Brzyska K, Zielinski P, Radwan J, Babik W (2012) Interspecific hybridization increases MHC class II diversity in two sister species of newts. *Molecular Ecology* **21**, 887-906.

O'Connor EA, Strandh M, Hasselquist D, Nilsson J-Å, Westerdah H (2016) The evolution of highly variable immunity genes across a passerine bird radiation. *Molecular Ecology* in press, 10.1111/mec.13530.

Pavey SA, Sevellec M, Adam W*, et al.* (2013) Nonparallelism in MHCII diversity accompanies nonparallelism in pathogen infection of lake whitefish (C*oregonus clupeaformis*) species pairs as revealed by next-generation sequencing. *Molecular Ecology* **22**, 3833-3849.

Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**, 7-21.

Promerová M, Babik W, Bryja J*, et al.* (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.

Radwan J, Kuduk K, Levy E, LeBas N, Babik W (2014) Parasite load and MHC diversity in undisturbed and agriculturally modified habitats of the ornate dragon lizard. *Molecular Ecology* **23**, 5966-5978.

Radwan J, Zagalska-Neubauer M, Cichon M*, et al.* (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Molecular Ecology* **21**, 2469-2479.

Ross MG, Russ C, Costello M *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome biology*, **14**, R51.

Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* **5**, 1289-1303.

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

Sepil I, Moghadam HK, Huchard E, Sheldon BC (2012) Characterization and 454 pyrosequencing of Major Histocompatibility Complex class I genes in the great tit reveal complexity in a passerine system. *BMC Evolutionary Biology* **12**.

Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in Zoology* **12**, 16.

Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. *Bmc Genomics* **14**, 542.

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B-Biological*

Stutz WE, Bolnick DI (2014) Stepwise Threshold Clustering: A new method for genotyping MHC loci using Next-Generation Sequencing technology. *Plos One* **9**, e100587.

Syed F, Grunenwald H, Caruccio N (2009) Optimized library preparation method for next-generation sequencing. *Nature Methods* **6**, I-II.

Wegner KM, Kalbe M, Kurtz J, Reusch TBH, Milinski M (2003) Parasite selection for immunogenetic optimality. *Science* **301**, 1343-1343.

Westerdahl H, Hansson B, Bensch S, Hasselquist D (2004a) Between-year variation of MHC allele frequencies in great reed warblers: selection or drift? *Journal of Evolutionary Biology* **17**, 485-492.

Westerdahl H, Wittzell H, von Schantz T (1999) Polymorphism and transcription of MHC class I genes in a passerine bird, the great reed warbler. *Immunogenetics* **49**, 158-170.

Westerdahl H, Wittzell H, von Schantz T, Bensch S (2004b) MHC class I typing in a songbird with numerous loci and high polymorphism using motif-specific PCR and DGGE. *Heredity* **92**, 534-542.

Winternitz JC, Minchey SG, Garamszegi LZ*, et al.* (2013) Sexual selection explains more functional variation in the mammalian major histocompatibility complex than parasitism. *Proceedings of the Royal Society B-Biological Sciences* **280**, 2013.1605.

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

Figures

Figure 1. Dot plot comparison of the per-individual number of alleles called using different methods: AVT – allele validation threshold (Radwan *et al.* 2012), RA – replicate amplicon (Sommer et al. 2013), DOC - degree of change (Lighten et al. 2014b), AC - adjustable clustering using AmpliSAS tool (Sebastian et al. 2016), and at different levels of coverage. Only the 1st replicate is shown for AVT, DOC and AC.

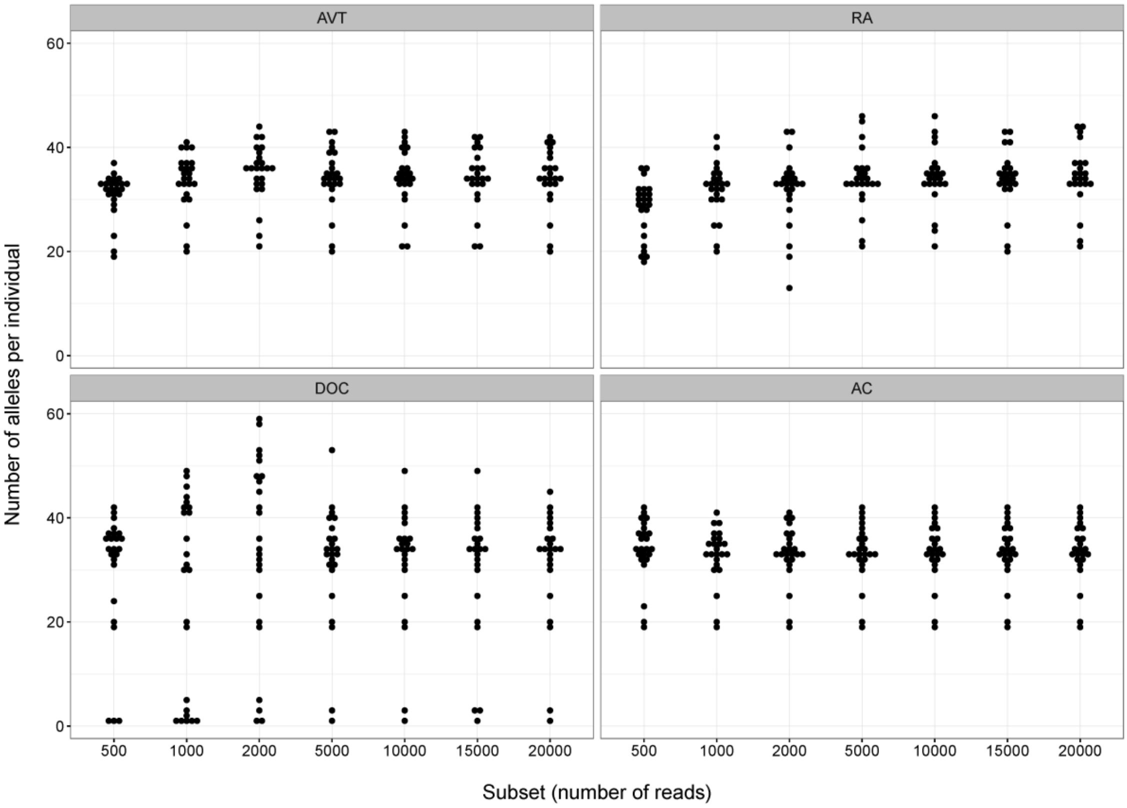


Figure 2. Comparison of genotyping agreement between two replicates using different methods: AVT – allele validation threshold (Radwan *et al.* 2012), DOC - degree of change (Lighten et al. 2014b), and AC - adjustable clustering using AmpliSAS tool (Sebastian et al. 2016). Within each subset, repeatability is calculated as the number of identical alleles divided by the total number of alleles called in both replicates in a given individual (both common and unique variants for each replicate). Figure represents Tukey's boxplot: the bottom and top of each box represents the 1st and 3rd quartiles, the band indicates the 2nd quartile (the median), whiskers represent the lowest/highest datum within 1.5 interquartile range of the lower/upper quartile and outliers are represented by full circles.

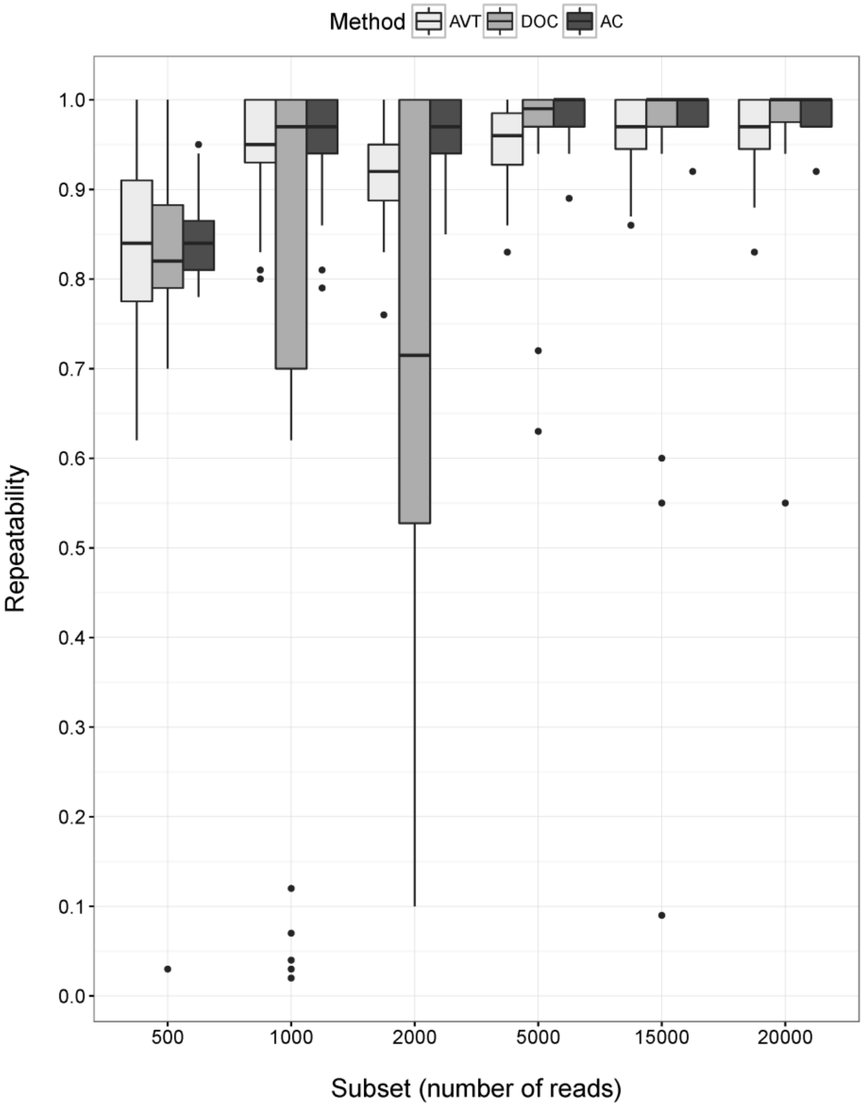


Figure 3. Pairwise comparison of genotyping agreement across subsets, between : RA – replicate amplicon (Sommer *et al.* 2013), DOC - degree of change (Lighten et al. 2014b), and AC - adjustable clustering using AmpliSAS tool (Sebastian *et al.* 2016). Agreement is calculated as the number of identical alleles dived by a summarized number of alleles called by the two methods (both common and unique alleles for each method) in a given individual. Common alleles from replicates are used. Figure represents Tukey's boxplot. For extended pairwise comparison among all methods (including AVT – allele validation threshold (Radwan *et al.* 2012) see Figure S3.

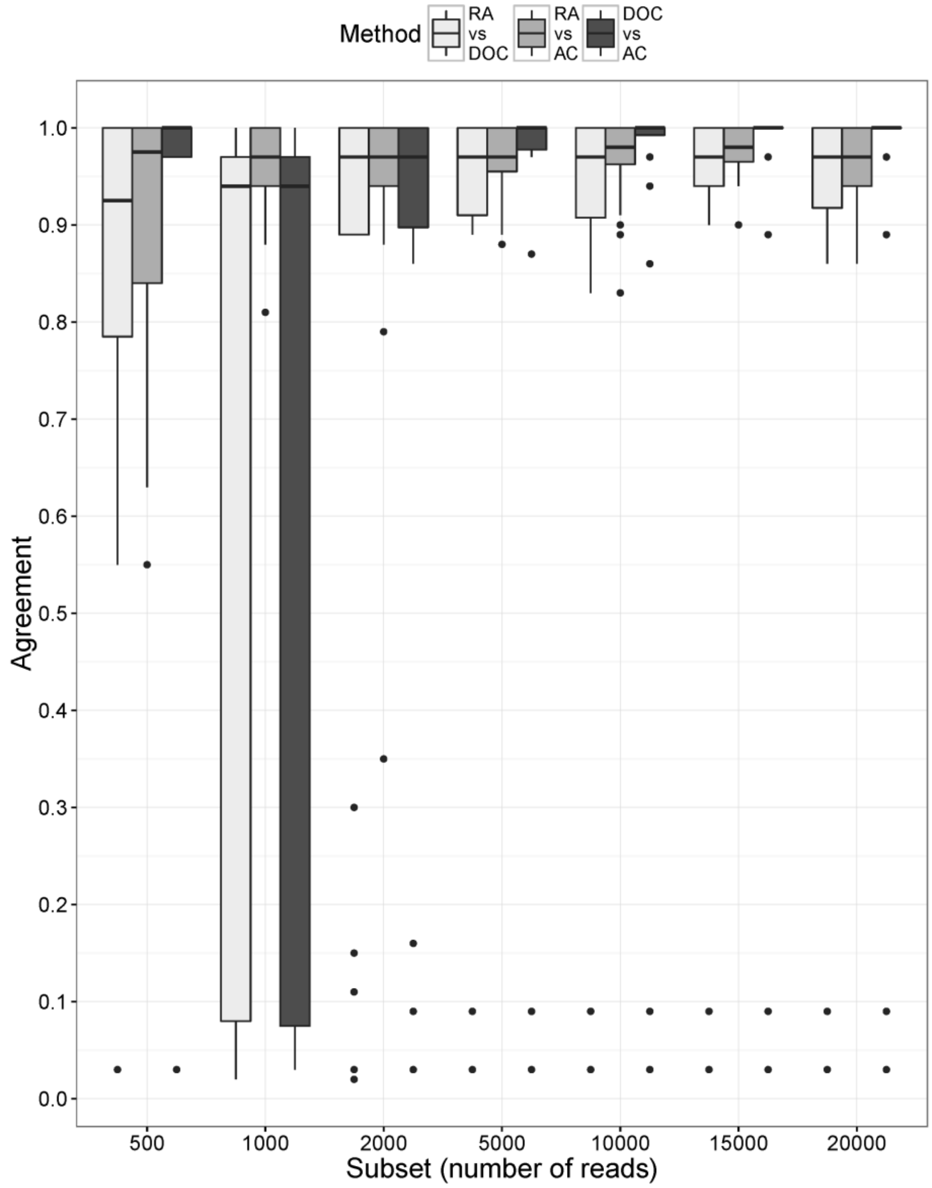
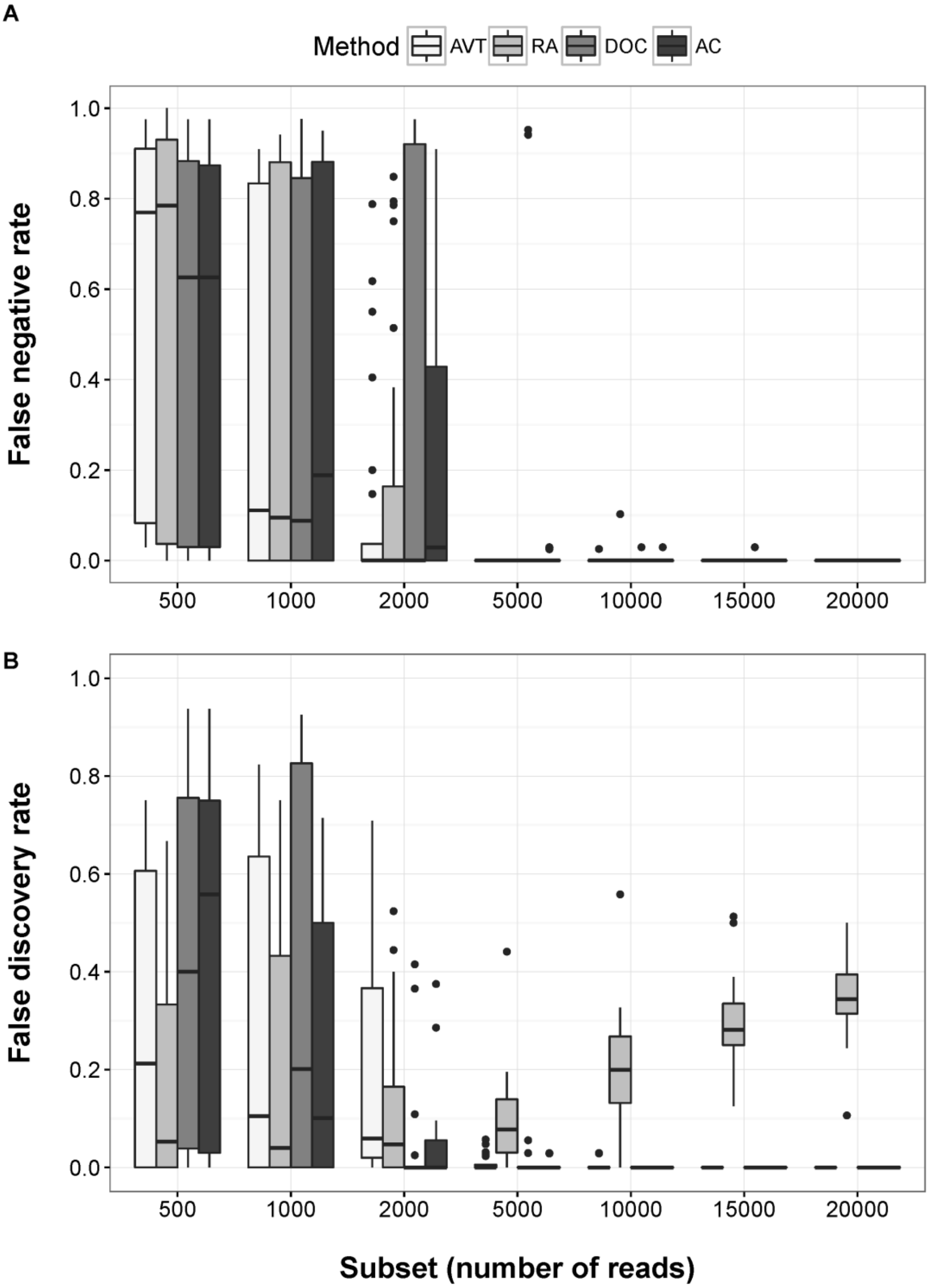


Figure 4. False negative (A) and false discovery (B) rates calculated for simulated datasets for different methods: AVT – allele validation threshold (Radwan *et al.* 2012), RA – replicate amplicon (Sommer et al. 2013), DOC - degree of change (Lighten et al. 2014b), AC - adjustable clustering using AmpliSAS tool (Sebastian et al. 2016), and at different levels of coverage. Only the 1st replicate is shown for AVT, DOC and AC.


A
Method
AVT
RA
DOC
AC
False negative rate
B
False discovery rate
Subset (number of reads)