Received: 24 February 2020 | Revised: 19 October 2020 | Accepted: 22 October 2020

DOI: 10.1111/1755-0998.13290

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

A novel workflow to improve genotyping of multigene families in wildlife species: An experimental set-up with a known model system

Mark A. F. Gillingham[1] | B. Karina Montero[1,2] | Kerstin Wihelm[1] | Kara Grudzus[1] | Simone Sommer[1] | Pablo S. C. Santos[1]

[1]Institute of Evolutionary Ecology and Conservation Genomics, Ulm Universität, Ulm, Germany

[2]Zoological Institute, Animal Ecology and Conservation, Biocenter Grindel, Universität Hamburg, Hamburg,, Germany

Correspondence
Mark A. F. Gillingham and Pablo S. C. Santos, Institute of Evolutionary Ecology and Conservation Genomics, Ulm Universität – Ulm, Germany.
Emails: mark.alan.gillingham@gmail.com; mail@psc-santos.com

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG Gi 1065/2-1

Abstract

Genotyping complex multigene families in novel systems is particularly challenging. Target primers frequently amplify simultaneously multiple loci leading to high PCR and sequencing artefacts such as chimeras and allele amplification bias. Most genotyping pipelines have been validated in nonmodel systems whereby the real genotype is unknown and the generation of artefacts may be highly repeatable. Further hindering accurate genotyping, the relationship between artefacts and genotype complexity (i.e. number of alleles per genotype) within a PCR remains poorly described. Here, we investigated the latter by experimentally combining multiple known major histocompatibility complex (MHC) haplotypes of a model organism (chicken, *Gallus gallus*, 43 artificial genotypes with 2–13 alleles per amplicon). In addition to well-defined 'optimal' primers, we simulated a nonmodel species situation by designing 'cross-species' primers based on sequence data from closely related Galliform species. We applied a novel open-source genotyping pipeline (ACACIA; https://gitlab.com/psc_santos/ACACIA), and compared its performance with another, previously published pipeline (AmpliSAS). Allele calling accuracy was higher when using ACACIA (98.5% versus 97% and 77.8% versus 75% for the 'optimal' and 'cross-species' data sets, respectively). Systematic allele dropout of three alleles owing to primer mismatch in the 'cross-species' data set explained high allele calling repeatability (100% when using ACACIA) despite low accuracy, demonstrating that repeatability can be misleading when evaluating genotyping workflows. Genotype complexity was positively associated with nonchimeric artefacts, chimeric artefacts (nonlinearly by levelling when amplifying more than 4–6 alleles) and allele amplification bias. Our study exemplifies and demonstrates pitfalls researchers should avoid to reliably genotype complex multigene families.

KEYWORDS

ACACIA, allele dropout, amplicon genotyping, high-throughput sequencing, MHC, multigene family, open-source genotyping pipeline, PCR amplification bias, sequencing bias

1 | INTRODUCTION

A key challenge for molecular ecologists is that they frequently work on systems with limited to no knowledge of their genomes. Multigene complexes such as resistance genes (R-genes) and self-incompatibility genes (SI-genes) in plants, immunoglobulin superfamily and major histocompatibility genes (MHC) in vertebrates, and homeobox genes in animals, plants and fungi, among many others are particularly challenging to genotype in nonmodel organisms. While a large number of de novo genomes have been published using short-sequencing technology, most traditional genome assembly of nonmodel organisms has not been able to assemble the highly repetitive genomic regions of multigene families (O'Connor et al., 2019). Recent long-read single-molecule sequencing technologies offer a very promising avenue to characterize multigene families in the future (Fuselli et al., 2018; Gordon et al., 2016; Larsen et al., 2014; Westbrook et al., 2015). However, the high sequencing errors and the high cost associated with required sequencing depth continue to constrain characterization of nonmodel organisms particularly when genotyping a large number of individuals with highly complex multigene families. Therefore, the development of a genotyping approach for specific multigene families (i.e. amplicon-based genotyping) typically continues to rely on information from closely related species available in genetic databases, which may have very different gene duplication and deletion events. As a result of high sequence similarity from recent gene duplication events, polymerase chain reaction (PCR) primers will frequently bind across multiple loci leading to the amplification of multiple allelic variants (Babik, 2010; Biedrzycka et al., 2017; Burri et al., 2014; Lighten et al., 2014; Lighten et al., 2014; Sebastian et al., 2016; Sommer et al., 2013). Assessing and validating genotyping methods can be particularly challenging when the number of loci targeted is unknown.

Unspecific locus amplification may lead to several biases during PCR. First, chimeric sequences (hereafter 'chimeras', which may arise because of incomplete extension of sequences during a PCR cycle and are subsequently completed with a different allele template) are likely to become more frequent as more loci are amplified within an amplicon simply because there will be more gene variants from which chimeras can be generated (Lenz & Becker, 2008). Second, amplification bias of some gene variants relative to others may occur because primers preferentially bind to some alleles/loci (hereafter referred to as 'PCR competition'; Marmesat et al., 2016; Sommer et al., 2013). Creative solutions in primer design and in PCR conditions, such as using pooled primers instead of degenerate primers (Marmesat et al., 2016), reducing the number of cycles and modifying elongation steps of PCRs (Judo et al., 1998; Lenz & Becker, 2008; Smyth et al., 2010), can significantly reduce amplification bias. However, even after the application of such methods, PCR biases will nonetheless persist and may lead to genotyping errors because: (a) chimeras may be difficult to distinguish from valid recombinant gene variants (frequent in multigene complexes (Chen et al., 2007)), resulting either in PCR artefacts being falsely validated as true allelic variants (type I errors, hereafter referred to as 'false positives') or in true allelic variants being falsely rejected as an artefact (type II errors, hereafter referred to as 'allele dropout'); and (b) poorly amplified allelic variants may not be sequenced resulting in allele dropout, particularly when the number of sequences per amplicon (a set of sequences of a target region generated within a PCR) is low (Biedrzycka et al., 2017; Galan et al., 2010; Lighten, Oosterhout, & Bentzen, 2014; Lighten, Oosterhout, Paterson, et al., 2014; Sommer et al., 2013).

The rapid dissemination of high-throughput DNA sequencing (HTS) platforms has provided molecular ecologists with an exciting opportunity to tackle the parallelized genotyping of multiple markers in numerous species since it has allowed the generation of thousands of sequences (termed 'reads') per amplicon at a fraction of the cost and time needed by previous methods, which typically involved laboriously isolating individual sequences via a cloning vector followed by Sanger sequencing (Babik, 2010; Lighten, Oosterhout, Paterson, et al., 2014; Sommer et al., 2013; Montero et al., 2018). However, HTS platforms have their own limitations, the most relevant being the relatively high amount of sequencing errors generated in a typical sequencing run (Glenn, 2011; Huse et al., 2007; McElroy et al., 2012; Ross et al., 2013; Sommer et al., 2013). For instance, Illumina, currently the mainstream technology for HTS amplicon sequencing, report an error rate (primarily substitutions of base pairs) of ≤0.1% per base for ≥75%–85% of bases (see Glenn, 2011 for details), although final error rates are likely to be much higher and can reach up to 6% (McElroy et al., 2012). Indeed, previous genotyping studies in multilocus systems (>10) reported average amplification and sequencing artefact rates of 1.5%–2.5% per amplicon (Promerová et al., 2012; Radwan et al., 2012; Sepil et al., 2012). Therefore, PCR competition when amplifying multiple loci per amplicon means that sequences from some genuine allelic variants occur at a similar frequency to PCR artefacts or sequencing errors (Biedrzycka et al., 2017; Galan et al., 2010; Lighten, Oosterhout, & Bentzen, 2014; Sommer et al., 2013). In this scenario, poorly amplified alleles cannot be easily distinguished from artefacts during allele validation leading to further false positives and allele dropout during genotyping.

The need to distinguish PCR and sequencing artefacts from valid allelic variants has led to the development of multiple bioinformatic workflows (i.e. a set of bioinformatic steps during processing of sequencing data, which eventually leads to genotyping, hereafter referred to as a 'genotyping pipeline'). While all genotyping pipelines rely to some degree on the assumption that artefacts are less frequent than genuine allelic variants, they vary in the approach used to discriminate poorly amplified allelic variants from artefacts. Genotyping pipelines for complex gene families have been extensively reviewed in Biedrzycka et al. (2017). Recently developed pipelines cluster artefacts to their putative parental sequences, thereby increasing the read depths of true variants (Lighten, Oosterhout, Paterson, et al., 2014; Pavey et al., 2013; Sebastian et al., 2016; Stutz & Bolnick, 2014). Currently, the most commonly used pipeline for MHC studies is the AmpliSAS web server pipeline (Sebastian et al., 2016). After chimera removal, AmpliSAS uses a clustering

algorithm to discriminate between artefacts and allelic variants, which take into account the error rate of a particular HTS technology and the expected lengths of the amplified sequences. This is achieved in a stepwise manner, whereby it first clusters the most common variant (according to specified error rates) and then moves on to the next most common variant, until no variant remains to be clustered. Microbiome studies, which typically amplify hypervariable regions of the 16S rRNA gene from very diverse bacterial communities within a single amplicon, have used a similar strategy to AmpliSAS, whereby potential artefactual variants are clustered to suspected parental sequences using Shannon entropy (referred to as 'Oligotyping' (Eren et al., 2013)) or other similar clustering methods (Amir et al., 2017; Callahan et al., 2016).

Most of the amplicon genotyping pipelines for multigene families available to molecular ecologists have only been tested on nonmodel organisms for which the real genotype is unknown (but see Sebastian et al. (2016)). As a consequence, studies have frequently depended on repeatability of duplicated samples to justify genotyping pipeline reliability (Biedrzycka et al., 2017; Galan et al., 2010; Lighten, Oosterhout, Paterson, et al., 2014; Radwan et al., 2012; Sebastian et al., 2016; Sommer et al., 2013). However, for a given set of PCR primers and sequencing technology, PCR and sequencing bias will be consistently repeatable and relying on repeatability to validate a genotyping pipeline may be misleading (Biedrzycka et al., 2017). For instance, the high rate of Illumina substitution errors are known to be nonrandom (see references within Sebastian et al. (2016)), and therefore, variants that result from substitution errors are highly repeatable between amplicons (Biedrzycka et al., 2017). Furthermore, while the generation of PCR and sequencing artefacts is well known, the precise relationship between artefacts and the number of alleles amplified within an amplicon for a given set of primers and sequencing technology has never been described. Yet, having a clear indication of this relationship is an important step in predicting what are the optimal pipelines settings (e.g. predicting error rates) for a given number of loci amplified within an amplicon. The latter can only be achieved by experimentally manipulating the number of loci of a priori known genotypes before PCR amplification and HTS sequencing.

In this study, we artificially generated genotypes of known combinations of MHC alleles of a model organism (the chicken, *Gallus gallus*) by mixing DNA samples from 7 haplotypes as an example of a target multigene region of interest to molecular ecologists and to assess the accuracy of amplicon-based genotyping. While we focus on the MHC hereafter, all methods and results are applicable to any multigene family. Like many multigene complexes, MHC genes are subject to multiple gene conversion, duplication and deletion events (Nei & Rooney, 2005; Nei et al., 1997; Parham & Ohta, 1996) and MHC gene copies vary considerably across and even within a species (reviewed in ref. Kelley et al. (2005)). Therefore, the number of MHC loci present in a nonmodel study system often remains unknown. For instance, the range in the number of MHC class IIB alleles within an individual was found to be as high as 19–42 in some passerine species (Biedrzycka et al., 2017). In contrast, the chicken MHC B complex is unusually simple, leading it to be coined as a 'minimal essential' system, with only two MHC class I loci and two MHC class IIB loci (Kaufman, Jacob, et al., 1999; Kaufman, Milne, et al., 1999; Kaufman et al., 1995). The latter is therefore an ideal system to validate MHC genotyping pipelines for the following reasons: (a) the structure of the B complex is well known with well-defined primers in conserved regions; (b) the well-characterized B complex haplotype lineages can be used so that the expected MHC genotyping results are known prior to sequencing and genotyping; and (c) the number of alleles amplified within an amplicon can be experimentally engineered by combining DNA samples from multiple MHC B complex haplotypes.

To perform the genotyping of known chicken MHC haplotypes and extract data concerning PCR and sequencing artefacts at each step of the genotyping workflow, we developed and calibrated our own genotyping pipeline (named ACACIA for Allele CAlling proCedure for Illumina Amplicon sequencing data). ACACIA is written in Python, and it takes advantage of several previously published software dedicated to genomics (detailed in the methods), as well as the widely used Biopython library (Cock et al., 2009) to handle genomic data. We experimentally generated a MHC data set with a range 2–13 alleles within amplicon by combining DNA samples from multiple chicken MHC B complex haplotypes. Since MHC B complex in chickens is well characterized, optimal primers to amplify the entire exons that code for the antigen binding regions have been developed (Goto et al., 2002; Shaw et al., 2007). However, in most wildlife species such extensive genomic information around the region of interest is unavailable. To avoid the problems associated with overfitting ACACIA to one specific data set and to replicate the challenge of designing primers for a nonmodel species, we additionally designed primers within the exons coding for antigen binding regions using sequence data from closely related Galliform species that were not chickens (hereafter referred to as 'cross-species' primers). The latter enabled us to gain insight into the relative amount of artefacts generated by an intentionally suboptimal set of primers, for which we expected allele dropout.

Specifically, this study aimed to:

1. Validate ACACIA using experimentally manipulated genotypes with a range of allelic variants (2–13) within an amplicon that are known a priori;
2. To investigate the relationship between multigene family complexity (i.e. number of alleles amplified within an amplicon) and artefacts generated by PCR and sequencing (i.e. chimeras and insertions/deletions)

2 | MATERIALS AND METHODS

2.1 | Samples and DNA extraction

Chicken blood samples originated from experimental inbred lines kept at the Institute for Animal Health at Compton, UK (lines 72, C, WL and N), and the Basel Institute for Immunology in Basel, Switzerland

(lines H.B15 and H.B19^{+}), as detailed in Jacob et al. (2000), Shaw et al. (2007) and Wallny et al. (2006). These lines carry seven common B haplotypes: B2 (line 72), B4 and B12 (line C), B14 (line WL, sometimes referred as W), B15 (H.B15), B19 (H.B19) and B21 (line N). All the lines used in this study are homozygotes (NCBI Accession nos.: AJ248572 to AJ248586). In each haplotype are two class IIB loci: BLB1 (previously known as BLBI or BLBminor) and BLB2 (BLBII or BLBmajor), with alleles now designated as BLB1*02 and BLB2*02 from the B2 haplotype, etc. All alleles have different nucleotide sequences, except BLB1*12 and BLB1*19. DNA was isolated from blood cells by a salting out procedure (Miller et al., 1988).

2.2 | Generating 43 artificial MHC genotypes

We artificially generated 43 genotypes of varying allelic complexity by combining equimolar amounts of DNA samples from the seven MHC haplotypes mentioned above (Table 1; created genotypes listed in Table S1).

2.3 | Optimal primers for chicken MHC class IIB

We targeted 241 bp of the 270 bp exon 2 of MHC class IIB, the polymorphic region known to code for antigen binding sites, using the primers OL284BL (5-GTGCCCGCAGCGTTCTTC-3) and RV280BL (5-TCCTCTGCACCGTGAAGG-3; Goto et al., 2002). The primers are not locus-specific and bind to both loci of the chicken B complex.

2.4 | Cross-species primer design for chicken MHC class IIB

To replicate designing primers without any a priori knowledge of the species' MHC Class IIB structure or sequences, we downloaded 61 exon 2 MHC class IIB sequences from seven Galliform species (*Coturnix japonica*, *Crossoptilon crossoptilon*, *Meleagris gallopavo*, *Numida meleagris*, *Pavo cristatus*, *Perdix perdix* and *Phasianus colchicus*, all accession numbers are listed in the Table S2) from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). We then used Primer3 (Rozen & Skaletsky, 1999; Untergasser et al., 2012) to design the forward primer GagaF1 (5-WTCTACAACCGG CAGCAGT-3) and the reverse primer GagaR2 (5-TCCTCTGCACC GTGAWGGAC-3) aiming at amplifying 151 bp of exon 2. No species were given more weight than others during primer design, and all default parameters of Primer3 (concerning melting temperatures and structural settings) were kept. The only exception is that we allowed up to two degenerate positions in the primer sequence. Note here that the design of primers was deliberately suboptimal to convey the challenges of primer design. Optimal primer design should include sequences from other avian groups than Galliform species and consider phylogenetic relationships to weigh variants (Babik, 2010; Burri et al., 2014; see Discussion for more details on primer design).

TABLE 1 In this study, we generated 43 genotypes that were amplified twice (duplicated). The number of alleles per genotype and the number of genotypes with that number of all alleles are shown. The list of haplotypes used to artificially create the genotypes is listed in the Table S1

Number of alleles per genotype	Number of genotypes
2	7
4	7
6	7
8	7
10	7
11	5
12	2
13	1
Total	43

2.5 | PCR amplification, library preparation and high-throughput sequencing

For both data sets, we replicated all individuals to estimate repeatability ($n_{individuals} = 43$ and $n_{amplicons} = 86$). Individual PCRs were tagged with a 10-base pair identifier, using a standardized Fluidigm protocol (Access Array™ System for Illumina Sequencing Systems, ©Fluidigm Corporation). We first performed a target-specific PCR with the CS1 adapter and the CS2 adapter appended. To enrich base pair diversity of our libraries during sequencing, we added four random bases to our forward primer. The CS1 and CS2 adapters were then used in a second PCR to add a 10-bp barcode sequence and the adapter sequences used by the Illumina instrument during sequencing.

The first PCR consisted of 3–5 ng of extracted DNA, 0.5 units FastStart Taq DNA Polymerase (Roche Applied Science), 1× PCR buffer, 4.5 mM $MgCl_2$, 250 M of each dNTP, 0.5 M primers and 5% dimethyl sulphoxide (DMSO). The PCR was carried out with an initial denaturation step at 95°C for 4 min followed by 30 cycles at 95°C for 30 s, 60°C for 30 s and 72°C for 45 s, and a final extension step at 72°C for 10 min. The second PCR contained 2 L of the product generated by the initial PCR, 80 nM per barcode primer, 0.5 units FastStart Taq DNA Polymerase, 1× PCR buffer, 4.5 mM $MgCl_2$ 250 M of each dNTP, and 5% dimethyl sulphoxide (DMSO) in a final volume of 20 L. Cycling conditions were the same as those outlined above, but the number of cycles was reduced to ten. Reducing the number of PCR cycles, the elongation time within PCR cycles and omitting the final extension step is recommended to reduce the number of chimeras when coamplifying multiple loci, because most incomplete primer extensions that generate chimeras are thought to be formed in the final cycles of PCRs and during the final extension step (see Discussion; Judo et al., 1998; Lenz & Becker, 2008; Smyth et al., 2010). However, we chose to process samples using conventional PCR conditions, because a high number of cycles may

be necessary in some study systems and we wanted to replicate conditions used in most MHC wildlife studies. Thus, we purposefully wanted to evaluate the robustness of our pipeline in the more challenging setting where a high number of artefacts might be generated due to suboptimal PCR conditions.

PCR products were purified using an Agilent AMPure XP (Beckman Coulter) bead cleanup kit. The fragment size and DNA concentration of the cleaned PCR products were estimated with the QIAxcel Advanced System (Qiagen) and by UV/VIS spectroscopy on an Xpose instrument (Trinean, Gentbrugge, Belgium). Samples were then pooled to equimolar amounts of DNA. The library was prepared as recommended by Illumina (MiSeq System Denature and Dilute Libraries Guide 15039740 v05) and was loaded at 7.5 pM on a MiSeq flow cell with a 10% PhiX spike. Paired-end sequencing was performed over 2 × 251 cycles.

2.6 | Data analysis with the ACACIA pipeline

ACACIA consists of 11 consecutive steps of data processing. The software requires two nonstandard python libraries (PANDAS (McKinney, 2010) and BIOPYTHON (Cock et al., 2009)) and six third-party software (FASTQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/), FLASH (Magoč & Salzberg, 2011), VSEARCH (Rognes et al., 2016), BLAST (Altschul et al., 1990), MAFFT (Katoh & Standley, 2013) and OLIGOTYPING (Eren et al., 2013)), which can all be installed with one command. The input files are any number of FASTq files, which are the current canonical output of the Illumina platform. The step-by-step workflow is described below:

1. Generating Quality Reports: Sequencing quality is assessed for each FASTq file yielded by the sequencing platform, with the FastQC tool. Reports for each file are produced in HTML format for visual inspection.
2. Trimming low-quality ends of forward and reverse reads (optional): The information generated in step #1 is crucial for an informed decision about how many (if any) bases should be trimmed out of each read. If trimming is performed here, step #1 is repeated. Shorter FASTq files are generated as output of this step.
3. Merging paired-end reads (optional): This concerns projects with paired-end sequencing only and should be skipped if using data from single-end sequencing (note: the names of the paired forward and reverse FASTq files should be identical prior to the first '_'character, e.g.: ID1S1L001_R1_001.fastq and ID1S1L001_R2_001.fastq). The reads of file pairs are merged using FLASh (Magoč & Salzberg, 2011). The minimum and maximum lengths of overlap during merging can be adjusted by the user to improve performance (defaults are zero and read length, respectively). New FASTq files with merged sequences are generated as output, as well as a series of.log files that allow users to monitor merging performance.
4. Trimming primers: After prompting users to enter the sequences of the primers used for target amplification, ACACIA trims primer sequences from both ends of the merged sequences (IUPAC nucleotide ambiguity codes are allowed). Primerless sequences are written into FASTq files, which are the output of this step. The Python functions for trimming primers and low-quality ends (step #2) are part of the core ACACIA pipeline. External tools were avoided here to decrease dependency on further software.
5. Quality control: Users are then prompted to enter the values of two parameters (q and p) to filter sequences based on their mean phred scores. First, q stands for quality and denotes a phred score threshold that can take values from 0 to 40. Second, p stands for percentage and denotes the proportion of bases, in any given sequence, that have to achieve at least the quality threshold q for that sequence to pass the quality filter. ACACIA uses the default values $q = 30$ and $p = 90$ if users do not explicitly change them. In practical terms, these thresholds correspond to an error probability lower than 10–3 in at least 90% of bases for each sequence. All information on quality data of sequences passing this filter is then removed, and FASTA files with high-quality sequences are given as the output of this step.
6. Removing singletons: A large proportion of sequences contain random errors inherent to the sequencing technology (Quail et al., 2012). To decrease file sizes without risking loss of relevant allele information, ACACIA removes all singletons (sequences that appear one single time) in an individual amplicon.
7. Removing chimeras: The chimera identification tool VSEARCH (Rognes et al., 2016) is employed here, with slightly altered settings (alignwidth = 0 and mindiffs = 1) aiming at increasing sensitivity to chimeras that diverge very little from one of the 'parent' sequences. FASTA files with nonchimeric sequences, along with log files for each individual amplicon, are given as output.
8. Removing unrelated sequences: All remaining sequences are then compared with a set of reference sequences chosen by users. This step aims at removing sequences that passed all filters so far but are products of unspecific priming during PCR. Typically, sequences phylogenetically related to those being analysed can be downloaded from the GenBank (www.ncbi.nlm.nih.gov/genbank/). Users are prompted to provide one FASTA file with reference sequences, which is converted by ACACIA to a local BLAST database (Altschul et al., 1990) and used for BLAST. Only sequences yielding high-scoring hits to the local database ($E \leq 10^{-10}$) are written into new FASTA files as an output of this step, which is the workflow's last filtering procedure.
9. Aligning: The MAFFT aligner (Katoh & Standley, 2013) is used to perform global alignments of sequences that have passed filters. Since all sequences are pooled into one single alignment output file, the individual IDs are now transferred from file names into the FASTA sequence headers. We have successfully aligned up to 603,513 sequences in a desktop computer with four CPUs and 32GB of RAM. Users with a significantly higher number of sequences might find it useful to increase the computational

parallelization of the aligner as described recently (Nakamura et al., 2018).

10. Calling candidate alleles: The OLIGOTYPING tool (Eren et al., 2013) is used to call candidate alleles. Although originally conceived as a tool for identifying variants from microbiome 16S rRNA amplicon sequencing projects, we recognized Oligotyping as ideal for other forms of highly variable amplicon sequencing projects. This step consists of concatenating high-information nucleotide positions (defined by entropy analysis of the alignment produced in the previous step) and subsequently using entropy information to cluster divergent variants, while grouping redundant information and filtering out artefacts. Although Oligotyping was conceived as a supervised tool, we automated the selection of parameter values aiming at high tolerance. This has the advantage of running an unsupervised instance of Oligotype as a pipeline step, at the cost of keeping potential false positives among the results. Report files with a list of candidate alleles grouped by individual amplicons are the output of this step.
11. Allele calling and final reporting: A Python script is used to perform the final allele calling by filtering out Oligotyping results according to the following criteria:

–. Removal of unique allele variants (Y/N). Setting Y (yes) removes all alleles identified in one single individual amplicon.

–. Absolute number of reads (abs_nor): minimum number of sequences that need to support an allele; otherwise, the allele is considered an artefact. Ranges between 0 and 1,000, with default = 10.

–. Lowest proportion of reads (low_por): to be called in an individual amplicon, an allele needs to be supported by at least the proportion of reads, within that individual amplicon, that is declared here. Ranges between 0 and 1, with default = 0, while a value greater than 0 is recommended for data sets with ultra-deep sequencing depth, which can suffer more from false positives (Biedrzycka et al., 2017).

Subsequently, putative alleles with very low frequency (both at the individual and population level) are scrutinized again. If the proportion of reads of a putative allele within an individual amplicon is less than 10 times lower than the next higher ranking allele, and if it is very similar (one single different base) to another, more frequent allele present in the same individual amplicon, that putative allele is considered an artefact and removed. Finally, if an individual amplicon has fewer than 50 sequences following all of the allele calling validation steps, it is eliminated. Users are able to change all parameter values but ACACIA recommends settings based on our benchmarking. The output of this step consists of four files:

- **allelereport.csv**: a brief allele report listing genotypes of all individual amplicons, as well as frequencies and abundances of all alleles found in the run;
- **allelereport_XL.csv**: a detailed allele report including the number of reads supporting each allele both within individuals and in the population;
- **pipelinereport.csv**: a pipeline report quantifying read counts and sequences failing or passing each pipeline step described above;
- **alleles.fasta**: a FASTA sequence file of all alleles identified in the run.

To evaluate the pipeline, we calculated both allele calling accuracy and allele calling repeatability. Allele calling accuracy was calculated as the percentage of alleles that have been correctly called across replicates. This was done by comparing the predicted genotype to the genotype generated by ACACIA. All alleles that were dropped out or false positives were marked as inaccurately called alleles. Allele calling repeatability on the other hand was calculated as the percentage of alleles called in both replicates (including false positives). Note here that allele calling accuracy and repeatability are not necessarily correlated if allele calling errors are highly repeatable, that is either false positives or allele dropout are consistent across replicates. However, allele calling accuracy can only be calculated if the genotype is known a priori, which will not be available to most wildlife studies. We have therefore calculated both measures to investigate the pitfalls of relying on allele repeatability to validate a genotyping pipeline.

We investigated the best abs_nor and low_por settings for our data sets by first looking at the allele calling accuracy and repeatability at varying abs_nor values (range: 0–40, with low_por set at 0) first, and at varying low_por values (range: 0–0.02, with the optimal abs_nor, in our case 10) second. The latter is how we recommend users to find their optimal settings, although the range of abs_nor and low_por values to be investigated may vary across different data sets, depending on where the 'peak' optimal setting lies.

The pipeline is supervised by a configuration text file (config.ini), which is appended every time users enter one of the settings mentioned above. Users can avoid running ACACIA interactively (and run the whole workflow in a 'hands-free' mode) by providing a complete config.ini file at the beginning of the workflow. A template of a config.ini file is given in ACACIA's repository (https://gitlab.com/psc_santos/ACACIA/blob/master/config.ini).

2.7 | Data analysis with the AmpliSAS pipeline

To compare how ACACIA performed relative to an existing relevant pipeline, we applied the web server AmpliSAS pipeline to our chicken data sets (Sebastian et al., 2016). The default AmpliSAS parameters of a substitution error rate of 1% and an indel error rate of 0.001% for Illumina data were used. We then tested for the optimal 'minimum dominant frequency' clustering threshold for a given filtering threshold (i.e. 0.5% for the 'minimum amplicon frequency'), by testing a set of thresholds of 10%, 15%, 20% and 25%. All clustering parameters tested gave an allele calling accuracy of 97%, but we chose

the 25% clustering threshold because it was the only parameter that resulted in no false positives.

Subsequently, AmpliSAS filters for clusters that are likely to be artefacts, including chimeras and other low-frequency artefacts that have filtered through the clustering step (Sebastian et al., 2016). The default setting for the filtering of low-frequency variants (i.e. 'minimum amplicon frequency') is 3%. However, this value was far too high for our data sets, and we tested a range of filtering threshold between 0% and 1% at 0.1% intervals (i.e. 0%, 0.1%, 0.2%, etc.). We assessed the optimal filtering threshold using both allele calling accuracy and repeatability.

2.8 | Statistical analyses

To analyse the relationship between the number of alleles amplified and amplification bias/artefacts, we used generalized additive mixed models using the 'MGCV' package (Wood, 2006) in R version 3.6.3 (R Core Team, 2020). The three response variables that were explored using a binomial error distributions corrected for overdispersion (aka quasibinomial) were as follows: proportion of reads assigned to an allele, proportion of reads that were nonchimeric artefacts and proportion of reads that were chimeras. The response variables number of chimeric variant and number of parental variants that were generating chimeric variants were analysed using a Poisson error distribution, with the latter corrected for overdispersion (aka quasi-Poisson). The fixed term number of alleles amplified was entered as a smoother and was limited to 6 estimated degrees of freedom. To control for pseudoreplication, sample ID was entered as a random factor.

3 | RESULTS

3.1 | Sequencing depth for each data set and proportion of artefacts detected using ACACIA

A total of 530,101 paired-end reads were generated for the optimal primer dataset, which amounted to an average of 6,164 reads per amplicon (n = 86). For the cross-species primer data set, 994,338 paired-end reads were generated, amounting to an average of 11,562 reads per amplicon (n = 86). The proportion of artefacts identified at each step of the ACACIA pipeline for the chicken data sets combined is illustrated in Figure 1. Workflow filtering removed the highest proportion of reads when filtering for singletons (13.6%) and chimeras (14.2%). After all filters, 66.4% of the original raw reads were used for allele calling.

3.2 | Optimal settings of different workflows

We compared allele calling repeatability across a range of different abs_nor and low_por settings when using the ACACIA workflow to identify the optimal settings according to genotyping accuracy for our data sets. We first fixed the abs_nor setting at 10 and tested different low_por values and found that the optimal setting (i.e. the highest accuracy values) was 0 across both data sets (Figure 2a.). Setting higher low_por values resulted in a higher allele dropout rate, which led to lower accuracy and repeatability scores. We then tested the optimal abs_nor setting for a fixed low_por value of 0 and found that the optimal setting was 10 across both data sets (Figure 2b). An abs_nor value of 0 increased the rate of false positives, while a value above 10 increased the rate of allele dropout.

For the AmpliSAS workflow, we investigated the optimal filtering threshold and found differing optimal values between data sets. For the optimal primer data set, we found that the optimal filtering threshold was 0.3, while 0.5 was found to be optimal for the cross-species primer data set (Figure 2c).

3.3 | AmpliSAS versus ACACIA: optimal primer data set

When using the optimal settings of the ACACIA workflow, comparison of results with expected genotypes revealed that nine alleles dropped out, no false positives were found (Table 2), and as a result, allele calling accuracy was 98.5% (Figure 2a,b). All instances of allele dropout derived from the B21 haplotype. For two genotypes, both BLB1*21 and BLB2*21 dropped out. For four genotypes, only BLB2*21 dropped out, and for one genotype, only BLB1*21 dropped out (Table 2). Allele calling repeatability was 97.7%.

Using the optimal settings in AmpliSAS across 86 genotypes, a total of 17 alleles dropped out and one false positive was found (Table 2), which resulted in an allele calling accuracy of 97% (Figure 2c). As with ACACIA, most allele dropouts (16 of 17) derived from the B21 haplotype. For three genotypes, both BLB1*21 and BLB2*21 dropped out. For nine genotypes, only BLB1*21 alleles dropped out, and for one genotype, only BLB2*21 allele dropped out. Finally for one genotype, the allele dropout was BLB1*04 and the same genotype had a false-positive allele (Table 2). Allele calling repeatability was 95.3%. Therefore, the ACACIA workflow resulted in higher allele calling accuracy and repeatability than the AmpliSAS workflow.

3.4 | AmpliSAS versus ACACIA: chicken cross-species primer dataset

Using the optimal settings of ACACIA, we found a total of 134 allele dropouts across the 86 genotypes and allele calling accuracy was 77.8% (Figure 2a,b). However, all dropouts were from the alleles BLB1*04, BLB1*15 or BLB1*21, which were never called in the genotypes they were predicted to occur. Further comparison between the allelic reads and the primers revealed two mismatches at the 1st bp and 16th bp within the forward primer. Across the whole data set, only 13 (0.001%), 114 (0.01%) and 11 (0.001%) reads prior to

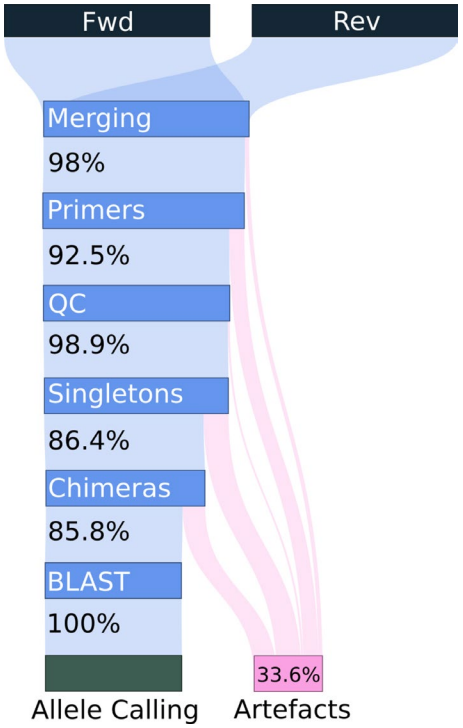


FIGURE 1 Flow diagram of reads and sequences from the two Illumina runs (a first run for the optimal primer data set and the second for the cross-species data set) analysed with ACACIA. Black bars denote the number of initial, raw reads (i.e. 100% of the reads generated by the Illumina runs). Blue bars correspond to filtering steps, and the percentages given correspond to the proportion of sequences from the previous step that were kept for the next stage of the workflow. The percentage given at the bottom right (Artefacts) refers to the total percentage of reads that were filtered from the total initial reads generated by Illumina, prior to any filtering steps. (Fwd & Rev) raw forward and reverse reads; (Merging) paired-end read merger, which includes a first quality filter; (Primers) primer trimming step, which also removes sequences lacking full primers; (QC) quality control; (Singletons) singleton removal; (Chimeras) chimera removal; (BLAST) BLAST filter

applying any downstream quality filtering steps after merging corresponded to BLB1*04, BLB1*15 or BLB1*21, respectively. By comparison, the range for all other alleles was between 25,812 (2.79%) and 115,489 (12.49%) and the range for all artefact reads were between 1 (0.0001%) and 5,535 (0.60%). Therefore, all allele dropouts in the cross-species data set when using the ACACIA workflow are explained by primer mismatch leading to very poor amplification and sequencing of these alleles, which were well within the lower range of artefact reads. Since BLB1*04, BLB1*15 and BLB1*21 dropped out in all genotypes, allele calling repeatability between both replicates was 100% when using the ACACIA workflow, which highlights that relying on allele calling repeatability when validating a genotyping workflow can be misleading.

Using the optimal settings of AmpliSAS, we found 152 allele dropouts across all genotypes and allele calling accuracy was 75% (Figure 2c.). As above, 134 dropouts were due to a mismatch with the forward primer. The remaining 17 alleles that dropped out were BLB1*12 or *19 (13 alleles) and BLB2*14 (4 alleles) (Table 2). Allele calling repeatability between both replicates was 96.1%. Therefore, as with the optimal primer data set the ACACIA workflow resulted in higher allele calling accuracy and repeatability than the AmpliSAS workflow.

3.5 | Relationship between number of alleles amplified and artefacts

In the optimal primer data set, when amplifying two alleles within an amplicon (i.e. within haplotype), all alleles amplified and the proportion of reads assigned to alleles ranged from 0.24 to 0.36 (Figure 3). The latter confirms the suitability of the primer set design for this model system. In contrast, in the cross-species data set, primer mismatch and systematic allele dropout for the alleles BLB1*04, BLB1*15 or BLB1*21 meant that three haplotypes had a single allele instead of two (Figure 3). In both data sets, the contribution of allelic variants to the proportion of reads decreased sharply with increasing number of alleles when amplifying less than 4–6 alleles but starts to level when amplifying more than 4–6 alleles (Figure 3, optimal data set GAMM: $F_{3.477,\ 542} = 237.3$, p-value < .001; cross-species data set GAMM: $F_{4.779,\ 420} = 99.73$, p-value < .001). Amplification efficiency was significantly different between alleles in both data sets (optimal data set GAMM: $F_{12,\ 542} = 10.63$, p-value < .001; cross-species data set GAMM: $F_{9,\ 420} = 35.53$, p-value < .001). Both alleles from the B4 and B21 haplotypes in the optimal data set and the BLB2*04 allele in the cross-species primer data set consistently amplified poorly when coamplifying with alleles from other haplotypes (Figure 3; see Figure S1 for multiple comparison post hoc test of allele amplification). In the optimal primer data set, the low amplification efficiency of the B21 haplotype when coamplifying with other haplotypes explains the high allele dropout of alleles from this haplotype in more complex genotypes (i.e. when coamplifying 10 or more alleles; Figure 3). In contrast, the higher sequencing depth of the cross-species data set meant that BLB2*04 allele did not drop out. However, we identified a primer mismatch between BLB2*04 allele and the second base pair of the reverse primer, explaining the lower amplification efficiency of this allele when coamplified with other alleles.

The proportion of sequences classified as artefacts was much higher for PCRs using the optimal primer set than when using the cross-species primer set (Figure 4a,b; nonchimeric artefact GAMM: $F_{1,\ 74} = 2{,}669.1$, p-value < .001; chimera: $F_{1,\ 74} = 180.4$, p-value < .001), which is likely due to the fact that the fragment length of the optimal primer data set was longer relative to the cross-species primer data set (241 bp versus 151 bp, respectively; see Discussion). For both data sets in this study, when considering nonchimeric artefacts, there was a positive relationship between the proportion of artefacts and the number of alleles amplified

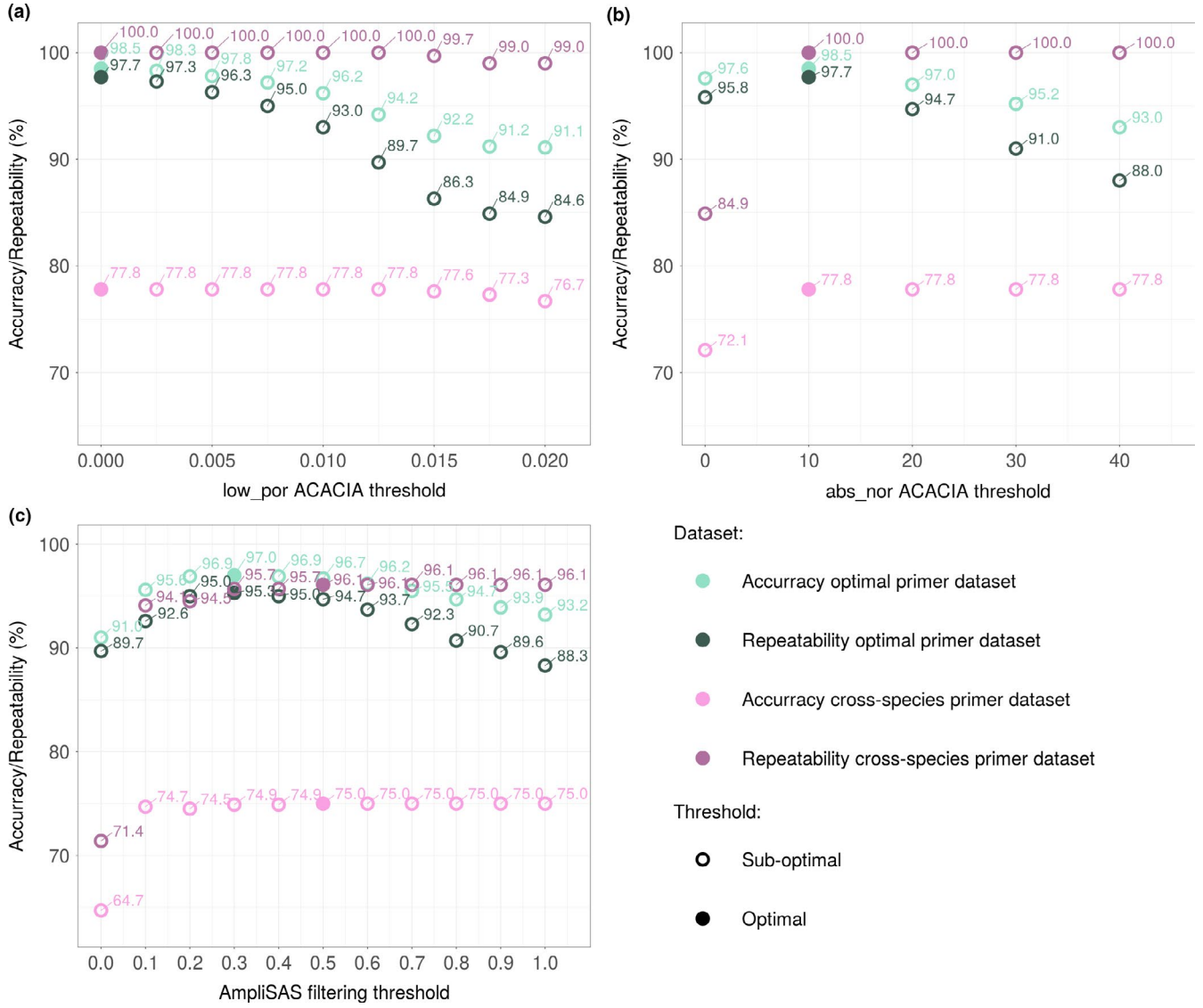


FIGURE 2 Allele calling accuracy and repeatability for the two data sets of this study (optimal primers or cross-species primers) at different low_por threshold settings with abs_nor set at 0 within the ACACIA pipeline (a); at different low_por threshold settings with abs_nor set at 0 within the ACACIA pipeline (b); and at different filtering thresholds (i.e. 'minimum amplicon frequency') within the AmpliSAS pipeline (c)

(Figure 4a.; GAMM: $F_{1,74} = 207.3$, p-value < .001). There is a logarithmic relationship between the proportion of chimeric artefacts and the number of alleles amplified whereby the proportion of chimeric reads no longer increased with number of alleles amplified when amplifying more than 4–6 alleles (Figure 4b, GAMM: $F_{4.857,74} = 35.77$, p-value < .001). The total number of unique chimeric reads also tended to follow a logarithmic relationship, whereby the number of unique chimeric variants seemed to no longer increase with the number of alleles amplified when amplifying more than 10 alleles (Figure 3c, GAMM: $F_{4.06,74} = 117.5$, p-value < .001). The relationship between the total number of parental variants generating chimeras and the number of alleles amplified also levelled when amplifying more than six alleles (Figure 4d.; GAMM: $F_{4.06,\ 74} = 117.5$, p-value < .001).

4 | DISCUSSION

Using known MHC genotypes for two data sets (chicken MHC Class IIB complex), we achieved higher allele calling accuracy (≥98.5%) and repeatability (≥97.7%) using ACACIA for the optimal primer data set. With fewer allele dropouts and false positives, the ACACIA pipeline performed better than AmpliSAS. We demonstrated the 'costs' of designing primers within MHC exon 2 in terms of allele dropout, with three common alleles failing to amplify when using primers designed from sequences of related Galliform species. We also explored the relationship between artefacts and the number of alleles amplified per amplicon, and, as expected, found heterogeneous amplification efficiency of allelic variants when amplifying multiple loci within a PCR. Surprisingly, the relationship between the proportion

TABLE 2 Specific genotypes within replicates that had a genotyping error using ACACIA and AmpliSAS genotyping workflows (excluding allele dropout due to primer mismatch in the cross-species primer data set). Genotypes, replicates (Rep.), predicted number of alleles (\# Pred.All.), allele dropouts (Dropout) and false positives (F.P.) using ACACIA and AmpliSAS are shown

Genotype	Rep.	# Pred.All.	Dropout ACACIA	Dropout AmpliSAS	F.P. AmpliSAS
a. Optimal primers dataset (BLB MHC class IIB)					
B2-B4-B12-B14-B19-B21	1	11	BLB2*21	BLB2*21	
				BLB1*21	
B4-B14-B15-B19-B21	1	10	BLB2*21	BLB1*21	
	2	10	BLB2*21	BLB2*21	
B4-B15-B19-B21	1	8	BLB2*21	BLB1*21	
B2-B4-B12-B14-B15-B19-B21	1	13	BLB2*21 BLB1*21	BLB2*21 BLB1*21	
B2-B4-B12-B14-B15-B21	1	12	BLB2*21 BLB1*21	BLB2*21 BLB1*21	
B2-B12-B14-B15-B19-B21	1	11	BLB1*21		
B2-B4-B12-B15-B19-B21	1	11		BLB1*21	
B2-B4-B12-B15-B21	1	10		BLB1*21	
B2-B4-B14-B15-B19-B21	1	12		BLB1*21	
B2-B4-B14-B15-B21	1	10		BLB1*21	
B2-B4-B15-B19-B21	1	10		BLB1*21	
	2	10		BLB1*21	
B4-B12-B21	1	6		BLB1*04	1
B4-B14-B15-B19-B21	2	10		BLB1*21	
b. Cross-species primer data set (BLB MHC class IIB)					
B12-B14-B15-B21	1	5		BLB1*12 or *19	
	2	5		BLB1*12 or *19	
B14-B15-B19-B21	1	8		BLB1*12 or *19	
B2-B12-B14-B15	1	6		BLB1*12 or *19	
	2	6		BLB1*12 or *19	
B2-B12-B14-B15-B19-B21	1	11		BLB2*14	
B2-B14-B15-B19-B21	1	10		BLB1*12 or *19	
B2-B4-B12-B14-B15	1	10		BLB1*12 or *19	
	2	10		BLB1*12 or *19	
B2-B4-B12-B14-B15-B19	1	11		BLB2*14	
B2-B4-B12-B14-B15-B19-B21	1	13		BLB2*14	
B2-B4-B12-B14-B15-B21	1	12		BLB1*12 or *19	
B2-B4-B12-B14-B19-B21	1	11		BLB2*14	
B2-B4-B14-B15-B19-B21	1	12		BLB1*12 or *19	
B4-B12-B14-B15	1	8		BLB1*12 or *19	
	2	8		BLB1*12 or *19	
B4-B14-B15-B19-B21	1	10		BLB1*12 or *19	

of chimeric artefacts and number of alleles amplified was not linear but rather levelled when amplifying more than 4–6 alleles. However, nonchimeric artefacts did increase linearly with increasing number of alleles amplified. Below we discuss in further detail ACACIA, AmpliSAS, primer design for nonmodel organisms, the relationship between the number of alleles amplified and artefacts, and the effect of chimera formation on genotyping pipelines, and finally, we conclude by advising users on important points to consider when genotyping complex multigene families in nonmodel organisms.

4.1 | AmpliSAS versus ACACIA

Experimentally generating variation in the number alleles amplified with amplicon of known chicken MHC class IIB genotypes allowed us to validate our ACACIA pipeline to genotype systems with high complexity at high accuracy and repeatability across replicates in the optimal primer data set. While we achieved higher allele calling accuracy and repeatability using ACACIA than the AmpliSAS web server pipeline, we do not claim that ACACIA will necessarily perform

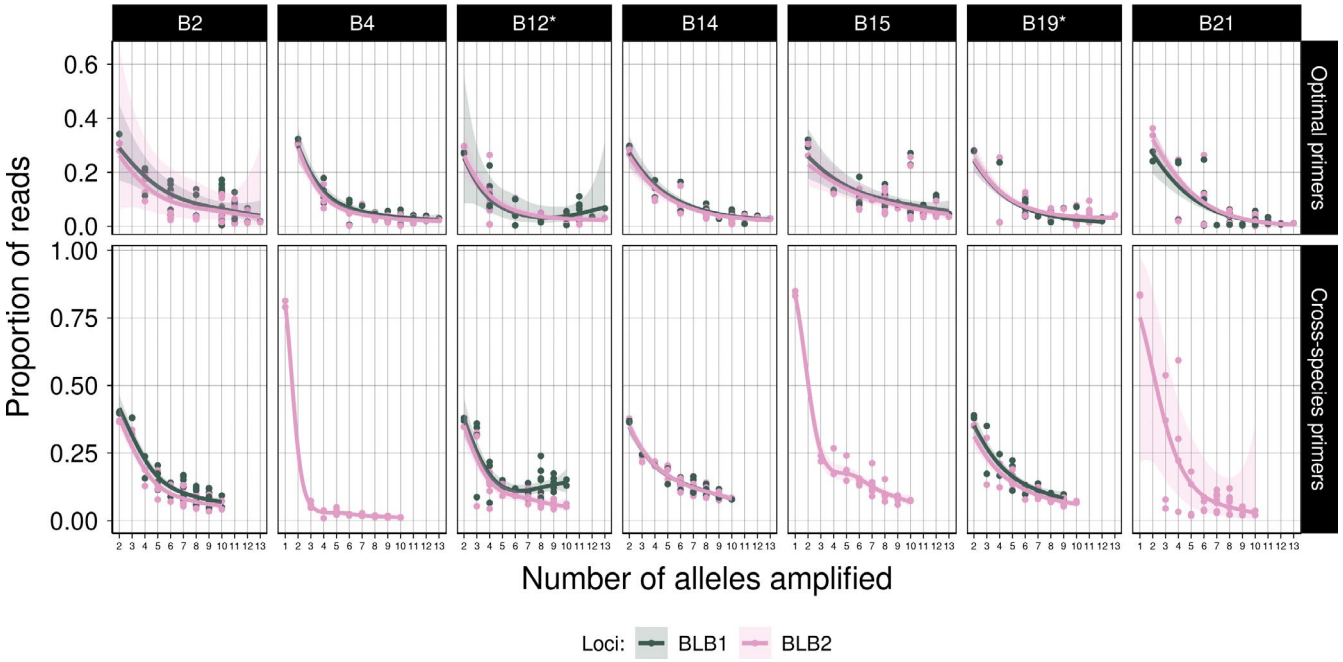
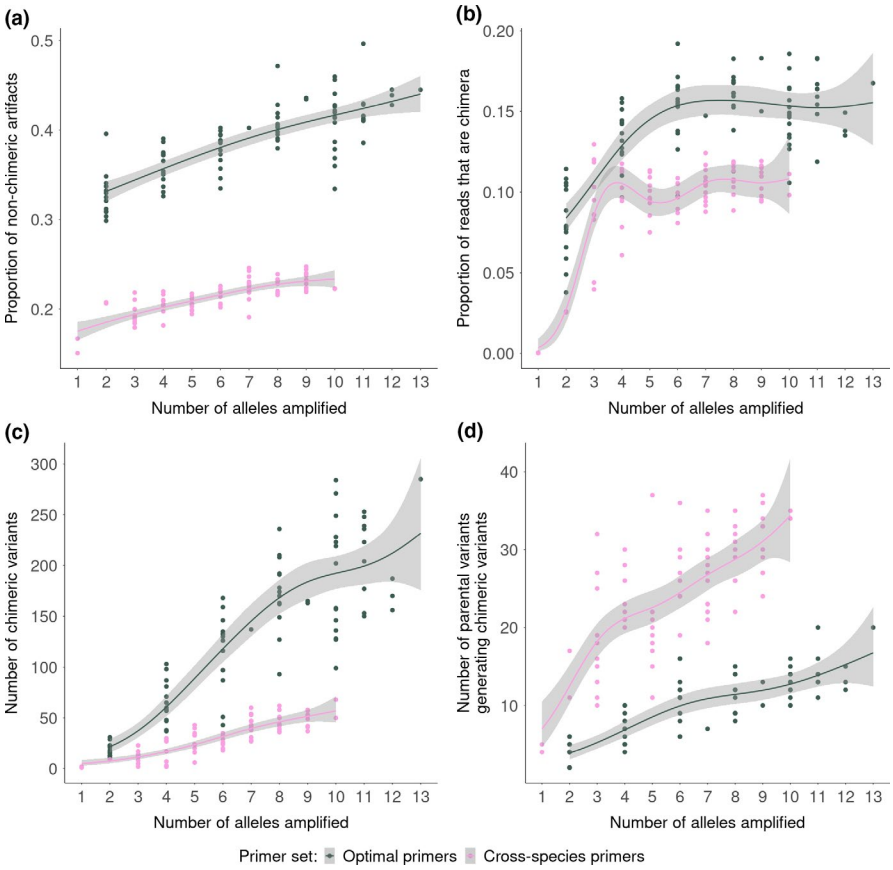


FIGURE 3 The relationship between the number of alleles amplified within a PCR and the proportion of reads assigned to alleles for each haplotype and locus. *Note that haplotype B12 and B19 have the same BLB01 allele, therefore for presentation purposes only, when both haplotypes were within the same genotype, BLB01*12 or *19 was assigned to haplotype B12 to avoid pseudoreplication. BLB1*04, BLB1*15 and BLB1*21 failed to amplify due to primer mismatch in the cross-species data set

FIGURE 4 The relationship between the number of alleles amplified and the proportion of reads that are nonchimeric artefacts (a); the proportion of chimeric reads (b); the absolute number of chimeric variants (c); and the absolute number of parental variants generating chimeric reads (d)



better than AmpliSAS with all data sets. To demonstrate the latter, we would need to test both pipelines on a larger number of data sets and/or on simulated data sets. In addition, while our pipeline should suit data generated with any high-throughput sequencing technologies, we have only tested ACACIA with paired-end Illumina sequencing technology. Here, we focus on ACACIA versus AmpliSAS;

however, further discussion on how ACACIA compares with other genotyping workflows can be found in the supplementary material.

Biedrzycka et al. (2017) reviewed extensively different genotyping workflows according to sequencing depth using AmpliSAS in highly complex MHC system (up to 42 alleles within an individual). The authors achieved an allele calling repeatability <90% when coverage <5,000 sequences regardless of the genotyping workflow used and reached 99% with a sequencing depth of 20,000. While our study does not allow to assess the relationship between the number of alleles amplified within amplicon and sequencing depth using ACACIA, our results were consistent with Biedrzycka et al. (2017) since allele calling repeatability was 97.7% for the optimal primer data set and 100% for the cross-species primer dataset, which had an average sequencing depth of 6,164 and 11,562 reads per amplicon, respectively. For the optimal primer data set, regardless of the genotyping pipeline used, allele dropout occurred in genotypes with a high number of alleles within amplicon (for ACACIA, 8 out of 9, and for AmpliSAS, 12 out of 14 genotypes with allele dropouts had 10 alleles or more). Our optimal primers amplified all alleles at a similar efficiency when amplified within single haplotypes suggesting that the primers are indeed optimally designed. For all instances, allele dropouts were alleles from the B4 and B21 haplotype that amplified poorly when coamplified with alleles from other haplotypes. Higher sequencing depth will reduce or even remove such allele dropout instances (Biedrzycka et al., 2017). Indeed for the cross-species primer data set, sequencing depth was nearly twice as high, and there were no instances of allele dropout due to the ACACIA pipeline (all allele dropouts were due to primer mismatch; see subsequent subsection of the discussion) and allele calling repeatability was 100%. Therefore, in order to reach allele calling repeatability values <99%, we advise researchers to aim for a sequencing depth of at least 10,000 reads per amplicon when amplifying more than 4 alleles per amplicon and of 20,000 reads when amplifying more than 15 alleles regardless of the bioinformatic workflow used (Biedrzycka et al., 2017).

The most apparent benefit of using the AMPLISAS web server is that it is relatively easy to use for users with limited knowledge of scripting languages (such as PYTHON, PERL, C++ or R). However, we have noticed that a number of studies report results using default settings when applying the AMPLISAS pipeline to their data set. We find this concerning since, as our study demonstrates, the default clustering and filtering parameters are unlikely to be optimal for most data sets. Indeed, allele calling accuracy was much lower when using the default settings (81.8%) as compared to the optimal settings (97%) in the optimal primer data set in our study, due to high allele dropout. We therefore strongly discourage users from using default settings and advise to permutate between different filtering and clustering parameters to find the best settings for their data set when using the AMPLISAS pipeline. As most wildlife studies cannot assess allele calling accuracy, duplicating samples and relying on repeatability is the only feasible method for most research to optimize their amplicon-based genotyping workflow. However, authors should bear in mind that due to the high recurrence of amplification and sequencing errors, high repeatability in allele calling between replicates does not necessarily entail an error-free workflow and may be misleading. Therefore, careful design of primers and PCR conditions to reduce artefacts during amplification is crucial to maximize amplicon-based genotyping accuracy regardless of the bioinformatic tools used (see further discussion on this below).

An important disadvantage of the AmpliSAS web server is that at the time of writing, sequencing depth per amplicon was limited to 5,000 reads. For data sets with sequencing depth above 5,000 reads, AmpliSAS can be run locally but we found that, unlike the web server, the local version of AmpliSAS had limited documentation and troubleshooting was time-consuming. Once installed, ACACIA does not require users to have experience with scripting languages, allows genotyping with virtually unlimited sequencing depth and provides output data reporting the number of reads kept at each step of the pipeline. The latter should aid users when deciding upon optimal parameters and thresholds. As for the AmpliSAS pipeline, we advise to not use default parameters of ACACIA without critically assessing different parameters for each data set. In particular, we urge users to permutate between different settings of abs_nor and low_por parameters. We advise to first search for the optimal abs_nor setting with a fixed low_por parameter of 0 because it is likely that it is only necessary to change the low_por parameter setting from 0 in data sets with ultra-deep sequencing depth. If it is subsequently found that the optimal low_por setting is greater than 0, users should repeat the permuting step of abs_nor until the optimal settings are found. Of course, finding optimal settings requires the inclusion of replicates for at least a subset of the data set. We therefore recommend that a sufficient number of replicates are always included in genotyping runs to obtain sufficiently accurate repeatability values.

4.2 | The challenge of designing primers for nonmodel organisms

A common approach for primer design in complex genomic regions of nonmodel organisms includes aligning multiple sequences of phylogenetically related species. By building primers on consensus sequences, researchers assume that oligos will also amplify the target region in the species of interest. However, knowledge about related species is often limited to very few individuals. This means that primers can be designed in regions that are polymorphic in the target species. As a consequence, certain allelic variants are not amplified and homozygosity is overestimated. Indeed, this proved to be the case in our cross-species primer data set, whereby two mismatches (1st bp and 16th bp) within the forward primer (19 bp long) were sufficient to prevent the amplification of three alleles (out of 13). Interestingly, a single base pair mismatch between the second base pair of the reverse primer and the BLB2*04 allele did not prevent the amplification of this allele, although it did suffer severely from low amplification efficiency when in competition with other alleles. High sequencing depth for the cross-species primer data set prevented this allele from dropping out, regardless of the genotyping pipeline

used. Our study therefore highlights the importance of carefully designed primers for amplicon-based genotyping.

Two nonmutually exclusive strategies can be used to decrease allele dropout in nonmodel organism with no a priori information on the target region. First, designing multiple primers and combining them within a PCR reaction are known to reduce allele dropout due to primer mismatch and allele amplification bias (Marmesat et al., 2016). In addition, combining multiple primers within PCRs reduces the need to sequence multiple primer sets separately, considerably reducing the cost of using multiple primer sets to genotype a novel target region. Second, isolating introns and exons flanking the highly polymorphic exons of interest can help inform primer design and even result in locus-specific primers (Babik et al., 2008, 2009; Biedrzycka & Radwan, 2008; Burri et al., 2008, 2014; Gaigher et al., 2016; Gillingham et al., 2016), which reduces allele dropout due to primer mismatch and allows assigning alleles to loci. Indeed, an important set-back of multilocus amplification is that assigning alleles to loci in complex systems is frequently impossible, even when using recent phasing algorithms (Huang et al., 2019), limiting the use of many population genetic analyses (Gillingham et al., 2017). In addition, the total number of loci is currently likely to be underestimated in many species since recent gene duplication means that different genes often carry identical alleles (Gaigher et al., 2016; Worley et al., 2008).

The characterization of flanking regions have typically relied on cloning followed by Sanger sequencing (Burri et al., 2008, 2014; Gaigher et al., 2016; Gillingham et al., 2016) or genome walking techniques such as vectorette PCRs (Babik et al., 2008, 2009; Biedrzycka & Radwan, 2008). Both of these approaches are time-consuming and thus only allow the characterization of a limited number of individuals, which may underrepresent sequence diversity during primer design. The development of long-read HTS of single molecules, such as Pacific Biosciences single-molecule real-time sequencing or Oxford Nanopore sequencing, offers much promise in characterizing longer regions of novel multigene families and consequently to enable more informed primer design (O'Connor et al., 2019). For instance, Fuselli et al. (2018) were recently able to develop an assembly pipeline that combined long-read HTS from a single sample with de novo short-read HTS assembly of six samples to characterize a 9Kb region of the MHC Class II (*DRB*) locus in Alpine chamois *Rupicapra rupicapra*. More complex multigene families have also been characterized in primate species (Larsen et al., 2014), including the MHC (Gordon et al., 2016; Westbrook et al., 2015), using long-read HTS. An important limitation with long-read HTS technologies is that they have higher sequencing error rates than short-read HTS (the reported error rates are 16% per base compared with 0.1% for Illumina HTS) and lower sequencing output (5–20 Gb/run compared with 200–600 for Illumina HTS) (Bansal & Boucher, 2019; Reinert et al., 2015). Therefore, while long-read HTS will undoubtedly improve our understanding of the MHC and other multigene complexes structure and consequently nonmodel species primer design, population studies genotyping a large number of individuals are likely to continue to rely on amplicon-based genotyping from short-read HTS for the foreseeable future.

4.3 | Relationship between number of alleles amplified and artefacts

By knowing the exact expected alleles for the chicken genotypes, we were able to quantify chimeric artefacts precisely (Figure 1). There were a higher proportion of chimeric and nonchimeric artefacts in the optimal primer data set than in the cross-species primer data set. The most likely explanation for the latter is the shorter sequence for the cross-species primer data set (151 bp) compared with the optimal primer data set (241 bp). A shorter fragment reduces the number of base pairs that can be erroneously substituted/deleted and the number of breaking points for chimera formation. In addition, it is likely that the probability of incomplete elongation is inversely related to fragment length. Thus, fragment length appears to be the dominant factor predicting the proportion of artefactual reads.

As expected, the proportion of reads that were nonchimeric artefacts increased linearly as the number of alleles amplified with an amplicon increased, which can be explained simply by the fact that there are an increasing number of possible artefacts that can be generated as the number of initial template variants increases. An unexpected result was that the proportions of chimeras did not increase with increasing number of alleles amplified with an amplicon, when amplifying more than 4–6 alleles. Similarly, when amplifying more than 10 alleles, the number of chimeric variants no longer increased with increasing number of alleles amplified within an amplicon. Such saturation in chimera generation beyond a threshold of alleles amplified is likely to be a by-product of allele PCR competition. Indeed, as demonstrated by our own data, there is amplification bias whereby some gene variants are amplified preferentially relative to others (Marmesat et al., 2016; Sommer et al., 2013). Therefore, a few gene variants (~3–6 gene variants) are preferentially amplified and most chimeras originate from these dominantly amplified variants and few chimeras are generated from the poorly amplified variants. Indeed, we found that the number of parental variants generating chimeras in our data set did not increase with increasing number of alleles amplified when amplifying more than 4–6 alleles. The nonlinear relationship between chimera generation and number of alleles amplified has important implications when considering sequencing depth needed to accurately genotype complex multigene families, since it suggests that linearly increasing sequencing depth for increasing number of alleles coamplified is not necessarily the optimal strategy. The challenges of dealing with chimeras in genotyping pipelines are discussed below in detail.

4.4 | Chimeras in genotyping pipelines

The formation of artificial chimeras during amplification is an important source of artefacts in amplicon sequencing projects (Lenz & Becker, 2008; Smyth et al., 2010), including those with newer sequencing technologies (Laver et al., 2016). Chimeras are challenging to identify as artefacts because they resemble real alleles generated by recombination, particularly in multigene families under high rates of interlocus genetic exchange ('concerted evolution'), which is

common in many MHC systems (Burri et al., 2008, 2010; Edwards et al., 1995; Gillingham et al., 2016; Hess & Edwards, 2002; Wittzell et al., 1999). Our results suggest that chimeras are more prevalent, harder to identify and potentially more reproducible across technical replicates than previously assumed. We expect the same to be true for similar projects with conserved, yet variable amplification targets such as the MHC.

One allele erroneously called as a real variant (i.e. a false positive) by the AmpliSAS pipeline in the optimal primer data set was actually a chimera between the BLB1*21 and BLB2*21 alleles. Furthermore, when using the AmpliSAS pipeline, 15 allele dropouts in the cross-species primer dataset were due to erroneous assignment of real allelic variants as chimera artefacts. Indeed, the BLB1*12 or *19 allele was identical to potential chimeric artefact sequences between BLB2*14 (85 possible breakpoints) and any of the following alleles: BLB1*04, BLB2*15, BLB2*19, BLB2*21 or BLB1*21 (Figure 5a.). In addition, BLB2*14 dropped out because it is identical to a chimera formed between the BLB1*02 and BLB1*12 or *19 alleles (33 breakpoints; Figure 5b.).

We have identified two factors that seemed to enhance chimera formation and challenge the distinction between artefact and real allelic variants. First, the combination of multiple real 'parent' sequences can yield the same chimeras, as illustrated in our examples in Figure 5a and Figure 5b, whereby any breakpoint in the shaded areas leads to the same chimeras. Second, peripheral breakpoints (Figure 5c) can generate chimeras that differ to parental sequences by as little as a single base pair. For instance, a chimera could be a product of the allele BLB2*21 combined with any of the other alleles shown in the alignment, with a breakpoint within the shaded area (Figure 5c.). Since the potential breakpoints are at the very end of the sequence, the chimera is very similar to one of its parents (in this example, it is different from BLB2*21 by only one base). In an attempt to deal with this issue as much as possible, we changed the default settings of VSEARCH so that chimeras can be detected even if they differ from one parent by one single base. Both the 'multiple parents' and the 'peripheral breakpoints' issues are likely to contribute to making chimeras reproducible across replicates.

Our study highlights the challenges of chimeras for amplicon-based genotyping. We purposefully used conventional PCR conditions to replicate methods used by most wildlife MHC studies. However, the formation of most chimeras is known to occur during the final cycles of PCR amplification when dNTP and primer concentrations are low and when incompletely elongated sequences are high (Judo et al., 1998; Lenz & Becker, 2008; Smyth et al., 2010). When target primers and dTNPs concentration are low during the latter stages of PCR cycles, incompletely elongated sequences act as primers and bind with the wrong sequences generating chimeras. Chimera formation during amplification can be simply reduced by adjusting the ratio of DNA template to dNTP and primer concentrations, reducing the number of cycles, increasing the extension step within PCR cycles and omitting the final extension step (which elongates high concentrations of incomplete chimeric sequences; Judo et al., 1998; Lenz & Becker, 2008; Smyth et al., 2010). Therefore, prior to any amplicon-based genotyping study, we advise researchers to reduce artefacts, including chimeras, during the wet laboratory stage of their workflow by applying carefully designed and optimal PCR conditions. Practices during the wet laboratory that reduce artefact generation in the first place are likely to be the most effective way of reducing genotyping errors regardless of the bioinformatic allele calling workflow used.

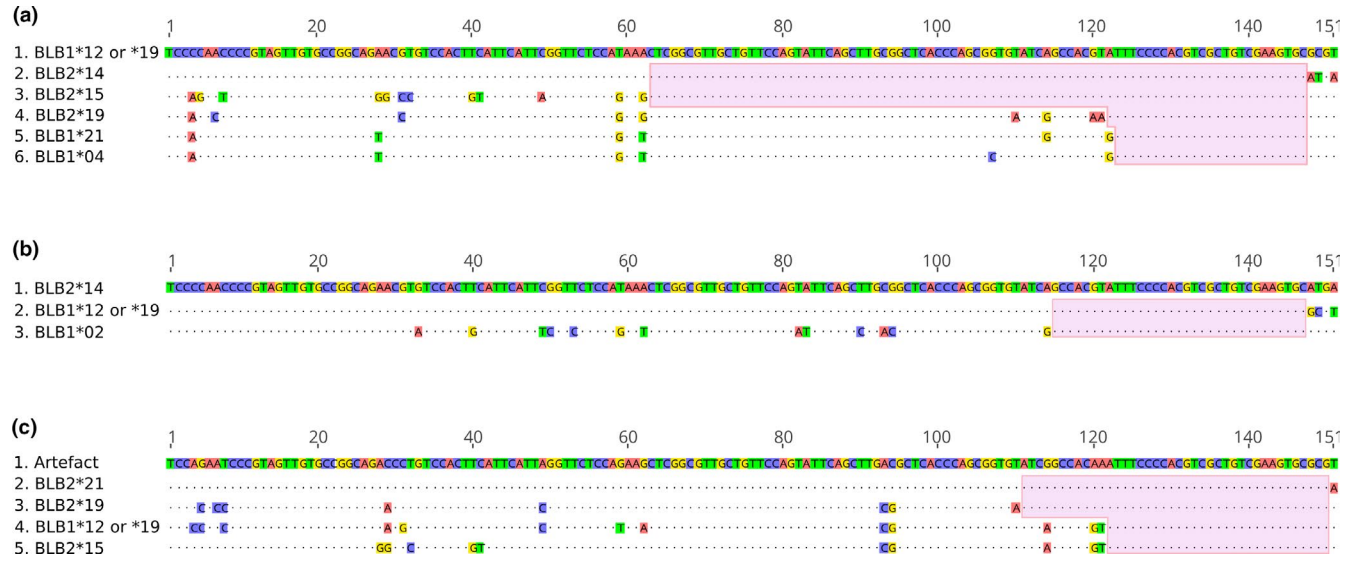


FIGURE 5 Three alignments with examples of sequences that can be classified as chimeras. The points denote identity to the first sequence in each alignment, while the differences to it are highlighted. The shaded areas indicate possible chimera-yielding breakpoints. (a) The allele BLB1*12 or *19 could be a chimera of BLB2*14 with any of the four other allele sequences depicted, in a case of multiple potential parent pairs. (b) BLB2*14 can be interpreted as a chimera between BLB1*12 or *19 minor and BLB1*02. (c) Actual chimera with multiple potential parents and a peripheral breakpoint, and therefore very similar to one of its parents

5 | CONCLUSION

We have demonstrated that the ACACIA genotyping pipeline provides high allele calling accuracy and repeatability. Regardless of the pipeline used, however, users should critically assess the optimal parameters to be used concerning both the wet laboratory and bioinformatic pipelines. We are convinced that universal default settings for optimal genotyping accuracy cannot be achieved, since optimal parameters will depend on data set-specific generation of artefacts. The latter, in turn, varies according to species-specific complexity (number of alleles coamplified), DNA quality and the conditions of PCR (e.g. extension time, number of cycles and the polymerase used) and sequencing (e.g. quality and depth). High sequencing depth allows detecting alleles that amplify poorly in complex (multigene) systems. Furthermore, simple steps prior to sequencing can greatly reduce the number of artefacts generated and improve genotyping accuracy: designing more than one PCR primer pair, reducing the number of PCR cycles, increasing PCR in-cycle extension time and omitting the final extension step. Reducing chimera formation during PCRs is particularly critical, because they are difficult to distinguish from real alleles generated by interlocus recombination.

ACKNOWLEDGEMENTS

MG was supported by a DFG grant (DFG Gi 1065/2-1). We are very grateful to Jim Kaufman and his laboratory members for providing the chicken DNA samples used in this study and for his comments on a previous version of this work. Version 3 of this preprint has been peer-reviewed and recommended by Peer Community In Evolutionary Biology (https://doi.org/10.24072/pci.evolbiol.100092). We thank the PCI reviewers Thomas Bigot, Helena Westerdahl and Sebastian Ernesto Ramos-Onsins, the PCI recommender François Rousset and two anonymous reviewers for their comments that improved our manuscript. Open access funding enabled and organized by ProjektDEAL.

CONFLICT OF INTEREST

The authors of this preprint declare that they have no financial conflict of interest with the content of this article. François Rousset is the recommender for PCI Evolutionary Biology.

AUTHOR CONTRIBUTIONS

M.G. and P.S. conceived the study. P.S. wrote ACACIA. M.G. did the data analysis in R. M.G., P.S. and B.K.M. ran the allele calling workflows. B.K.M. did the AmpliSAS analysis. K.W. participated in and supervised the laboratory work. K.G. did the laboratory work. S.S. instigated the study and heads the laboratory where the work was carried out. M.G., B.K.M. and P.S. wrote the first draft of the paper. All authors commented and approved subsequent versions.

DATA AVAILABILITY STATEMENT

Raw sequences of all data sets, example input files, suggested settings and the source code at the time of this publication are available at FIGSHARE (https://figshare.com/projects/ACACIA/66485 and 10.6084/m9.figshare.9952520). ACACIA is freely available on the GitLab at https://gitlab.com/psc_santos/ACACIA (this paper's code is available as a snapshot tagged as V1.0, https://gitlab.com/psc_santos/ACACIA/-/tags/V1.0), under an MIT licence.

ORCID

Mark A. F. Gillingham https://orcid.org/0000-0002-7935-9539
B. Karina Montero https://orcid.org/0000-0003-4246-6004
Simone Sommer https://orcid.org/0000-0002-5148-8136
Pablo S. C. Santos https://orcid.org/0000-0002-3008-014X

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, *2*(2), e00191–e00216. https://doi.org/10.1128/mSystems.00191-16

Babik, W. (2010). Methods for MHC Genotyping in Non-Model Vertebrates. *Molecular Ecology Resources*, *10*(2), 237–251. https://doi.org/10.1111/j.1755-0998.2009.02788.x

Babik, W., Pabijan, M., Arntzen, J. W., Cogalniceanu, D., Durka, W., & Radwan, J. (2009). Long-term survival of a urodele amphibian despite depleted major histocompatibility complex variation. *Molecular Ecology*, *18*(5), 769–781. https://doi.org/10.1111/j.1365-294X.2008.04057.x

Babik, W., Pabijan, M., & Radwan, J. (2008). Contrasting patterns of variation in MHC loci in the alpine newt. *Molecular Ecology*, *17*(10), 2339–2355. https://doi.org/10.1111/j.1365-294X.2008.03757.x

Bansal, V., & Boucher, C.. (2019). Sequencing technologies and analyses: where have we been and where are we going? *iScience*, *18*, 37–41. https://doi.org/10.1016/j.isci.2019.06.035

Biedrzycka, A., & Radwan, J. (2008). Population fragmentation and major histocompatibility complex variation in the spotted suslik, *Spermophilus Suslicus*. *Molecular Ecology*, *17*(22), 4801–4811. https://doi.org/10.1111/j.1365-294X.2008.03955.x

Biedrzycka, A., Sebastian, A., Migalska, M., Westerdahl, H., & Radwan, J. (2017). Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. *Molecular Ecology Resources*, *17*(4), 642–655. https://doi.org/10.1111/1755-0998.12612

Burri, R., Hirzel, H. N., Salamin, N., Roulin, A., & Fumagalli, L. (2008). Evolutionary patterns of MHC class II B in owls and their implications for the understanding of avian MHC evolution. *Molecular Biology and Evolution*, *25*(6), 1180–1191. https://doi.org/10.1093/molbev/msn065

Burri, R., Promerová, M., Goebel, J., & Fumagalli, L. (2014). PCR-based isolation of multigene families: lessons from the avian MHC class IIB. *Molecular Ecology Resources*, *14*(4), 778–788. https://doi.org/10.1111/1755-0998.12234

Burri, R., Salamin, N., Studer, R. A., Roulin, A., & Fumagalli, L. (2010). Adaptive divergence of ancient gene duplicates in the avian MHC class II β. *Molecular Biology and Evolution*, *27*(10), 2360–2374. https://doi.org/10.1093/molbev/msq120

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: Mechanisms, evolution and human disease. *Nature Reviews Genetics*, *8*(10), 762–775. https://doi.org/10.1038/nrg2193

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Edwards, S. V., Grahn, M., & Potts, W. K. (1995). Dynamics of Mhc evolution in birds and crocodilians: amplification of class II genes with degenerate primers. *Molecular Ecology*, *4*, 719–729. https://doi.org/10.1111/j.1365-294X.1995.tb00272.x

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, *4*(12), 1111–1119. https://doi.org/10.1111/2041-210X.12114

Fuselli, S., Baptista, R. P., Panziera, A., Magi, A., Guglielmi, S., Tonin, R., Benazzo, A., Bauzer, L. G., Mazzoni, C. J., & Bertorelle, G. (2018). A new hybrid approach for MHC genotyping: high-throughput NGS and long read MinION nanopore sequencing, with application to the non-model vertebrate alpine chamois (*Rupicapra Rupicapra*). *Heredity*, *121*(4), 293–303. https://doi.org/10.1038/s41437-018-0070-5

Gaigher, A., Burri, R., Gharib, W. H., Taberlet, P., Roulin, A., & Fumagalli, L. (2016). Family-assisted inference of the genetic architecture of major histocompatibility complex variation. *Molecular Ecology Resources*, *16*(6), 1353–1364. https://doi.org/10.1111/1755-0998.12537

Galan, M., Guivier, E., Caraux, G., Charbonnel, N., & Cosson, J.-F. (2010). A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, *11*(1), 296. https://doi.org/10.1186/1471-2164-11-296

Gillingham, M. A. F., Béchet, A., Courtiol, A., Rendón-Martos, M., Amat, J. A., Samraoui, B., Onmuş, O., Sommer, S., & Cézilly, F. (2017). Very high MHC Class IIB diversity without spatial differentiation in the mediterranean population of greater flamingos. *BMC Evolutionary Biology*, *17*, 56. https://doi.org/10.1186/s12862-017-0905-3

Gillingham, M. A. F., Courtiol, A., Teixeira, M., Galan, M., Bechet, A., & Cezilly, F. (2016). Evidence of gene orthology and trans-species polymorphism, but not of parallel evolution, despite high levels of concerted evolution in the major histocompatibility complex of flamingo species. *Journal of Evolutionary Biology*, *29*(2), 438–454. https://doi.org/10.1111/jeb.12798

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759–769. https://doi.org/10.1111/j.1755-0998.2011.03024.x

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., Malig, M., Raja, A., Fiddes, I., Hillier, L. W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R. K., Haussler, D., Chin, C.-S., & Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science*, *352*(6281), aae0344. https://doi.org/10.1126/science.aae0344

Goto, R. M., Afanassieff, M., Ha, J., Iglesias, G. M., Ewald, S. J., Briles, W. E., & Miller, M. M. (2002). Single-strand conformation polymorphism (SSCP) assays for major histocompatibility complex B genotyping in chickens. *Poultry Science*, *81*(12), 1832–1841. https://doi.org/10.1093/ps/81.12.1832

Hess, C. M., & Edwards, S. V. (2002). The evolution of the major histocompatibility complex in birds. *BioScience*, *52*(5), 423–431.

Huang, K., Zhang, P., Dunn, D. W., Wang, T., Mi, R., & Li, B. (2019). Assigning alleles to different loci in amplifications of duplicated loci. *Molecular Ecology Resources*, *19*(5), 1240–1253. https://doi.org/10.1111/1755-0998.13036

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, *8*(7), R143. https://doi.org/10.1186/gb-2007-8-7-r143

Jacob, J. P., Milne, S., Beck, S., & Kaufman, J. (2000). The major and a minor class II β-chain (B-LB) gene flank the tapasin gene in the B-F /B-L region of the chicken major histocompatibility complex. *Immunogenetics*, *51*(2), 138–147. https://doi.org/10.1007/s002510050022

Judo, M. S. B., Wedel, A. B., & Wilson, C. (1998). Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Research*, *26*(7), 1819–1825. https://doi.org/10.1093/nar/26.7.1819

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kaufman, J., Jacob, J., Shaw, I., Walker, B., Milne, S., Beck, S., & Salomonsen, J. (1999). Gene organisation determines evolution of function in the chicken MHC. *Immunological Reviews*, *167*(February), 101–117. https://doi.org/10.1111/j.1600-065X.1999.tb01385.x

Kaufman, J., Milne, S., Göbel, T. W. F., Walker, B. A., Jacob, J. P., Auffray, C., Zoorob, R., & Beck, S. (1999). The chicken B locus is a minimal essential major histocompatibility complex. *Nature*, *401*(6756), 923–925. https://doi.org/10.1038/44856

Kaufman, J., Völk, H., & Wallny, H.-J. (1995). A 'Minimal Essential Mhc' and an 'Unrecognized Mhc': two extremes in selection for polymorphism. *Immunological Reviews*, *143*(1), 63–88. https://doi.org/10.1111/j.1600-065X.1995.tb00670.x

Kelley, J., Walter, L., & Trowsdale, J. (2005). Comparative genomics of major histocompatibility complexes. *Immunogenetics*, *56*(10), 683–695. https://doi.org/10.1007/s00251-004-0717-7

Larsen, P. A., Heilman, A. M., & Yoder, A. D. (2014). The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. *BMC Genomics*, *15*(1), 720. https://doi.org/10.1186/1471-2164-15-720

Laver, T. W., Caswell, R. C., Moore, K. A., Poschmann, J., Johnson, M. B., Owens, M. M., Ellard, S., Paszkiewicz, K. H., & Weedon, M. N. (2016). Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Scientific Reports*, *6*, 21746. https://doi.org/10.1038/srep21746

Lenz, T. L., & Becker, S. (2008). Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci — Implications for evolutionary analysis. *Gene*, *427*(1), 117–123. https://doi.org/10.1016/j.gene.2008.09.013

Lighten, J., Oosterhout, C., Paterson, I. G., McMullan, M., & Bentzen, P. (2014). 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*, *14*(4), 753–767. https://doi.org/10.1111/1755-0998.12225

Lighten, J., van Oosterhout, C., & Bentzen, P. (2014). Critical review of NGS analyses for de novo genotyping multigene families. *Molecular Ecology*, *23*(16), 3957–3972. https://doi.org/10.1111/mec.12843

Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

Marmesat, E., Soriano, L., Mazzoni, C. J., Sommer, S., & Godoy, J. A. (2016). PCR strategies for complete allele calling in multigene families using high-throughput sequencing approaches. *PLoS One*, *11*(6), e0157402. https://doi.org/10.1371/journal.pone.0157402

McElroy, K. E., Luciani, F., & Thomas, T. (2012). GemSIM: General, error-model based simulator of next-generation sequencing data. *BMC Genomics*, *13*(1), 74. https://doi.org/10.1186/1471-2164-13-74

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science* Conference, 51–56.

Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A Simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*, *16*(3), 1215. https://doi.org/10.1093/nar/16.3.1215

Montero, B. K., Refaly, E., Ramanamanjato, J.-B., Randriatafika, F., Rakotondranary, S. J., Wilhelm, K., Ganzhorn, J. U., & Sommer, S. (2018). Challenges of next-generation sequencing in conservation management: Insights from long-term monitoring of corridor effects on the genetic diversity of mouse lemurs in a fragmented landscape. *Evolutionary Applications*, *12*(3), 425–442. https://doi.org/10.1111/eva.12723

Nakamura, T., Yamada, K. D., Tomii, K., & Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*, *34*(14), 2490–2492. https://doi.org/10.1093/bioinformatics/bty121

Nei, M., & Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, *39*, 121–152. https://doi.org/10.1146/annurev.genet.39.073003.112240

Nei, M., Xun, G. U., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences*, *94*(15), 7799–7806. https://doi.org/10.1073/pnas.94.15.7799

O'Connor, E. A., Westerdahl, H., Burri, R., & Edwards, S. V. (2019). Avian MHC evolution in the era of genomics: Phase 1.0. *Cells*, *8*(10), 1152. https://doi.org/10.3390/cells8101152

Parham, P., & Ohta, T. (1996). Population biology of antigen presentation by MHC class I molecules. *Science*, *272*(5258), 67–74. https://doi.org/10.1126/science.272.5258.67

Pavey, S. A., Sevellec, M., Adam, W., Normandeau, E., Lamaze, F. C., Gagnaire, P.-A., Filteau, M., Hebert, F. O., Maaroufi, H., & Bernatchez, L. (2013). Nonparallelism in MHCIIβ diversity accompanies nonparallelism in pathogen infection of lake whitefish (*Coregonus Clupeaformis*) species pairs as revealed by next-generation sequencing. *Molecular Ecology*, *22*(14), 3833–3849. https://doi.org/10.1111/mec.12358

Promerová, M., Babik, W., Bryja, J., Albrecht, T., Stuglik, M., & Radwan, J. (2012). Evaluation of Two approaches to genotyping major histocompatibility complex class I in a passerine—CE-SSCP and 454 pyrosequencing. *Molecular Ecology Resources*, *12*(2), 285–292. https://doi.org/10.1111/j.1755-0998.2011.03082.x

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Yong, G. U. (2012). A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341. https://doi.org/10.1186/1471-2164-13-341

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Radwan, J., Zagalska-Neubauer, M., Cichoń, M., Sendecka, J., Kulma, K., Gustafsson, L., & Babik, W. (2012). MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Molecular Ecology*, *21*(10), 2469–2479. https://doi.org/10.1111/j.1365-294X.2012.05547.x

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of next-generation sequencing reads. *Annual Review of Genomics and Human Genetics*, *16*(1), 133–151. https://doi.org/10.1146/annurev-genom-090413-025358

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5), R51. https://doi.org/10.1186/gb-2013-14-5-r51

Rozen, S., & Skaletsky, H. (1999). Primer3 on the WWW for general users and for biologist programmers. In S. Misener, & S. A. Krawetz (Eds.), *Bioinformatics Methods and Protocols* (365–386). Methods in Molecular Biology™. : Humana Press.

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*(2), 498–510. https://doi.org/10.1111/1755-0998.12453

Sepil, I., Moghadam, H. K., Huchard, E., & Sheldon, B. C. (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*(1), 68. https://doi.org/10.1186/1471-2148-12-68

Shaw, I., Powell, T. J., Marston, D. A., Baker, K., van Hateren, A., Riegert, P., Wiles, M. V., Milne, S., Beck, S., & Kaufman, J. (2007). Different evolutionary histories of the two classical class I genes BF1 and BF2 illustrate drift and selection within the stable MHC haplotypes of chickens. *The Journal of Immunology*, *178*(9), 5744–5752.

Smyth, R. P., Schlub, T. E., Grimm, A., Venturi, V., Chopra, A., Mallal, S., Davenport, M. P., & Mak, J. (2010). Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, *469*(1), 45–51. https://doi.org/10.1016/j.gene.2010.08.009

Sommer, S., Courtiol, A., & Mazzoni, C. J. (2013). MHC genotyping of non-model organisms using next-generation sequencing: A new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, *14*(1), 542. https://doi.org/10.1186/1471-2164-14-542

Stutz, W. E., & Bolnick, D. I. (2014). Stepwise threshold clustering: a new method for genotyping MHC loci using next-generation sequencing technology. *PLoS One*, *9*(7), e100587. https://doi.org/10.1371/journal.pone.0100587

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), e115. https://doi.org/10.1093/nar/gks596

Wallny, H.-J., Avila, D., Hunt, L. G., Powell, T. J., Riegert, P., Salomonsen, J., Skjodt, K., Vainio, O., Vilbois, F., Wiles, M. V., & Kaufman, J. (2006). Peptide motifs of the single dominantly expressed class I molecule explain the striking MHC-determined response to rous sarcoma virus in chickens. *Proceedings of the National Academy of Sciences of the United States of America 103* (5): 1434–1439. https://doi.org/10.1073/pnas.0507386103

Westbrook, C. J., Karl, J. A., Wiseman, R. W., Mate, S., Koroleva, G., Garcia, K., Sanchez-Lockhart, M., O'Connor, D. H., & Palacios, G. (2015). No assembly required: Full-length MHC class I allele discovery by pacbio circular consensus sequencing. *Human Immunology*, *Single-molecule DNA Sequencing*, *76*(12), 891–896. https://doi.org/10.1016/j.humimm.2015.03.022

Wittzell, H., Bernot, A., Auffray, C., & Zoorob, R. (1999). Concerted evolution of two Mhc class II B loci in pheasants and domestic chickens. *Molecular Biology and Evolution*, *16*(4), 479–490. https://doi.org/10.1093/oxfordjournals.molbev.a026130

Wood, S. (2006). *Generalized additive models: An introduction with R*. CRC Press.

Worley, K., Gillingham, M., Jensen, P., Kennedy, L. J., Pizzari, T., Kaufman, J., & Richardson, D. S. (2008). Single locus typing of MHC class I and class II B loci in a population of red jungle fowl. *Immunogenetics*, *60*(5), 233–247. https://doi.org/10.1007/s00251-008-0288-0

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Gillingham MAF, Montero BK, Wihelm K, Grudzus K, Sommer S, Santos PSC. A novel workflow to improve genotyping of multigene families in wildlife species: An experimental set-up with a known model system. *Mol Ecol Resour.* 2020;00:1–17. https://doi.org/10.1111/1755-0998.13290