A novel workflow to improve multi-locus genotyping of wildlife species: an experimental set-up with a known model system

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

[1]*University of Ulm, Institute of Evolutionary Ecology and Conservation Genomics, Albert-Einstein-Allee 11, 89069 Ulm, Germany* mark.gillingham@uni-ulm.de kerstin.wilhelm@uni-ulm.de kara.grudzus@uni-ulm.de simone.sommer@uni-ulm.de pablo.santos@uni-ulm.de

[2]*Zoological Institute, Animal Ecology & Conservation, Biocenter Grindel, Universität Hamburg, Hamburg, Germany* karina.montero@uni-hamburg.de

Keywords: open-source genotyping pipeline, ACACIA, next generation sequencing, amplicon genotyping, allele dropout, PCR amplification bias, sequencing bias, multigene family, MHC

Corresponding authors:
Mark Gillingham and Pablo Santos
Evolutionary Ecology and Conservation Genomics
Albert-Einstein-Allee 11
89069 Ulm
Germany
Phone: +49 731 50 22692
e-mail: mark.gillingham@uni-ulm.de; pablo.santos@uni-ulm.de

Abstract

Genotyping novel complex multigene systems is particularly challenging in non-model organisms. Target primers frequently amplify simultaneously multiple loci leading to high PCR and sequencing artefacts such as chimeras and allele amplification bias. Most next-generation sequencing genotyping pipelines have been validated in non-model systems whereby the real genotype is unknown and artefacts generated may be highly repeatable. Further hindering accurate genotyping, the relationship between artefacts and copy number variation (CNV) within a PCR remains poorly described. Here we investigate the latter by experimentally combining multiple known major histocompatibility complex (MHC) haplotypes (chicken, *Gallus gallus*, 43 artificial genotypes with 2-13 alleles per amplicon). In addition to well defined "optimal" primers, we simulated a non-model species situation by designing "naive" primers, with sequence data from closely related Galliform species. We applied a novel open-source genotyping pipeline (ACACIA) to the data, and compared its performance with another, previously published, pipeline. Finally, we applied ACACIA on a non-model system (grey-brown mouse lemurs, *Microcebus griseorufus*) with high CNV (MHC Class I exon 2 with up to 11 loci). ACACIA yielded very high allele calling accuracy (>98%). Non-chimeric artefacts increased linearly with increasing CNV but chimeric artefacts leveled when amplifying more than 4-6 alleles. As expected, we found heterogeneous amplification efficiency of allelic variants when co-amplifying multiple loci. Using our validated ACACIA pipeline and the example data of this study, we discuss in detail the pitfalls researchers should avoid in order to reliably genotype complex multigene systems. ACACIA is publicly available at https://gitlab.com/psc_santos/ACACIA.

Introduction

A key challenge for molecular ecologists is that they frequently work on systems with limited to no knowledge of their genomes. This means that the development of a genotyping approach often relies on information from closely related species available in genetic databases. Furthermore, assessing and validating genotyping methods can be particularly challenging when the structure of the target region is unknown.

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 non-model organisms. 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, Oosterhout, & Bentzen, 2014; Sebastian et al., 2016; Sommer, Courtiol, & Mazzoni, 2013). Unspecific locus amplification may lead to several biases during PCR since 1) chimeric sequences (hereafter “chimeras”; which may arise because of incomplete extension of sequences during a PCR cycle which 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 chimeric sequences can be generated (Lenz & Becker, 2008); 2) 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, Courtiol, & Mazzoni, 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, Wedel, & Wilson, 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: 1) chimeric sequences 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 a 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 2) 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 et al., 2014; Lighten, Oosterhout, & Bentzen, 2014; Sommer, Courtiol, & Mazzoni, 2013).

The recent rapid dissemination of next generation DNA sequencing (NGS) 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 cost and time needed previously (Babik, 2010; Sommer, Courtiol, & Mazzoni, 2013; Lighten et al., 2014). However, NGS 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; Sommer, Courtiol, & Mazzoni, 2013Liu, Keller, & Heckel, 2012; McElroy, Luciani, & Thomas, 2012; Ross et al., 2013). For instance, Illumina, currently the mainstream technology for NGS amplicon sequencing, report an error rate (primarily substitutions of base pairs) of $\leq$ 0.1% per base for $\geq$ 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 multi-locus-systems (>10) reported average amplification and sequencing artefact rates of 1.5% to 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, Courtiol, & Mazzoni, 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 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 NGS 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 non-model 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 et al., 2014; Radwan et al., 2012; Sebastian et al., 2016; Sommer, Courtiol, & Mazzoni, 2013). However for a given set of PCR primers and sequencing technology, PCR and sequencing bias, and thus in turn the rate of false positives and allele dropout, will be consistently repeatable (Biedrzycka et al., 2017). For instance, the high rate of Illumina substitution errors are known to be not random (see references within Sebastian et al., 2016) and therefore variants which 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 CNV of *a priori* known genotypes before PCR amplification and NGS sequencing.

In this study, we manipulated known combinations of the MHC alleles of a model organism (the chicken, *Gallus gallus*) as an example of a target multigene region of interest to molecular ecologists, in order to accurately quantify the effects of PCR and sequencing artefacts on genotyping pipelines. 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 (Nei, Gu, & Sitnikova, 1997; Nei & Rooney, 2005; Parham & Ohta, 1996) and MHC gene copies vary considerably across and even within a species (reviewed in Kelley, Walter, & Trowsdale, 2005). Therefore, the number of MHC loci present in a non-model study system often remains unknown. For instance, MHC Class IIB CNV was found to be as high as 21 in some passerine species, resulting in up to 42 allelic variants amplified within an amplicon and strong CNV between individuals (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 II loci (Kaufman, Jacob, et al., 1999; Kaufman, Milne, et al., 1999; Kaufman, Völk, & Wallny, 1995). The latter is therefore an ideal system to validate MHC genotyping pipelines for the following reasons: 1.) the structure of the B complex is well known with well-defined primers in conserved regions; 2.) 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 3.) CNV within an amplicon can be experimentally engineered by combining DNA samples from multiple MHC B complex haplotypes.

In order 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 **A**llele **CA**lling

pro**C**edure for **I**llumina **A**mplicon sequencing data). We experimentally generated a MHC dataset with a range of CNVs by combining DNA samples from multiple chicken MHC B complex haplotypes. Since MHC B complex in chickens is well characterised, optimal primers to amplify the entire exons which code for the antigen binding regions have been developed within the introns (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 replicate the challenge of designing primers for a non-model 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 "naïve primers"). The latter enabled us to test and quantify the relative amount of artefacts generated by naïve primer design as compared to optimal primers. We further tested our pipeline on a non-model system (an MHC Class I dataset from grey-brown mouse lemurs, *Microcebus griseorufus*) which varied significantly in CNV (up to 11 loci with considerable CNV).

Specifically, this study aimed to:

1. validate ACACIA using experimentally manipulated genotypes with different CNV that are known *a priori*;
2. accurately describe the relationship between PCR/sequencing artefacts and CNV by experimentally varying CNV and primer design in a model system;
3. test ACACIA in wildlife species with unknown genotypes of varying CNV (within and between species).

Materials and Methods

Samples and DNA extraction

Chicken blood samples originated from experimental inbred lines kept at the Institute for Animal Health at Compton UK (lines 7_2, C, WL and N) and the Basel Institute for Immunology in Basel Switzerland (lines H.B15 and H.B19+), as discussed (Jacob et al., 2000; Shaw et al., 2007; Wallny et al., 2006). These lines carry seven common B haplotypes: B2 (line 7_2), 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 are homozygotes at the MHC except line C, which was not used in this study. In each haplotype are two class II B 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, Dykes, & Polesky, 1988).

Grey-brown mouse lemurs were caught in Tsimanampetsotsa National Park in southwestern Madagascar. Field work was conducted between 2013 and 2015 following an established trapping protocol described in detail in Scheel et al. (2015). Genomic DNA was isolated from ear biopsies preserved in 70% ethanol. We performed DNA extractions using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, USA).

Generating 41 artificial MHC genotypes

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

Optimal primers for chicken MHC Class II

We targeted the entire 241 bp of exon 2 of MHC Class II, 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.

Naïve primer design for chicken MHC Class II

In order to naïvely design primers, we downloaded 61 exon 2 MHC Class II sequences from seven Galliform species (*Coturnix japonica*, *Crossoptilon crossoptilon*, *Meleagris gallopavo*, *Numida meleagris*, *Pavo cristatus*, *Perdix perdix* and *Phasianus colchicus*) from the 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′-WTCTACAACCGGCAGCAGT-3′) and the reverse primer GagaR2 (5′-TCCTCTGCACCGTGAWGGAC-3′) aiming at amplifying 151 bp of exon 2.

MHC Class I primer design of the grey-brown mouse lemur

Target-specific primers (MHCI-W04F: 5'CCCAGGCTCCCACTCCCT-3' and MHCI-W04R: 5'-GCGTCGCTCTGGTTGTAGT-3') were designed to flank the classical MHC class I exon 2 gene W04 (fragment length = 236 bp), previously described as a functional gene coding for antigen-binding sites (Averdam et al., 2009; Flügge et al., 2002). We designed primers from the

consensus sequence of the MHC class I gene from three primate species deposited in GenBank: grey mouse lemur (*Microcebus murinus,* accession numbers FP236833, AJ302085, AJ297588-AJ297590), ring-tailed lemur (*Lemur catta,* KC506599, AB098452), and rhesus monkey (*Macaca mulatta,* NM_001048245).

PCR Amplification, Library Preparation, and High-Throughput Sequencing

For all datasets we replicated all individuals in order to estimate repeatability ($n_{individuals}$ = 43 and $n_{amplicons}$ = 86 for the chicken datasets; and, $n_{individuals}$ = 147 and $n_{amplicons}$ = 294 for the grey-brown lemur dataset).

Individual PCR reactions 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 10bp barcode sequence and the adapter sequences used by the Illumina instrument during sequencing.

For the chicken datasets, the first PCR consisted of 3–5 ng of extracted DNA, 0.5 units FastStart Taq DNA Polymerase (Roche Applied Science, Mannheim, Germany), 1x PCR buffer, 4.5 mM $MgCl_2$, 250 μM each dNTP, 0.5 μM primers, and 5% dimethylsulfoxide (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, 72°C for 45 s, and a final extension step at 72°C for 10 min. The second PCR contained 2 μl of the product generated by the initial PCR, 80 nM per barcode primer, 0.5 units FastStart Taq DNA Polymerase, 1x PCR buffer, 4.5 mM $MgCl_2$, 250 μM each

dNTP, and 5% dimethylsulfoxide (DMSO) in a final volume of 20 µl. Cycling conditions were the same as those outlined above but the number of cycles was reduced to ten.

For the grey-brown mouse lemurs the first PCR round was carried out in a 10-µl reaction volume, including 1 µl DNA template, 0.5 µM primers, 1 µl GC enhancer and 1 unit AmpliTaq Gold 360 Master Mix. PCR cycling included an activation step at 95°C for 10 min followed by 25 cycles consisting of a denaturation step at 95 °C for 30 s, annealing at 64 °C for 30 s and extension at 72 °C for 60 s. A final extension step was omitted to reduce artefact formation (Smyth et al., 2010). The second PCR contained 2 µl of the product generated by the initial PCR, 1 µl GC enhancer, 80 nM per barcode primer and 0.5 units AmpliTaq Gold 360 Master Mix in a final volume of 20 µl. Cycling conditions were the same as those outlined above but the number of cycles was reduced to seven.

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.

Data analysis with the ACACIA pipeline

ACACIA is written in Python 2.7 and consists of 11 consecutive steps of data processing. One dedicated script performs the administration of all commands, handles data from one step to the

next and gathers information along the way. ACACIA uses a total of seven external programs (described below) within its workflow.

1. **Generating Quality Reports.** Sequencing quality is assessed for each FASTq file yielded by the sequencing platform, with the FastQC tool (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Visual reports for each file are produced as an output.
2. **Trimming of low quality ends of forward and reverse reads.** 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, new quality reports are generated as in step #1, and step #2 can be repeated. FASTq files are generated as an output.
3. **Merging of paired-end reads.** This step is for projects with paired-end sequencing only, and users can skip this step if using data from single-end sequencing (Note: the name of the matching forward and reverse FASTq files should be identical prior to the first "_", e.g.: File1_S1_L001_R1_001.fastq and File1_S1_L001_R2_001.fastq). The reads of these files are merged using FLASH (Magoč & Salzberg, 2011). The minimum and maximum lengths of overlap during merging can be changed by the user (defaults are zero and the length of the reads) (Magoč & Salzberg, 2011). As an output, FASTq files with merged sequences are generated, as well as a series of monitoring (log) files that allow users to check merging performance.
4. **Primer trimming.** After prompting users to enter the sequences of the primers used for amplicon sequencing, ACACIA searches the oligos at both ends of the merged sequences (IUPAC nucleotide ambiguity codes are allowed). When a perfect match for both primers is

297 found, they are trimmed and new, primer-less sequences, are written into FASTq files which
298 are the output of this step.

299 5. **Quality-control filter.** Users are then prompted to enter the values of two parameters (*q* and
300 *p*) to filter sequences based on their phred-score quality flags. First, *q* stands for *quality* and
301 denotes a phred-score threshold that can take values from 0 to 40. Second, *p* stands for
302 *percentage* and denotes the proportion of bases, in any given sequence, that need to achieve
303 at least the quality threshold *q* for that sequence to pass the quality filter. ACACIA uses the
304 default values $q = 30$ and $p = 90$ if users do not explicitly change them. In practical terms,
305 this threshold combination corresponds to an error probability lower than 10^{-3} in at least 90%
306 of bases for each sequence. All quality scores of sequences passing this filter are removed to
307 decrease file sizes and FASTA files with high-quality sequences are given as output.

308 6. **Singleton removal.** A large proportion of sequences contain random errors inherent to the
309 sequencing technology (Quail et al., 2012). In order to decrease file sizes without risking loss
310 of relevant allele information, ACACIA removes all singletons (sequences that appear one
311 single time) in an individual amplicon.

312 7. **Chimera removal.** The chimera identification tool VSEARCH (Rognes et al., 2016) is
313 employed here, with slightly altered settings (*alignwidth* = 0 and *mindiffs* = 1) aiming at
314 increasing sensitivity to chimeras that diverge very little from one of the "parent" sequences.
315 FASTA files with non-chimeric sequences, along with log files for each individual amplicon,
316 are given as output.

317 8. **Removal of unrelated sequences.** A local BLAST tool (Altschul et al., 1990) is used in this
318 step in order to compare all sequences with a set of reference sequences chosen by users.
319 This steps aims at removing sequences that passed all filters so far but are products of

unspecific priming during PCR. Typically, reference sequences can be downloaded from GenBank (www.ncbi.nlm.nih.gov/genbank/) depending on the gene family and taxonomy relevant for each project. Users are prompted to provide one FASTA file with reference sequences, which is converted to a local BLAST database and used for the BLAST search. Only sequences yielding high-scoring hits to the database (expectation value threshold = 10) are written into new FASTA files as an output of this step. This is the workflow's last filtering step.

9. **Aligning.** The MAFFT aligner (Katoh & Standley, 2013) is used to perform global alignment of sequences that have passed all 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 of eight cores and 32GB of RAM. Users with a significantly higher number of sequences might find it useful to increase parallelization of the aligner as described recently (Nakamura et al., 2018).
10. **Entropy analysis and calling of candidate alleles.** The Oligotyping tool (Eren et al., 2013) is used here to call candidate alleles. 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 is conceived as a supervised tool, we automated the selection of parameter values aiming at high tolerance. This has the advantage of running Oligotype unsupervised 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 amplicon are the output of this step. The automation

of parameter values is accomplished by a Python script that considers the number of sequences present in the alignment. Although originally conceived as a tool for identifying variants from microbiome 16S rRNA amplicon sequencing projects, Oligotyping is also ideal for other forms of highly variable amplicon sequencing projects.

11. **Allele Calling.** 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 1000, with default = 10.
    - Lowest proportion of reads (low_por): in order 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, but 0.01 is recommended for large data sets, 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 once 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. Before ACACIA assigns names to alleles, users are asked to provide a FASTA file with previously known names and sequences of alleles to be taken into consideration while assigning names. 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;
- alleles.fasta: a FASTA sequence file of all alleles identified in the run;
- pipelinereport.csv: a pipeline report quantifying reads and sequences failing and/or passing each pipeline step described above.

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 datasets (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 was 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 clustering threshold 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 which 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 much too high for our datasets, and we tested a range of filtering threshold between 0% and 1% at 0.1% intervals (i.e. 0%, 0.1%, 0.2% etc., supplementary Figure S1). We found that the optimal threshold for the AmpliSAS

394 filtering step for the optimal primer chicken dataset was 0.3% (Figure S1), whilst we found
395 higher allele calling accuracy with a filter threshold of 0.5% for the naïve primers dataset.

Results

Sequencing depth for each dataset and proportion of artefacts detected using ACACIA

A total of 530,101 paired-end reads were generated for the chicken optimal primers dataset, which amounted to an average of 6,164 reads per amplicon (*n* = 86). For the naïve primers dataset, 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 datasets 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. Finally for the MHC Class I exon 2 grey-brown mouse lemur dataset, a total of 15,050,630 reads were generated which amounted to an average of 51,192 reads per amplicon (*n* = 294). All negative controls (*n* = 2 chicken optimal primer; *n* = 2 chicken naïve primer; *n* = 4 grey-brown mouse lemurs) were clean (fewer than 500 sequences).

AmpliSAS vs ACACIA: chicken optimal primers dataset

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

Using the optimal settings in AmpliSAS 17 alleles dropped out, one false positive was found (Table 2) and allele calling accuracy was 97% (Figure S1). As with ACACIA, most allele dropouts (16 of 17) derived from the B21 haplotype. For three genotypes, both BLB2*21 and

BLB1*21 dropped out. For nine genotypes, only BLB2*21 alleles dropped out and for one genotype only BLB1*21 allele dropped out. Finally for one genotype the allele dropout was BLB2*04 and the same genotype had a false positive allele (Table 2). Allele calling repeatability was 95.5%.

AmpliSAS vs ACACIA: chicken naïve primers dataset

Using ACACIA, we found 134 allele dropouts and allele calling accuracy was 77.9%. However, all dropouts were from the alleles BLB2*04, BLB2*15 or BLB2*21, for which a primer mismatch was present. Therefore, all allele dropouts could be explained by primer design. Furthermore, allele calling repeatability between both replicates was 100%.

Using AmpliSAS, we found 149 allele dropouts and allele calling accuracy was 75.2%. As above, 134 dropouts were due to a mismatch with the forward primer. The remaining 15 alleles that dropped out were BLB2*12 or *19 (11 alleles) and BLB1*14 (4 alleles). Allele calling repeatability between both replicates was 98.3%.

ACACIA using wildlife species dataset: MHC Class I in grey-brown mouse lemurs

Using ACACIA we were able to identify 279 exon 2 MHC Class I alleles in 147 grey-brown mouse lemur individuals. Allele calling repeatability was 99.6%, with only six individuals where an allele was called in only one of the two replicates.

Relationship between number of alleles amplified and artefacts

The proportion of sequences classified as artefacts was much higher for PCRs using the optimal primer set than when using the naïve primer set (Figure 2a). For all chicken data sets, there is a

logarithmic relationship between the total proportion of artefacts and the number of alleles amplified (Figure 2a). However when considering non-chimeric artefacts, there was a positive relationship between the proportion of artefacts and the number of alleles amplified (Figure 2b). The proportion of chimeric reads no longer increased with number of alleles amplified when amplifying more than 4-6 alleles (Figure 2c). 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 2d). The total number of parental variants generating chimeras also did not increase with CNV when amplifying more than six alleles (Figure 2e). Finally, the contribution of allelic variants to the proportion of reads decreased sharply with increasing number of alleles when amplifying less than 4-6 alleles (Figure 2f). However the contribution of allele variants to the proportion of reads stabilised when amplifying more than 4-6 alleles (Figure 2f). Both alleles from the B21 haplotype in the optimal dataset and the BLB1*04 allele in the naïve dataset consistently amplified poorly when co-amplifying with alleles from other haplotypes (Figure 2f).

The proportion of chimeric artefacts was much smaller for the lemur dataset compared to the chicken datasets. The lemur dataset had few individuals with less than six alleles and no individuals with fewer than four alleles. Similarly to the chicken dataset, we observed a weak positive relationship between the total proportion of artefacts and the number of alleles amplified when amplifying more than 4-6 alleles (Figure 2a). However, the relationship between the proportion of non-chimeric artefacts and the number of alleles amplified was weaker than the chicken datasets (Figure 2b). In addition, the proportion of reads that were chimeric was much smaller for the lemur dataset than for chicken datasets (Figure 2c). The relationship between the

number of chimeric variants and the number of alleles amplified was similar to the chicken datasets (Figure 2d).

Discussion

Using known MHC genotypes for two datasets (chicken MHC Class II B complex), we achieved high allele calling accuracy (>98%) and repeatability (>97%) using ACACIA. With fewer allele dropouts and false positives, the ACACIA pipeline performed better than AmpliSAS. We additionally achieved very high allele calling repeatability (99.6%), when applying ACACIA to a wildlife species with a complex MHC class I system (grey-brown mouse lemur MHC Class I with a CNV of 4-21). 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 naïvely designed from sequences of related Galliform species. We also explored the relationship between artefacts and CNV, and found that surprisingly, the relationship between the proportion of chimeric artefacts and CNV was not linear but rather leveled when amplifying more than 4-6 alleles. However, non-chimeric artefacts did increase linearly with increasing CNV. As expected we found heterogeneous amplification efficiency of allelic variants when amplifying multiple loci within a PCR. Below we discuss in further detail the AmpliSAS and ACACIA genotyping pipelines, primer design for non-model organisms, the relationship between CNV and artefacts, the effect of chimera formation on genotyping pipelines and, finally, we conclude by advising users on important points to consider when genotyping complex multigene systems in non-model organisms.

AmpliSAS vs ACACIA

Experimentally generating CNV of known chicken MHC Class II genotypes allowed us to validate our ACACIA pipeline to genotype systems with high CNV complexity at high accuracy and repeatability across replicates. 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 better than AmpliSAS with all datasets. To demonstrate the latter we would need to test both pipelines on a larger number of datasets and/or on simulated datasets. In addition, while our pipeline should suit data generated with any next-generation sequencing technologies, we have only tested ACACIA with paired-end Illumina sequencing technology.

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 using default settings when applying the AmpliSAS pipeline to their dataset. We find this concerning since, as our study demonstrates, the default clustering and filtering parameters are unlikely to be optimal for most datasets. 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 dataset in our study, due to high allele dropout when using the default settings. We therefore strongly discourage users from using default settings and advise to permutate between different filtering and clustering parameters in order to find the best settings when using the AmpliSAS pipeline.

An important disadvantage of the AmpliSAS web server is that at the time of writing, sequencing depth per amplicon was limited to 5000 reads. The latter is particularly problematic when wishing to genotype systems with complex CNV, which require high sequencing depth to genotype with high repeatability (Biedrzycka et al., 2017). For datasets with sequencing depth above 5000 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 any understanding of 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 dataset. In particular, we urge users to permutate between different settings of *abs_nor* and *low_por* parameters.

The challenge of designing primers for non-model organisms

A common approach for primer design in complex genomic regions of non-model organisms includes downloading and aligning multiple sequences of phylogenetically related species. By building primers on consensus sequences, researchers hope that oligos will amplify the target region also in the species of interest. However, knowledge about related species is often limited to very few individuals. This means that, inevitably, 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 naïve primers dataset, 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 BLB1*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 (Figure 2f). However, high sequencing depth for the naïve primer dataset prevented this allele from dropping out, regardless of the genotyping pipeline used. Our study therefore highlights the importance of designing

multiple primers when wishing to genotype a novel target region in non-model organisms to limit allele dropout due to primer mismatch.

Relationship between number of alleles amplified and artefacts

By knowing the exact alleles to expect for the chicken genotypes, we were able to quantify chimeric artefacts precisely (Figure 1). There was a higher proportion of chimeric and non-chimeric artefacts in the optimal primer dataset than in the naïve primer dataset. The most likely explanation for the latter is the shorter sequence for the naïve primer dataset (151 bp) compared to the optimal primer dataset (241 bp). A shorter fragment reduces the number of base pairs that can be erroneously substituted 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.

For the lemur dataset, PCR conditions were modified to avoid chimera formation. The extension step within PCR cycles was increased and the final extension step was omitted. Such modification to PCR cycles are recommended to reduce the number of chimeras when co-amplifying multiple loci, because most incomplete primer extensions which generate chimeras are thought to be formed during the final extension step (Judo et al., 1998; Lenz & Becker, 2008; Smyth et al., 2010). Our data further supports the latter, since the lemur dataset had a much smaller proportion of reads that were chimeras than the two chicken datasets.

As expected the proportion of reads that were non-chimeric artefacts increased linearly as CNV increased, which can be explained simply by the fact that there is an increasing number of possible artefacts that can be generated as the number of initial template variants increases. The

slower rate of increase in the number of artefacts with increasing CNV for the lemur dataset compared to the chicken data can again be explained by the modified PCR conditions for the lemur dataset mentioned previously. Thus, once again most reads that failed to be completely elongated within the PCR cycles are more likely to be erroneously elongated during the final extension step.

A more unexpected result was that the proportions of reads that were chimera did not increase with increasing CNV when amplifying more than 4-6 alleles. Similarly, when amplifying more than 10 alleles, the number of chimeric variants no longer increased with increasing CNV. Such saturation in chimera generation beyond a CNV threshold is likely to be a by-product of allele PCR competition. Indeed, as demonstrated by our own data (Figure 2f), there is amplification bias whereby some gene variants are amplified preferentially relative to others (Marmesat et al., 2016; Sommer, Courtiol, & Mazzoni, 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 dataset did not increase with increasing CNV when amplifying more than 4-6 alleles. The non-linear relationship between chimera generation and CNV have important implications when considering sequencing depth needed to accurately genotype complex multigene system. Below we discuss in further detail, the challenges of dealing with chimeras in genotyping pipelines.

The challenge of dealing with 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 systems under high rates of interlocus genetic exchange ("concerted evolution"), which is common in many MHC systems (Reto et al., 2008; Reto et al., 2010; Edwards, Grahn, & Potts, 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.

For the optimal primer dataset, regardless of the genotyping pipeline used, allele dropout occurred in genotypes with high CNV (for ACACIA 8 out of 9 and for AmpliSAS 12 out of 14 haplotypes had a CNV < 10). For all instances bar one, allele dropout were alleles from the B21 haplotype which amplified poorly when CNV was greater than 6 (Figure 2f). Higher sequencing depth will reduce or even remove such allele dropout instances (Biedrzycka et al., 2017). Indeed for the naïve primer dataset, sequencing depth was twice as high, and there were no instances of allele dropout due to the ACACIA pipeline (all allele dropouts were due to primer mismatch). One allele erroneously called as a real variant (i.e. a false positive) by the AmpliSAS pipeline in the optimal primer dataset was actually a chimera between the BLB1*21 and BLB2*21 alleles. Furthermore, when using the AmpliSAS pipeline, 15 allele dropouts in the naïve primer dataset were due to erroneous assignment of real allelic variants as chimera artefacts. Indeed, the B BLB2*12 or *19 minor allele was identical to potential chimeric artefact sequences between BLB1*14 (85 possible breakpoints) and any of the following alleles: BLB2*04, BLB1*15, BLB1*19, BLB1*21 or BLB2*21 (Figure 3a). In addition, BLB1*14 dropped out because it is

identical to a chimeric sequence between the BLB2*02 minor and BLB2*12 or *19 alleles (33 breakpoints; Figure 3b).

We have identified two factors which 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 3a and Figure 3b, whereby any breakpoint in the shaded areas leads to the same chimeras. Second, peripheral breakpoints (Figure 3c) can generate chimeric sequences that differ to parental sequences by as little as a single base pair. For instance, a chimera could be a product of the allele BLB1*21 combined with any of the other alleles shown in the alignment, with a breakpoint within the shaded area (Figure 3c). Since the potential breaking points 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 BLB1*21 by only one base). In 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.

Conclusion

Genotyping accuracy and artefacts are intrinsically linked. 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. We are convinced that universal default settings for optimal genotyping accuracy cannot be achieved, since optimal parameters will depend on dataset-specific generation of artefacts. The latter, in turn, varies according to species-specific CNV, 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 inter-locus recombination.

Author contributions

MG and PS conceived the study. PS designed ACACIA. MG did the data analysis in R. MG, PS and KM ran the allele calling workflows. KM did the AmpliSAS analysis and the lab work for the lemur samples. KW participated in and supervised the lab work. KG did the lab work for the chicken samples. SS instigated the study and heads the lab where the work was carried out. MG and PS wrote the first draft of the paper. All authors participated in the writing of the manuscript and contributed after the first version of the manuscript.

Data accessibility

Raw sequences of all datasets are available in the NCBI Sequence Read Archive (SRA; accession number: X) under BioProject X. ACACIA is freely available on the GitLab at https://gitlab.com/psc_santos/ACACIA, under an MIT license.

Acknowledgments

MG was supported by a DFG grant (DFG Gi 1065/2-1). We are very grateful to Jim Kaufman providing the chicken DNA samples used in this study and for his comments on a previous version of this work.

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. doi: 10.1016/S0022-2836(05)80360-2

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., … Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *MSystems*, *2*(2), e00191-16. doi: 10.1128/mSystems.00191-16

Averdam, A., Petersen, B., Rosner, C., Neff, J., Roos, C., Eberle, M., … Walter, L. (2009). A Novel System of Polymorphic and Diverse NK Cell Receptors in Primates. *PLOS Genetics*, *5*(10), e1000688. doi: 10.1371/journal.pgen.1000688

Babik, W. (2010). Methods for MHC genotyping in non-model vertebrates. *Molecular Ecology Resources*, *10*(2), 237–251. doi: 10.1111/j.1755-0998.2009.02788.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. doi: 10.1111/1755-0998.12612

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. doi: 10.1111/1755-0998.12234

Burri, Reto, 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), 1180–91. doi: 10.1093/molbev/msn065

Burri, Reto, 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. doi: 10.1093/molbev/msq120

678 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P.
679 (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature*
680 *Methods*, *13*(7), 581–583. doi: 10.1038/nmeth.3869
681 Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene
682 conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*,
683 *8*(10), 762–775. doi: 10.1038/nrg2193
684 Edwards, S., Grahn, M., & Potts, W. (1995). Dynamics of Mhc evolution in birds and
685 crocodilians: amplification of class II genes with degenerate primers. *Molecular Ecology*,
686 *4*, 719–729.
687 Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L.
688 (2013). Oligotyping: differentiating between closely related microbial taxa using 16S
689 rRNA gene data. *Methods in Ecology and Evolution*, *4*(12), 1111–1119. doi:
690 10.1111/2041-210X.12114
691 Flügge, P., Zimmermann, E., Hughes, A. L., Günther, E., & Walter, L. (2002). Characterization
692 and Phylogenetic Relationship of Prosimian MHC Class I Genes. *Journal of Molecular*
693 *Evolution*, *55*(6), 768–775. doi: 10.1007/s00239-002-2372-7
694 Galan, M., Guivier, E., Caraux, G., Charbonnel, N., & Cosson, J.-F. (2010). A 454 multiplex
695 sequencing method for rapid and reliable genotyping of highly polymorphic genes in
696 large-scale studies. *BMC Genomics*, *11*(1), 296. doi: 10.1186/1471-2164-11-296
697 Gillingham, M. a. F., Courtiol, A., Teixeira, M., Galan, M., Bechet, A., & Cezilly, F. (2016).
698 Evidence of gene orthology and trans-species polymorphism, but not of parallel
699 evolution, despite high levels of concerted evolution in the major histocompatibility
700 complex of flamingo species. *Journal of Evolutionary Biology*, *29*(2), 438–454. doi:
701 10.1111/jeb.12798
702 Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology*
703 *Resources*, *11*(5), 759–769. doi: 10.1111/j.1755-0998.2011.03024.x

704 Goto, R. M., Afanassieff, M., Ha, J., Iglesias, G. M., Ewald, S. J., Briles, W. E., & Miller, M. M.
705 (2002). Single-strand conformation polymorphism (SSCP) assays for major
706 histocompatibility complex B genotyping in chickens. *Poultry Science*, *81*(12), 1832–
707 1841. doi: 10.1093/ps/81.12.1832
708 Hess, C., & Edwards, S. (2002). The evolution of the major histocompatibility complex in birds.
709 *Bioscience*, *52*(5), 423–431.
710 Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007). Accuracy and
711 quality of massively parallel DNA pyrosequencing. *Genome Biology*, *8*(7), R143. doi:
712 10.1186/gb-2007-8-7-r143
713 Jacob, J. P., Milne, S., Beck, S., & Kaufman, J. (2000). The major and a minor class II β-chain
714 (B-LB) gene flank the Tapasin gene in the B-F /B-L region of the chicken major
715 histocompatibility complex. *Immunogenetics*, *51*(2), 138–147. doi:
716 10.1007/s002510050022
717 Judo, M. S. B., Wedel, A. B., & Wilson, C. (1998). Stimulation and suppression of PCR-
718 mediated recombination. *Nucleic Acids Research*, *26*(7), 1819–1825. doi:
719 10.1093/nar/26.7.1819
720 Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
721 Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4),
722 772–780. doi: 10.1093/molbev/mst010
723 Kaufman, J, Jacob, J., Shaw, I., Walker, B., Milne, S., Beck, S., & Salomonsen, J. (1999). Gene
724 organisation determines evolution of function in the chicken MHC. *Immunological*
725 *Reviews*, *167*, 101–17.
726 Kaufman, J, Milne, S., Göbel, T. W., Walker, B. a, Jacob, J. P., Auffray, C., … Beck, S. (1999).
727 The chicken B locus is a minimal essential major histocompatibility complex. *Nature*,
728 *401*(6756), 923–5. doi: 10.1038/44856

729 Kaufman, Jim, Völk, H., & Wallny, H.-J. (1995). A “Minimal Essential Mhc” and an
730 “Unrecognized Mhc”: Two Extremes in Selection for Polymorphism. *Immunological*
731 *Reviews*, *143*(1), 63–88. doi: 10.1111/j.1600-065X.1995.tb00670.x
732 Kelley, J., Walter, L., & Trowsdale, J. (2005). Comparative genomics of major histocompatibility
733 complexes. *Immunogenetics*, *56*(10), 683–695. doi: 10.1007/s00251-004-0717-7
734 Laver, T. W., Caswell, R. C., Moore, K. A., Poschmann, J., Johnson, M. B., Owens, M. M., …
735 Weedon, M. N. (2016). Pitfalls of haplotype phasing from amplicon-based long-read
736 sequencing. *Scientific Reports*, *6*, 21746. doi: 10.1038/srep21746
737 Lenz, T. L., & Becker, S. (2008). Simple approach to reduce PCR artefact formation leads to
738 reliable genotyping of MHC and other highly polymorphic loci — Implications for
739 evolutionary analysis. *Gene*, *427*(1), 117–123. doi: 10.1016/j.gene.2008.09.013
740 Lighten, J., Oosterhout, C., Paterson, I. G., McMullan, M., & Bentzen, P. (2014). Ultra-deep
741 Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for
742 copy number variation in the guppy (Poecilia reticulata). *Molecular Ecology Resources*,
743 *14*(4), 753–767. doi: 10.1111/1755-0998.12225
744 Lighten, J., Oosterhout, C. van, & Bentzen, P. (2014). Critical review of NGS analyses for de
745 novo genotyping multigene families. *Molecular Ecology*, *23*(16), 3957–3972. doi:
746 10.1111/mec.12843
747 Liu, Y., Keller, I., & Heckel, G. (2012). Breeding site fidelity and winter admixture in a long-
748 distance migrant, the tufted duck (Aythya fuligula). *Heredity*, *109*(2), 108–116. doi:
749 10.1038/hdy.2012.19
750 Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast Length Adjustment of Short Reads to Improve
751 Genome Assemblies. *Bioinformatics*, btr507. doi: 10.1093/bioinformatics/btr507
752 Marmesat, E., Soriano, L., Mazzoni, C. J., Sommer, S., & Godoy, J. A. (2016). PCR Strategies
753 for Complete Allele Calling in Multigene Families Using High-Throughput Sequencing
754 Approaches. *PLOS ONE*, *11*(6), e0157402. doi: 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. doi: 10.1186/1471-2164-13-74

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.

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

Nei, M., Gu, X., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences*, *94*(15), 7799–7806.

Nei, M., & Rooney, A. P. (2005). Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics*, *39*, 121–152. doi: 10.1146/annurev.genet.39.073003.112240

Parham, P., & Ohta, T. (1996). Population Biology of Antigen Presentation by MHC Class I Molecules. *Science*, *272*(5258), 67–74. doi: 10.1126/science.272.5258.67

Pavey, S. A., Sevellec, M., Adam, W., Normandeau, E., Lamaze, F. C., Gagnaire, P.-A., … 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. doi: 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. doi: 10.1111/j.1755-0998.2011.03082.x

780 Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. (2012).
781 A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific
782 Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341. doi:
783 10.1186/1471-2164-13-341
784 R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved
785 from https://www.R-project.org/
786 Radwan, J., Zagalska-Neubauer, M., Cichoń, M., Sendecka, J., Kulma, K., Gustafsson, L., &
787 Babik, W. (2012). MHC diversity, malaria and lifetime reproductive success in collared
788 flycatchers. *Molecular Ecology*, *21*(10), 2469–2479. doi: 10.1111/j.1365-
789 294X.2012.05547.x
790 Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open
791 source tool for metagenomics. *PeerJ*, *4*, e2584. doi: 10.7717/peerj.2584
792 Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., … Jaffe, D. B.
793 (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5),
794 R51. doi: 10.1186/gb-2013-14-5-r51
795 Rozen, S., & Skaletsky, H. (1999). Primer3 on the WWW for General Users and for Biologist
796 Programmers. In S. Misener & S. A. Krawetz (Eds.), *Bioinformatics Methods and*
797 *Protocols* (pp. 365–386). doi: 10.1385/1-59259-192-2:365
798 Scheel, B. M., Henke-von der Malsburg, J., Giertz, P., Rakotondranary, S. J., Hausdorf, B., &
799 Ganzhorn, J. U. (2015). Testing the Influence of Habitat Structure and Geographic
800 Distance on the Genetic Differentiation of Mouse Lemurs (Microcebus) in Madagascar.
801 *International Journal of Primatology*, *36*(4), 823–838. doi: 10.1007/s10764-015-9855-z
802 Sebastian, A., Herdegen, M., Migalska, M., & Radwan, J. (2016). amplisas: a web server for
803 multilocus genotyping using next-generation amplicon sequencing data. *Molecular*
804 *Ecology Resources*, *16*(2), 498–510. doi: 10.1111/1755-0998.12453

805 Sepil, I., Moghadam, H. K., Huchard, E., & Sheldon, B. C. (2012). Characterization and 454
806 pyrosequencing of Major Histocompatibility Complex class I genes in the great tit reveal
807 complexity in a passerine system. *BMC Evolutionary Biology*, *12*(1), 68. doi:
808 10.1186/1471-2148-12-68
809 Shaw, I., Powell, T. J., Marston, D. A., Baker, K., van Hateren, A., Riegert, P., … Kaufman, J.
810 (2007). Different evolutionary histories of the two classical class I genes BF1 and BF2
811 illustrate drift and selection within the stable MHC haplotypes of chickens. *The Journal of*
812 *Immunology*, *178*(9), 5744–5752.
813 Smyth, R. P., Schlub, T. E., Grimm, A., Venturi, V., Chopra, A., Mallal, S., … Mak, J. (2010).
814 Reducing chimera formation during PCR amplification to ensure accurate genotyping.
815 *Gene*, *469*(1), 45–51. doi: 10.1016/j.gene.2010.08.009
816 Sommer, S., Courtiol, A., & Mazzoni, C. J. (2013). MHC genotyping of non-model organisms
817 using next-generation sequencing: a new methodology to deal with artefacts and allelic
818 dropout. *BMC Genomics*, *14*(1), 542. doi: 10.1186/1471-2164-14-542
819 Stutz, W. E., & Bolnick, D. I. (2014). Stepwise Threshold Clustering: A New Method for
820 Genotyping MHC Loci Using Next-Generation Sequencing Technology. *PLOS ONE*,
821 *9*(7), e100587. doi: 10.1371/journal.pone.0100587
822 Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S.
823 G. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, *40*(15),
824 e115–e115. doi: 10.1093/nar/gks596
825 Wallny, H.-J., Avila, D., Hunt, L. G., Powell, T. J., Riegert, P., Salomonsen, J., … Kaufman, J.
826 (2006). Peptide motifs of the single dominantly expressed class I molecule explain the
827 striking MHC-determined response to Rous sarcoma virus in chickens. *Proceedings of*
828 *the National Academy of Sciences of the United States of America*, *103*(5), 1434–1439.
829 doi: 10.1073/pnas.0507386103
830 Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

831 Wittzell, H., Bernot, A., Auffray, C., & Zoorob, R. (1999). Concerted evolution of two Mhc class II
832 B loci in pheasants and domestic chickens. *Molecular Biology and Evolution*, *16*(4), 479–
833 490.

834
835

836 **Figure 1**: Flow diagram of reads and sequences from two Illumina runs analysed with ACACIA.
837 Blue bars correspond to filters, and the percentages given correspond to the sequences kept at
838 each step for further analyses. The percentage given at the bottom for artefacts refers to the total
839 amount of reads in the beginning of the process. (Fwd & Rev) raw forward and reverse reads;
840 (Mrg) Paired-end read merger; (Prm) primer filter; (QC) quality control; (Sgt) Singleton
841 removal; (Chm) chimera removal; (Blt) BLAST filter.

842

843 **Figure 2:** The relationship between the number of alleles amplified and: the total (chimeric +
844 non-chimeric reads) proportions of reads that are artefacts (a.); the proportion of non-chimeric
845 reads (b.); the proportion of chimeric reads (c.); the absolute number of chimeric variants (d.);
846 the absolute number of parental variants generating chimeric reads (e.); and, the proportion of
847 reads for each real allelic variant. The lemur dataset was excluded from (e.) because few
848 individuals had fewer than 6 alleles and none had fewer than 4 (however like the chicken
849 datasets the absolute number of parental variants generating chimeric reads also did not increase
850 with increasing number of alleles amplified). All relationships were fitted with general additive
851 model using the ggplot package (Wickham, 2016) in R (R Core Team, 2018) using a binomial
852 distribution for (a.), (b.), (c.) and (f.), and a Poisson distribution corrected for over-dispersion for
853 (d.) and (e.).

854

855

856 **Figure 3:** Three alignments with examples of sequences which can be classified as chimeras.
857 The points denote identity to the first sequence in each alignment, while the differences to it are
858 highlighted. The shaded areas indicate possible chimera-yielding breakpoints. (a) The allele
859 BLB2*12 or *19 could be a chimera of BLB1*14 with any of the four other allele sequences
860 depicted, in a case of multiple potential parent pairs. (b) BLB1*14 can be interpreted as a
861 chimera between BLB2*12 or *19 minor and BLB2*02. (c) Actual chimeric sequence with
862 multiple potential parents and a peripheral breakpoint, which is therefore very similar to one of
863 its parents.

864

865 **Figure S1:** AmpliSAS accuracy and repeatability for the optimal primer chicken dataset at
866 different filtering thresholds (i.e. ‘minimum amplicon frequency’).

867 **Table 1:** The number of alleles per genotype, the number of
868 genotypes with a certain number alleles and the number of
869 amplicons with a certain number alleles (all genotypes were
870 duplicated) for the chicken datasets used in this study. The
871 list haplotypes used to artificially create the genotypes are
872 listed in supplementary Table S1.

Number of alleles per genotype	Number of genotypes	Number of amplicons
2	7	14
4	7	14
6	7	14
8	7	14
10	7	14
11	5	10
12	2	4
13	1	2
Total	**43**	**86**

873

874 **Table 2:** Genotypes with allele dropouts and false positives using ACACIA and AmpliSAS

875 (excluding allele dropout due to primer mismatch in the naïve primers dataset).

Genotype	Replicate	Number of predicted alleles	Allele dropout using ACACIA	Allele dropout using AMPLISAS	False positive using AMPLISAS
a. Chicken optimal primers dataset (BLB MHC Class II)					
B2-B4-B12-B14-B19-B21	1	11	BLB1*21	BLB1*21 BLB2*21	
B4-B14-B15-B19-B21	1	10	BLB1*21	BLB2*21	
	2	10	BLB1*21	BLB1*21	
B4-B15-B19-B21	1	8	BLB1*21	BLB2*21	
B2-B4-B12-B14-B15-B19-B21	1	13	BLB1*21 BLB2*21	BLB1*21 BLB2*21	
B2-B4-B12-B14-B15-B21	1	12	BLB1*21 BLB2*21	BLB1*21 BLB2*21	
B2-B12-B14-B15-B19-B21	1	11	BLB2*21		
B2-B4-B12-B15-B19-B21	1	11		BLB2*21	
B2-B4-B12-B15-B21	1	10		BLB2*21	
B2-B4-B14-B15-B19-B21	1	12		BLB2*21	
B2-B4-B14-B15-B21	1	10		BLB2*21	
B2-B4-B15-B19-B21	1	10		BLB2*21	
	2	10		BLB2*21	
B4-B12-B21	1	6		BLB2*04	1 false positive
B4-B14-B15-B19-B21	2	10		BLB2*21	
b. Chicken naïve primers dataset (BLB MHC Class II)					
B12-B14-B15-B21	1	5		BLB2*12 or *19	
	2	5		BLB2*12 or *19	
B2-B12-B14-B15	1	6		BLB2*12 or *19	
	2	6		BLB2*12 or *19	
B2-B14-B15-B19-B21	1	7		BLB1*14	
B2-B4-B12-B14-B15	1	7		BLB2*12 or *19	
	2	7		BLB2*12 or *19	
B2-B4-B12-B14-B15-B19	1	8		BLB1*14	
B2-B4-B12-B14-B15-B19-B21	1	9		BLB1*14	
B2-B4-B12-B14-B15-B21	1	8		BLB2*12 or *19	
B2-B4-B12-B14-B19-B21	1	8		BLB1*14	
B4-B12-B14-B15	1	5		BLB2*12 or *19	
	2	5		BLB2*12 or *19	
B4-B14-B15-B19-B21	1	6		BLB2*12 or *19	
	2	6		BLB2*12 or *19	

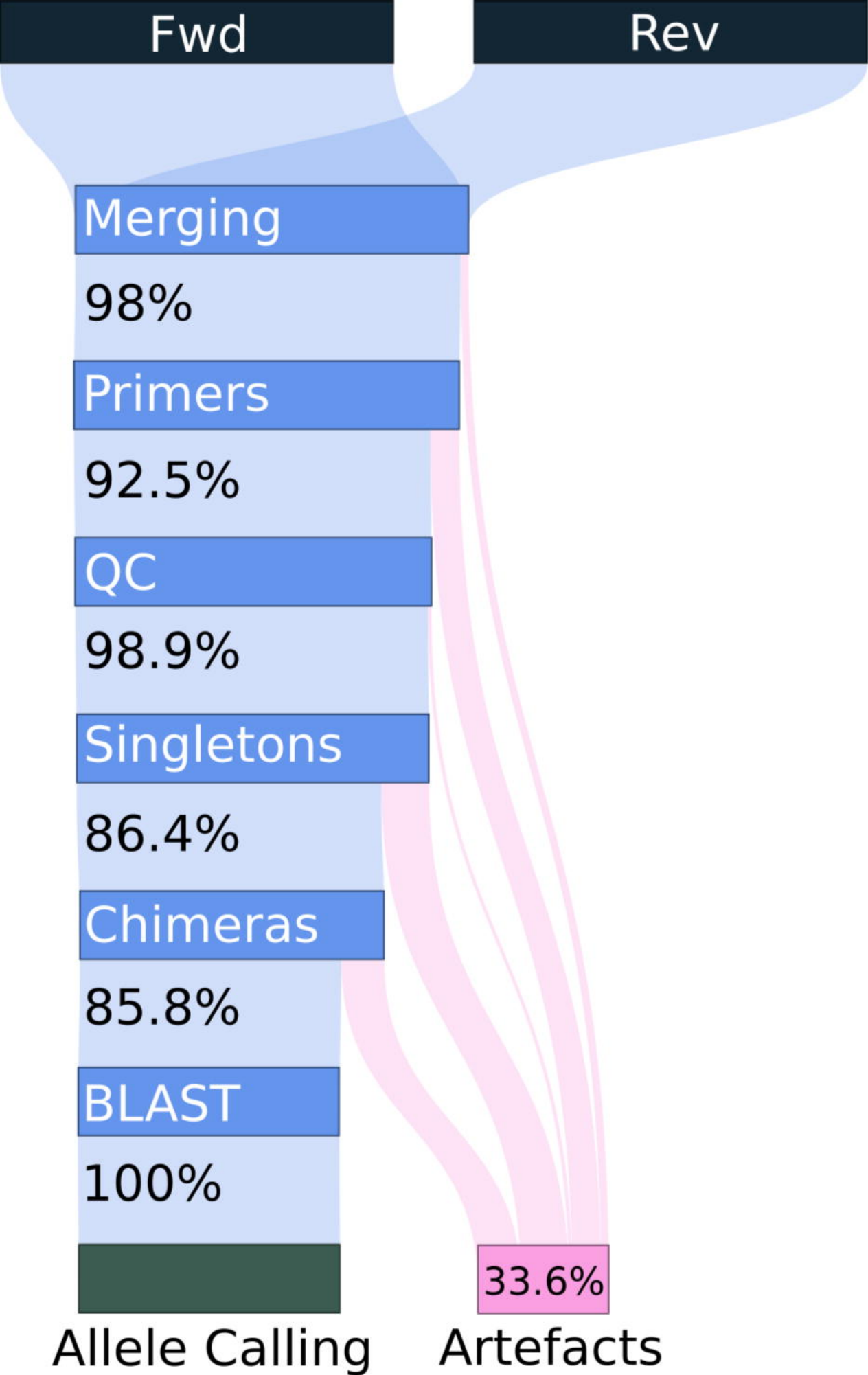
876

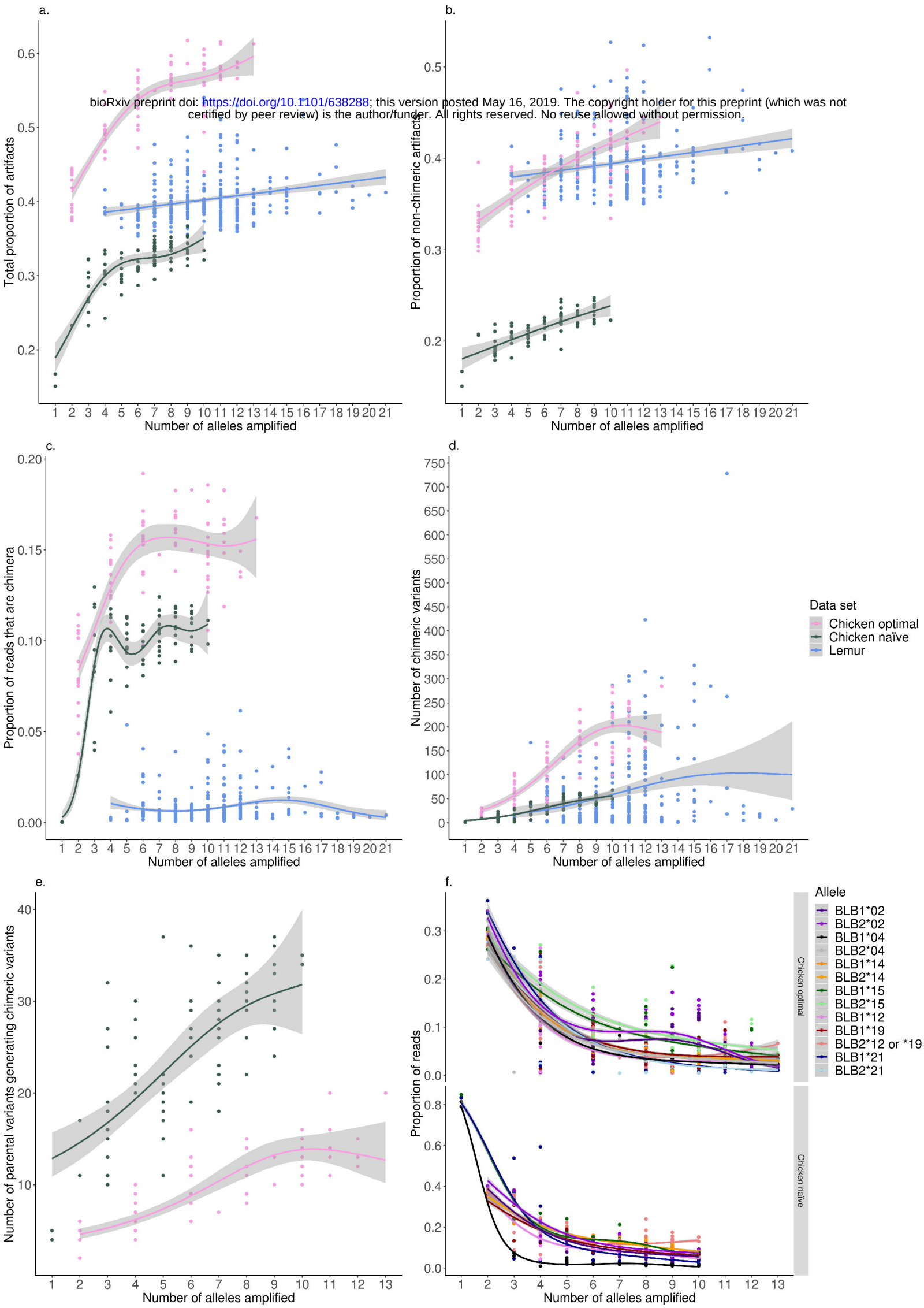
877

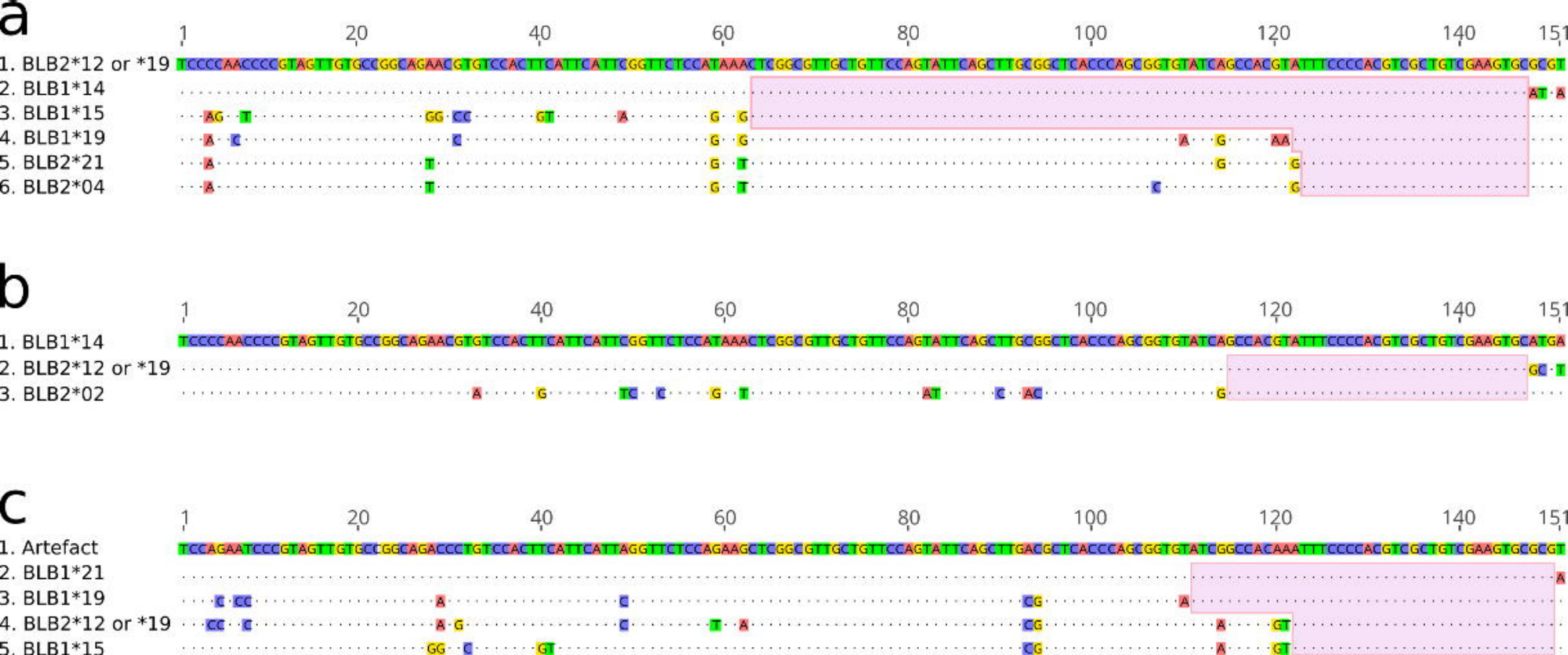
878 **Table S1:** The chicken MHC *B* complex haplotypes and combined
879 haplotypes which formed experimental genotypes with varying
880 copy number variation (CNV).

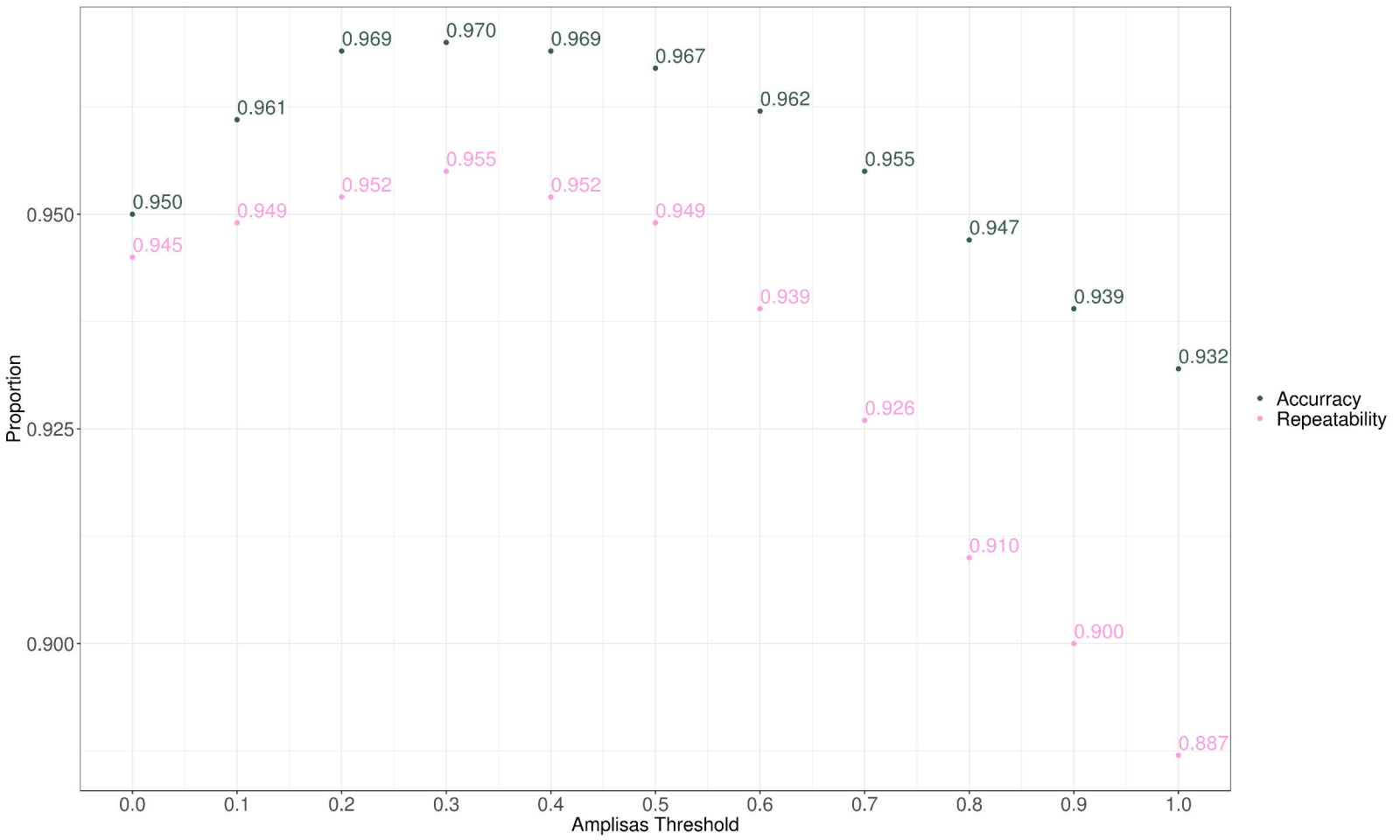
Combined haplotypes	Number of alleles
B2	2
B4	2
B12	2
B14	2
B15	2
B19	2
B21	2
B2-B4	4
B2-B12	4
B4-B12	4
B12-B14	4
B12-B21	4
B14-B15	4
B19-B21	4
B2-B4-B19	6
B2-B14-B19	6
B2-B15-B19	6
B4-B12-B21	6
B4-B14-B19	6
B12-B14-B21	6
B15-B19-B21	6
B2-B4-B12-B14	8
B2-B12-B14-B15	8
B2-B14-B19-B21	8
B4-B12-B14-B15	8
B4-B15-B19-B21	8
B12-B14-B15-B21	8
B14-B15-B19-B21	8
B2-B4-B12-B14-B15	10
B2-B4-B12-B14-B21	10
B2-B4-B12-B15-B21	10
B2-B4-B14-B15-B21	10
B2-B4-B15-B19-B21	10
B2-B14-B15-B19-B21	10
B4-B14-B15-B19-B21	10
B2-B4-B12-B14-B15-B19	11
B2-B4-B12-B14-B15-B21	12
B2-B4-B12-B14-B19-B21	11
B2-B4-B12-B15-B19-B21	11
B2-B4-B14-B15-B19-B21	12
B2-B12-B14-B15-B19-B21	11
B4-B12-B14-B15-B19-B21	11
B2-B4-B12-B14-B15-B19-B21	13

881


Fwd
Rev
Merging
98%
Primers
92.5%
QC
98.9%
Singletons
86.4%
Chimeras
85.8%
BLAST
100%
33.6%
Allele Calling
Artefacts




a
1 20 40 60 80 100 120 140 151
1. BLB2*12 or *19
2. BLB1*14
3. BLB1*15
4. BLB1*19
5. BLB2*21
6. BLB2*04
b
1 20 40 60 80 100 120 140 151
1. BLB1*14
2. BLB2*12 or *19
3. BLB2*02
c
1 20 40 60 80 100 120 140 151
1. Artefact
2. BLB1*21
3. BLB1*19
4. BLB2*12 or *19
5. BLB1*15


0.969
0.970
0.969
0.967
0.961
0.962
0.955
0.950
0.952
0.952
0.955
0.949
0.949
0.947
0.945
0.939
0.939
0.932
0.926
0.910
0.900
0.887
0.950
0.925
0.900
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Amplisas Threshold
Accurracy
Repeatability