Heredity (2015) 114, 485–493

www.nature.com/hdy

ORIGINAL ARTICLE

Large-scale genotyping of highly polymorphic loci by next-generation sequencing: how to overcome the challenges to reliably genotype individuals?

M Ferrandiz-Rovira[1,2], T Bigot[1], D Allainé[1], M-P Callait-Cardinal[1,2] and A Cohas[1]

Studying the different roles of adaptive genes is still a challenge in evolutionary ecology and requires reliable genotyping of large numbers of individuals. Next-generation sequencing (NGS) techniques enable such large-scale sequencing, but stringent data processing is required. Here, we develop an easy to use methodology to process amplicon-based NGS data and we apply this methodology to reliably genotype four major histocompatibility complex (MHC) loci belonging to MHC class I and II of Alpine marmots (*Marmota marmota*). Our post-processing methodology allowed us to increase the number of retained reads. The quality of genotype assignment was further assessed using three independent validation procedures. A total of 3069 high-quality MHC genotypes were obtained at four MHC loci for 863 Alpine marmots with a genotype assignment error rate estimated as 0.21%. The proposed methodology could be applied to any genetic system and any organism, except when extensive copy-number variation occurs (that is, genes with a variable number of copies in the genotype of an individual). Our results highlight the potential of amplicon-based NGS techniques combined with adequate post-processing to obtain the large-scale highly reliable genotypes needed to understand the evolution of highly polymorphic functional genes.
Heredity (2015) **114,** 485–493; doi:10.1038/hdy.2015.13; published online 11 March 2015

INTRODUCTION

Understanding the evolution of highly polymorphic functional genes is a major issue in evolutionary biology, ecology and medical science (for example, genes involved in immunity in insects (Ellis *et al.*, 2012), in plants (Jones, 2001) and in vertebrates (Hedrick, 1994; Akira *et al.*, 2001); genes regulating self-incompatibility in plants (Charlesworth and Awadalla, 1998) or in fungi (Wu *et al.*, 1998)). Reaching this goal requires the comprehension of the role of different selective pressures acting on functional genes. For this purpose, information regarding the genetic makeup of free-living individuals at various temporal and spatial scales is needed (Vasemägi and Primmer, 2005).

Amplicon-based next-generation sequencing (NGS) techniques allow the sequencing of large numbers of individuals at affordable costs (the term amplicon refers here to a pool of sequenced reads obtained for a given individual, locus and PCR). Moreover, amplicon-based NGS techniques are highly suitable to genotype genes with high allelic diversity and provide a viable method to obtain an accurate genotyping of all copies for genes characterized by a variable number of copies (that is, with copy-number variation; Babik, 2010; Promerová *et al.*, 2012). Indeed, amplicon-based NGS techniques have been widely used to genotype individuals at a wide range of highly polymorphic functional loci such as the major histocompatibility complex (MHC) in rodents (Babik *et al.*, 2009; Galan *et al.*, 2010; Sommer *et al.*, 2013), primates (Huchard *et al.*, 2012), carnivores (Oomen *et al.*, 2013), birds (Radwan *et al.*, 2012; Sepil *et al.*, 2012; Dunn *et al.*, 2013) and fish (Herdegen *et al.*, 2014; Lighten *et al.*, 2014b); the *CYP7B1* and *SPG7* genes involved in the hereditary spastic paraplegia in humans (Schlipf *et al.*, 2011); toll-like receptor genes in New Zealand robin (*Petroica australis*; Grueber *et al.*, 2012); and the S locus receptor kinase gene involved in plant self-incompatibility in rock cress (*Arabidopsis sp.*; Jørgensen *et al.*, 2012).

Owing to their wide range of applications, amplicon-based NGS techniques have undergone an extremely fast development (see Shokralla *et al.* (2012) for details on development of NGS technologies) leading to the emergence of several NGS platforms such as 454, Illumina, SOLiD or Ion Torrent. Although these NGS platforms may differ in terms of the number of obtained reads per run, the number of base pairs (bps) per read, the run time or the cost per run (see Glenn, 2011; Shokralla *et al.*, 2012 for reviews of available NGS techniques), all amplicon-based NGS techniques generate multiple reads for a given amplicon and are prone to errors such as artifacts (substitutions and chimeras) during PCR or indels during the sequencing process (Glenn, 2011). Discerning true alleles versus sequencing errors is thus challenging, especially where several functionally distinct alleles may differ by a single nucleotide (Babik, 2010), and post-processing procedures are then needed. How to deal with these errors is a fundamental challenge for large-scale sequencing studies and two approaches have been proposed (Lighten *et al.*, 2014a; Table 1).

The first approach aims to classify each variant as an allele or an artifact based on two assumptions: (1) that variants corresponding to alleles will be overrepresented compared with artifacts, and (2) that

[1]Laboratoire Biométrie et Biologie Evolutive, Université de Lyon, CNRS, UMR5558, Université Lyon 1, F-69622, Villeurbanne, F-69000 Lyon, France and [2]Université Lyon, VetAgro Sup Campus Vet, Marcy-L'Étoile, France
Correspondence: Dr A Cohas, Laboratoire Biométrie et Biologie Evolutive, Université de Lyon, CNRS, UMR5558, Université Lyon 1, F-69622, Villeurbanne, F-69000 Lyon, France.
E-mail: aurelie.cohas@univ-lyon1.fr
Received 7 May 2014; revised 19 January 2015; accepted 23 January 2015; published online 11 March 2015

Table 1 Comparison of large-scale studies of amplicon-based NGS using 454 platforms and post-processing for characterizing MHC of non-model vertebrates

Taxon	*Species*	*MHC class*	*Loci*	*Number of loci*	*Number of alleles/locus*	*Number of amplicons*	*Number of genotyped amplicons*	*% Assigned genotypes*	*Mean ± s.d. coverage per amplicon (max)*	*Cost per amplicon*[a] *(€)*	*Cost per genotyped amplicon*[b] *(€)*	*References*
Mammals	*Marmota marmota*	I	UB	4	2	1152	834	72%	134 ± 148 (1647)	1.58	2.37	Present study
			UD		3	1152	790	69%	100 ± 109 (850)			
		II	DRB1		8	1152	719	63%	69 ± 85 (707)			
			DRB2		3	1152	726	63%	99 ± 124 (1008)			
	Myodes glareolus	II	DRB	>8	<9.12	96	79	82%	94 ± 44	76.03	92.39	Babik *et al.*, 2009
	Gulo gulo	II	DRB1/	2	2/8	10	10	100%	5199 ± 1586	91.01	91.01	Oomen *et al.*, 2013
			DRB2			78	78	100%	5491 ± 1704	46.67	46.67	
	11 rodent Genera	II	[c]	15	NA	1566 (22–650)[d]	1407 (4–602)[d]	90%	74 ± NA	4.66	5.19	Galan *et al.*, 2010
	Microcebus murinus	II	DQB	2	61	672	643	96%	394 ± 429	5.07	5.63	Huchard *et al.*, 2012
			DRB		60	768	654	85%	394 ± 477			
Fish	*Gasterosteus aculeatus*	II	IIb	NA	244[e]	364	295	81%	442 ± NA (4687)	20	24.68	Stutz and Bolnick 2014
Birds	*Ficedula albicollis*	II	[f]	[f]	11.9 ± 3.2	249	222	89%	1244 ± 474 (2666)	29.24	32.80	Radwan *et al.*, 2012
	Parus major	I	[f]	[f]	23.8	1536	871	56%	286 ± 173	4.75	8.38	Sepil *et al.*, 2012

Abbreviations: MHC, major histocompatibility complex; NA, not applicable; NGS, next-generation sequencing.
[a]Overall cost of 454 sequencing divided by the number of amplicons excluding DNA extraction and PCR cost.
[b]Overall cost of 454 sequencing divided by the number of genotyped amplicons excluding DNA extraction and PCR cost.
[c]See Galan *et al.* (2010) for more details on the sequenced loci and species.
[d]Number of samples depending on the studied Genera.
[e]Number of alleles at the overall MHC genotyped loci.
[f]Birds have an overall locus organization differing from that of mammals and should not be compared with mammals, see Radwan *et al.* (2012) and Sepil *et al.* (2012) for more details on how to name bird's loci.

artifacts will be more similar to alleles than alleles will be from each other. Protocols then filter out obvious artifacts (too short or too long reads), filter variants based on their frequency of occurrence within an entire run and/or within an amplicon and apply criteria based on sequence similarity to differentiate artifacts from alleles (for example, Babik *et al.*, 2009; Galan *et al.*, 2010; Huchard *et al.*, 2012; Radwan *et al.*, 2012). The second and very recent approach (Stutz and Bolnick, 2014; Lighten *et al.*, 2014b), rather than applying frequency or similarity criteria to determine whether a given variant is an allele or not, groups variants into clusters based on the similarity of the corresponding sequences and then identify alleles on a cluster by cluster basis rather than on a variant by variant basis. Both these approaches suffer from validation over a very limited number of samples using cloning or duplicate PCRs of the same samples (generally <30 samples, but see Sommer *et al.*, 2013 for a counter example).

Building on these approaches, we propose a genotyping protocol suitable for any genes which does not present extensive copy-number variations between individuals of a given species and for data generated by amplicon-based NGS techniques. Compared with previously published protocols, this protocol (1) can be applied to amplicons with very low coverage, (2) can be readily implemented thanks to a web interface, (3) has been extensively validated (1450 validated genotypes) and (4) is highly reliable (error rate = 0.21%). This protocol allows reliable genotyping of amplicons with low coverage thanks to the assignation of reads based only on a few bps of the primers instead of the whole sequence and to clustering reads containing sequencing errors such as indels. Given that amplicon-based NGS techniques remain expensive (Glenn, 2011; Shokralla *et al.*, 2012), the proposed protocol contributes to the reduction of the sequencing costs.

The proposed protocol has been applied to data generated by 454 sequencing (suitable since the targeted MHC sequences range from 175 bp to 230 bp) on four MHC loci (*Mama-UB*, *Mama-UD*, *Mama-DRB1* and *Mama-DRB2*; Kuduk *et al.*, 2012) of 1096 wild Alpine marmots (*Marmota marmota*). The MHC is a multi-gene family present in all jawed vertebrates and the MHC genes are the most polymorphic loci known in the vertebrate genome (reviewed in Kelley *et al.*, 2005). These genes, divided into three main families named class I, II and III (Kelley *et al.*, 2005), encode for proteins involved in antigen presentation and have a critical role in vertebrate disease resistance by initiating the immune response (Hedrick, 1994). The quality of the assigned genotypes was assessed through: (1) checking intra-individual repeatability of amplicon-based NGS sequencing; (2) comparing genotypes of a given individual obtained by amplicon-based NGS and Sanger sequencing; and (3) assessing the consistency of the obtained genotypes using the parentage relationships previously established based on a panel of 16 microsatellites.

MATERIALS AND METHODS

Amplicon-based NGS data processing

NGS data processing to obtain individual genotypes can be divided into the four following steps:

- Step 1. Assignment of reads to loci and individuals, elimination of singletons and elimination of reads with inappropriate sizes;
- Step 2. Elimination of amplicons with insufficient coverage;
- Step 3. Determination of alleles; and
- Step 4. Determination of homozygous and heterozygous amplicons.

All steps are fully described below. These steps are carried out using two custom and heavily commented Python scripts that are freely available at https://github.com/tbigot/alFinder, accompanied by example files and data set. A user friendly web interface (http://pbil.univ-lyon1.fr/software/alFinder/config_generator/alFinder_config_step1) is available to apply the procedure.

In the case where alleles have been previously described, the user can process directly through all the steps described below. In the case where alleles have not been previously described, the user should conduct a pre-treatment to identify alleles' sequences. To this purpose, the user should follow the steps described below, but retaining only reads equal to the expected allele length in step 1.3 and skipping step 3. The post-processing procedure can then be conducted (following all steps described below) to maximize the number of retained reads and amplicons thanks to indels' assignment (see step 3).

Step 1. Assignment of reads to loci and individuals, elimination of singletons and elimination of reads with inappropriate sizes. Only reads containing perfect analogous individual tags (forward and reverse) and the minimal number of bp allowing to distinguish with no ambiguity between the forward and reverse primers (for example, the first x-bp of the different primers used or any x-bp within the primer sequence) instead of complete primers that are retained from the FASTA files produced as a result of the sequencing run (Step 1.1). The use of x-bp of the primers instead of the complete primers was designed to maximize the number of retained reads per individual. The number and position of bp used is flexible and depends on the design of the primers and on the degree of stringency needed. After cutting off tags and primers, the library file is compressed by removing singletons (that is, variants represented by a single read; Step 1.2) as well as reads with $<95\%$ or $>105\%$ of the expected allele length (Step 1.3) to decrease the opportunity for inclusion of PCR chimeras and to maximize the number of reads with indels retained.

Step 2. Elimination of amplicons with insufficient coverage. Given that a minimal number of reads is required to obtain a reliable genotype and that several identical reads must be present within a given amplicon to obtain a reliable allele (Galan *et al.,* 2010), the minimal number of reads per amplicon needs to be calculated to ensure a negligible probability of missing alleles. The model proposed by Galan *et al.* (2010) was used to assess this number per amplicon. In this model, the confidence level (f) to determine a correct genotype depends on three components: (1) r, the minimum required number of copies of the given allelic variant within an amplicon; (2) n, the total number of reads for a given amplicon; and (3) m, the maximum number of alleles within an amplicon (that is, m is fixed to two for one locus in a diploid species). The program 'Negative Multinomial' (implemented by Galan *et al.* (2010) and freely available online) is used to determine the minimum value of n for a confidence level (f) of at least 95%. Thus, amplicons with less reads than the total number of reads for a given amplicon (n) are discarded at this stage.

Step 3. Determination of alleles. Variants corresponding to previously described alleles are identified and named following the original names of the alleles. Remaining variants are classified in two groups as follows: (1) variants with a length corresponding to the one of the previously described alleles (that is, correct length variants) and (2) variants with a length different than the one corresponding to the previously described alleles (that is, incorrect length variants).

To increase the number of assigned reads thanks to the use of sequencing errors, all variants are aligned using the progressive alignment (Feng and Doolittle, 1987) with the default aligning parameters of the CLC Sequence Viewer software free trial version 6.7.1 (CLC Bio, Aarhus, Denmark). As the primary artifacts generated by amplicon-based NGS techniques differ among platforms used (for example, 454 and Ion Torrent primary generate indels, SOLiD generates A-T bias and Illumina generates substitutions (Glenn, 2011)), this step is more efficient in retaining additional reads for 454 and Ion Torrent instruments than for other technologies. Studies using SOLiD or Illumina instruments could thus skip this step until further implementations may allow the exploitation of artifacts generated by these techniques.

First, among correct length variants, those presenting indels that lead to a change in all amino acids following the indels (that is, variants with one insertion and one deletion resulting from sequencing errors) are assigned manually to a previously identified alleles or to another correct length variant. Other correct length variants are not assigned as they could be either artifacts (that is, indels) or true alleles produced by mutation-selection effects. Assigning such variants could lead to the underestimation of allelic diversity. Second, incorrect length variants presenting indels are assigned manually to previously identified alleles or to correct length variants. Assignation of variants with indels allows to increase the evidence that the most frequent variant present in a given amplicon is a true allele (Stutz and Bolnick, 2014). This procedure could be especially useful when true alleles are represented by a few reads due to a lower sequencing efficiency relative to other alleles (Sommer *et al.,* 2013).

Step 4. Determination of homozygous and heterozygous amplicons. To determine homozygous and heterozygous amplicons, the model proposed by Hohenlohe *et al.* (2010), and later used by Etter *et al.* (2011), is used to calculate, given the sequencing error rate, the likelihood of each possible genotype given all the retained variants of an amplicon. An homozygous or an heterozygous genotype is assigned to each amplicon based on a likelihood ratio test between the most likely homozygous and the most likely heterozygous genotypes with one degree of freedom. If the likelihood ratio test is not significant, no genotype is assigned.

Field methods and sample collection

DNA samples were collected in three wild populations of Alpine marmots located in the Nature Reserve of the Grande Sassière (at 2340 m.a.s.l., French Alps, 45°29′N, 65°90′E), the Gran Paradiso National Park (at 2190 m.a.s.l., Italian Alps, 45°35′N, 7°11′E) and the Llosa valley (at 1900 m.a.s.l., Catalan Pyrenees, 42° 26′N, 1°42′E).

In the Grande Sassière Nature Reserve, Alpine marmot samples have been collected since 1990. Individuals from 26 family groups were captured, tranquillized with Zolétil 100 (0.1 ml kg^{-1}), sexed and their age was determined from their size (up to 3 years). Their social status was determined according to scrotal development for males and teat development for females, and hair samples and skin biopsies were collected for molecular analysis. The composition of family groups was assessed from both capture–recapture data and intensive observations (see Cohas *et al.,* 2006 for details on the observation protocol). Similar protocols have been used since 2008 in the Gran Paradiso and the Llosa valley populations.

From the captured individuals, 1036, 30 and 30 genetic samples were obtained from the Grande Sassière, the Gran Paradiso and the Llosa valley populations, respectively (Table 2).

DNA extraction

For these 1096 individuals (Table 2), genomic DNA was extracted from 15 to 30 hairs or skin biopsies by incubation at 66 °C for 80 min for hairs and at 56 °C for 120 min for skin biopsies in 50 µl lysis buffer (20 mM Tris-HCl, 1.5 mM $MgCl_2$, 25 mM KCl, 0.5% Tween20 and 0.1 mg ml^{-1} proteinase K), followed by 20 min of proteinase K inactivation at 96 °C.

MHC characterization

Six MHC loci have yet been characterized for the Alpine marmot: four MHC class I loci (*Mama-UA*, *Mama-UB*, *Mama-UC* and *Mama-UD*) and two MHC class II DRB loci (*Mama-DRB1* and *Mama-DRB2*; Kuduk *et al.,* 2012). Among these loci, *Mama-UB*, *Mama-UD*, *Mama-DRB1* and *Mama-DRB2* were polymorphic (two alleles for *Mama-UB*, three for *Mama-UD*, eight for *Mama-DRB1* and three for *Mama-DRB2*), whereas polymorphism was not detected at *Mama-UA* and *Mama-UC*. Thus, only the four polymorphic loci were sequenced in this study.

Amplicon-based NGS

Amplification of MHC loci used specific primers for Alpine marmots designed to separate each locus and to avoid sequencing several duplicated genes with a single pair of primers (see Kuduk *et al.,* 2012 for more details). Amplification of MHC class I used the primers MarmMF1/MarmR2 for *Mama-UB* and MarmMF1/MarmR4 for *Mama-UD* (Kuduk *et al.,* 2012). Amplification of MHC class II DRB loci used the primers MM_DRB_F1/MM_DRB_R3 for *Mama-DRB1* and MM_DRB_F2a/MM_DRB_R2 for *Mama-DRB2*

Table 2 Summary of the individuals and amplicons used for MHC and microsatellite sequencing

	Number of captured individuals	*Number of individuals with microsatellite genotypes*	*Number of individuals to be sequenced at 4 MHC loci*	*Number of amplicons to be sequenced per MHC locus*	*Number of amplicons to be sequenced for all 4 MHC loci*	*Number of amplicons successfully sequenced at all 4 MHC loci*
Total	1096	1036	1096	1152[a]	4608[a]	3069[a]
The Sassière	1036	1036	1036	1092[a]	4368[a]	2894[a]
Grand Paradiso	30	0	30	30	120	91
The Llosa	30	0	30	30	120	84

Abbreviation: MHC, major histocompatibility complex.
[a]Includes replicated individuals for validation procedures.

(Kuduk *et al.,* 2012). MM_DRB_F2a (5′-ACGATTCCTGCAGCAGATGA-3′) was used instead of MM_DRB_F2 (5′-GAGTGTCATTTTTTCAATrGGA-3′; Kuduk *et al.,* 2012), because the high number of homopolymers in the MM_DRB_F2 primer region interferes with the extraction of reads. A GS FLX Titanium Primer Adapter (5′-CGTATCGCCTCCCTCGCGCCATCAG-3′ for forward primer and 5′-CTATGCGCCTTGCCAGCCCGCTCAG-3 for reverse primer) necessary for the 454 sequencing were added to these locus-specific primers. A 6-bp tag used for barcoding individuals was also added between the adapters and the locus-specific primers. Sixteen different tags in the forward primers and 12 different tags in the reverse primers (Supplementary Table S1) with a minimum of 3-bp differences between tags were used to ensure a very low probability for a read to be assigned to the wrong individual owing to a typing mistake in the individual tag and to allow us to pool 144 amplicons per locus.

PCR was performed in a Mastercycler (Eppendorf, Hamburg, Germany) in 10 µl of reaction mixture containing 5 µl HotStarTaq Polymerase Master Mix (QIAGEN, Hilden, Germany), 0.2 µl of both primers at 100 µM, 3.6 µl water and 1 µl DNA at a concentration of 30 ng µl^{-1}. The cycling scheme was 95 °C for 15 min, followed by 34 cycles at 95 °C for 30 s, primer-specific annealing temperature (50 °C for *Mama-UD* and *Mama-DRB2* and 55 °C for *Mama-UB* and *Mama-DRB1*) for 30 s, 72 °C for 60 s and a final extension step at 72 °C for 10 min. The concentration of the PCR products was measured by fluorometry using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Alcobendas, Madrid, Spain). Equimolar amounts of 144 amplicons were pooled for a given locus, and purified using MinElute PCR Purification Kit (QIAGEN). The amplicons for each of the four loci were subsequently pooled and sequenced on an eighth of a PicoTiterPlate of a Roche (Indianapolis, IN, USA) 454 FLX sequencing instrument (576 amplicons/eighth). The total of the 4608 amplicons were thus sequenced on a full PicoTiterPlate (Table 2).

Validation procedures

Three validation procedures were conducted on the 1036 sequenced individuals from the Grande Sassière only (Table 2): validation through NGS repeatability, through Sanger sequencing and through parentage relationships.

For validation through NGS repeatability, 30 individuals were amplified, for each locus, in 2–5 independent PCRs (mean number of PCR/individual/loci ± s.d.: 2.87 ± 0.82) using different tag combinations, yielding a total of 86 amplicons per locus. Consequently, from the 1096 individuals for whom DNA was extracted, 1152 amplicons at each locus were included in the sequencing process (Table 2).

For validation through Sanger sequencing, DNA extractions and PCR amplifications were carried out for nine individuals at four MHC loci (36 genotypes) following the same protocols as above. 30 µl of PCR products were purified using the Axygen Cleanup Kit for PCR (Union City, CA, USA). Purified PCR products were single-strand sequenced using Dideoxynucleotide Terminator (Dyeterminator, kit BigDye v.3.1 provided by LifeTechnologies, Alcobendas, Madrid, Spain). Sanger sequences are prone to sequencing errors in the 5′ region. So, for the polymorphic regions to be on the 3′ of the Sanger sequences, we sequenced with the reverse primer, when the polymorphism was at the end of the sequence of interest, and with the forward primer, when the polymorphism was at the beginning of the sequence of interest (MarmR2 (forward) for *Mama-UB*, MarmR4 (forward) for *Mama-UD*, MM_DRB_R3 (forward) for *Mama-DRB1* and MM_DRB_F2 (reverse) for *Mama-DRB2*). Sequenced products were then purified using AxyPrep Mag Dye Clean (Axygen, following manufacturer's instructions). The DNA sequencing reactions were then analyzed on an ABI3730XL 96 caps DNA Analyzer (LifeTechnologies). Obtained reads were aligned using the same procedure as stated earlier. Genotypes were considered as homozygous when they perfectly matched with a unique described allele, whereas they were considered as heterozygous when ambiguous positions in the DNA sequence were found and perfectly matched with two described alleles.

For validation through parentage relationships, 1036 individuals (Table 2) were genotyped at 16 microsatellites: SS-Bibl1, SS-Bibl18, SS-Bibl20, SS-Bibl31, SS-Bibl4 (Klinkicht, 1993); MS41, MS45, MS47, MS53, MS56, MS6, ST10 (Hanslik and Kruckenhauser, 2000); Ma002, Ma018, Ma066, Ma091 (Da Silva *et al.,* 2003; see Supplementary Data S1 for details on protocol and microsatellite characteristics). Parentage analyses (see Cohas *et al.,* 2008 and Supplementary Data S1) enabled us to confirm maternity and paternity for 663 pups. These 663 mother–father–offspring triads were then used to check for consistency of the obtained MHC genotypes.

RESULTS

454 sequencing output

The 454 sequencing run yielded a total of 584 745 reads. The 6-bp total length of the individual tags (forward and reverse) and the first 3-bp of the locus-specific forward and reverse primers were used to assign reads to a given individual and locus. Among the obtained reads, 459 143 reads perfectly matched the individual tags and the 3-bp primer markers. Between 966 and 1069 amplicons were obtained by retaining 3-bp primers, whereas between 922 and 1056 amplicons would have been obtained by retaining reads with complete primers. The mean coverage per amplicon varied across amplicons (mean coverage per amplicon ranged between 69 and 134, see Table 1 for details on each locus) and was lower than expected (expected coverage per amplicon: 217 reads).

After removing singletons and reads of $<95\%$ or $>105\%$ of the correct length variants (Table 3), 245 178 reads were retained and the coverage was similar for all loci with respect to the 1152 genotype calls for each locus (948 retained amplicons (82%) for *Mama-UB*; 944 (82%) for *Mama-UD*; 903 (78%) for *Mama-DRB1* and 901 (78%) for *Mama-DRB2*).

The minimum number of reads per amplicon (n) was found to be 12 (corresponding to a confidence level of 96.1%) after fixing r to 3 following Galan *et al.* (2010) and m to 2 (the Alpine marmot being a diploid species and each locus being amplified separately). A total of 242 067 reads were retained after discarding amplicons for which the total number of reads per locus was lower than 12 (Table 3). Again, similar numbers of retained amplicons were found at each locus with respect to the 1152 genotype calls for each locus (835 retained

amplicons (72%) for *Mama-UB*; 793 (69%) for *Mama-UD*; 720 (63%) for *Mama-DRB1* and 726 (63%) for *Mama-DRB2*).

At this stage, 578 variants were identified. Sixteen out of these variants corresponded to previously described alleles (Kuduk *et al.,* 2012), 55 variants had correct length (all belonging to *Mama-UD* locus) and the remaining 507 had incorrect length (77 for *Mama-UB*, 2 for *Mama-UD*, 233 for *Mama-DRB1* and 195 for *Mama-DRB2*). Among the 55 correct length variants, 44 *Mama-UD* variants contained indels and were assigned to previously identified alleles. Among the 507 incorrect length variants, 318 contained indels and were assigned to previously identified alleles (53 for *Mama-UB*, 104 for *Mama-DRB1* and 117 for *Mama-DRB2*). The remaining 11 correct length variants and the remaining 233 incorrect length variants could either be PCR artifacts/chimeras or new alleles differing from those described in Kuduk *et al.* (2012). After the determination of homozygous and heterozygous amplicons, all these 244 remaining variants were removed.

The assignment of indels to previously described alleles allowed us to increase the number of retained reads by 9.56% (194 996 reads were obtained by assigning indels, whereas 215 601 would have been obtained without assigning indels) and to increase the total number of retained amplicons (3069 amplicons were obtained by assigning indels, whereas 2866 would have been obtained without assigning indels).

The distribution of variant frequencies per sample at each locus (Figure 1), indicated two main peaks corresponding to frequencies of 90–100% and 25–75%. This distribution is consistent with the expected distribution of homozygous and heterozygous amplicons. The sequencing error rate for 454 FLX sequencing instrument has been estimated to range between 0.18% and 0.03% (Huse *et al.,* 2007). Sequencing error rate was assumed to be constant at each locus (but see Hohenl*ohe et al*., 2010; Et*ter et a*l., 2011). Indeed, at this stage of the post-processing most of the sequencing errors should have been eliminated thanks to the previous post-processing steps (for example, the elimination of singletons and reads of $<95\%$ and $>105\%$ correct length variants or the use of indels). We thus fixed the sequencing error rate to 0.03%. After determining the most likely genotype for each amplicon, a total of 239 879 reads were retained (Table 3). Similar numbers of retained amplicons were found at each locus with respect to the 1152 genotype calls for each locus (834 retained amplicons (72%) for *Mama-UB*; 790 (69%) for *Mama-UD*; 719 (63%) for *Mama-DRB1* and 726 (63%) for *Mama-DRB2*). Thus, from a total of 1152 genotype calls at each locus (4608 at the four MHC

Table 3 Amplicon-based next generation processing steps to minimize the impact of sequencing errors among all MHC studied loci

Processing description	*Total number of reads (N. amp[a])*	*Mama-UB reads (N. amp[a])*	*Mama-UD reads (N. amp[a])*	*Mama-DRB1 reads (N. amp[a])*	*Mama-DRB2 reads (N. amp[a])*
Number of reads (all mixed loci)	584 745	—	—	—	—
Elimination of reads without both forward and reverse individual tags and 3-bp locus-specific primers (Step 1.1)	459 143 (4103)	152 960 (1067)	113 862 (1069)	79 114 (966)	113 207 (1001)
Elimination of singletons (Step 1.2) and reads with $<95\%$ or $>105\%$ of correct length variants (Step 1.3)	245 178 (3696)	72 214 (948)	59 642 (944)	49 212 (903)	64 110 (901)
Elimination of amplicons with <12 reads for a given locus (Step 2)	242 067 (3074)	71 611 (835)	58 911 (793)	48 291 (720)	63 254 (726)
Elimination of variants after determining the most likely genotype for each amplicon (Step 3 and 4)	239 879 (3069)	70 817 (834)	58 424 (790)	47 251 (719)	63 387 (726)

Abbreviation: MHC, major histocompatibility complex.
[a]N. amp: number of sequenced amplicons.

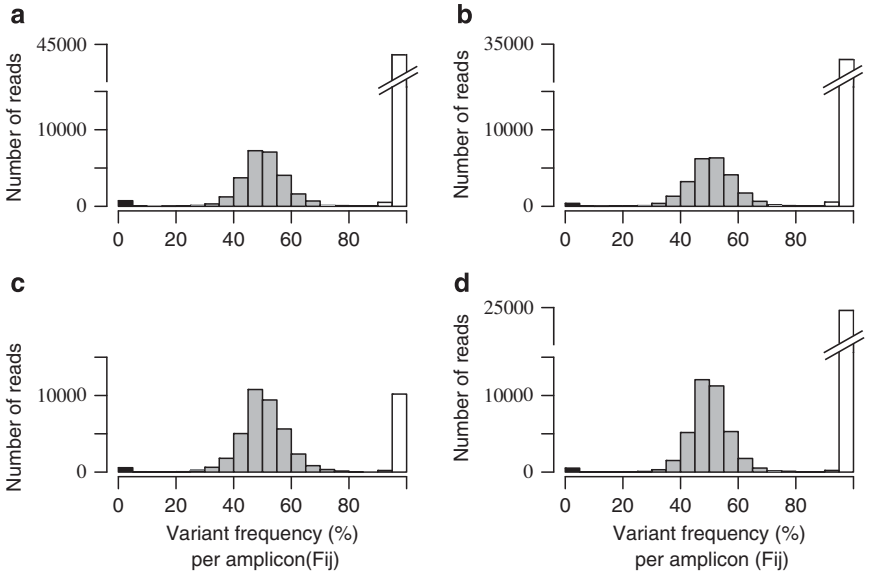

Figure 1 Distribution of the reads according to the frequency of the variant (*i*) within the amplicon *j* (F_{ij}) they belong over all amplicons with 12 or more reads for *Mama-UB* (**a**), *Mama-UD* (**b**), *Mama-DRB1* (**c**) and *Mama-DRB2* (**d**). White bars likely correspond to homozygous amplicons (F_{ij} of 90–100%), light gray bars likely correspond to heterozygous amplicons (F_{ij} of 25–75%) and black bars likely correspond to sequencing errors (Fij of 0–5%).

loci), between 719 and 834 genotypes were retained for each locus (3069 genotypes at the four MHC loci, see Table 3). At the final stage, no new variants were found suggesting that no new alleles were present in the studied population and that PCR artifacts/chimeras had been successfully removed during the post-processing.

Details on MHC allelic frequencies for the Grande Sassière population are summarized in Table 4. The number of genotypes obtained for each individual was found to significantly differ ($\chi^2 = 5330$; d.f. = 4; $P < 0.0001$) from a binomial distribution with a probability of being genotyped at a given locus of 0.67 (that is, 3069 obtained genotypes divided by 4608 genotype calls) due to an inflated number of individuals for which no genotype was obtained (257 observed vs 14 expected individuals with no genotype).

Conducting the proposed procedure without using *a priori* knowledge of the alleles leads to retain 2866 amplicons and to identify the 16 variants corresponding to the previously described alleles (Kuduk *et al.*, 2012) as potential alleles. The use of the pre-treatment does not affect the obtained results as no differences were found between these results and the ones obtained when using *a priori* knowledge of the allele.

Investigation of different sequencing error rates revealed that when the error rate was increased to 0.18, 156 additional genotypes were not assigned and 111 genotypes were found homozygous instead of heterozygous. Increasing the sequencing error rate further increased the genotyping error rate to 0.95% (13 identified erroneous genotypes) instead of 0.21% (see below for genotype validation results).

Validation of MHC genotype assignment

A total of 1450 out of 2894 assigned genotypes from the Grande Sassière population (Table 5) were used to validate the above procedure. Regarding intra-individual repeatability of 454 sequencing, no difference was found among the 30 replicated individuals (16 replicates for *Mama-UB*, 13 for *Mama-UD*, 12 for *Mama-DRB*1 and 10 for *Mama-DRB2*, all validated). Concerning genotypes obtained through 454 sequencing and Sanger sequencing, no difference was found among the 36 duplicated genotypes (9 for each studied locus). In the last validation, 1363 assigned genotypes (1016 from offspring, 161 from mothers and 186 from fathers) were used to check for genotype consistency with the parentage relationships established with microsatellites (Table 5). From a total of 1363 checked genotypes, only 3 mismatches (0.21% of all checked genotypes) within mother–father–offspring triads (Table 5) were found. Given the three validation procedures, genotypes obtained after 454 post-processing were found to be subject to a very small error rate (0.21%).

MHC sequencing cost

The expected cost of sequencing a single locus per individual was lower for 454 than for Sanger sequencing. DNA extraction costs 1.50 € (2.02 $) and PCR costs 4.50 € (6.07 $), which is the same whatever the sequencing method used. Although using numerous tagged primers necessary for amplicon-based NGS was expected to be costly, the extra-cost compared with using a single pair of primers necessary for Sanger sequencing was negligible when genotyping such a large number of individuals.

The cost of sequencing was 1.58 € (2.15 $) and 4.50 € (6.07 $), respectively, for 454 and Sanger. However, after sequencing, 33% of genotype calls could not be retrieved with 454 sequencing, which represents a real cost of 2.37 € (3.20 $) per genotype obtained without taking into account DNA extraction and PCR cost (see Table 1 to compare expected and obtained costs for similar studies).

DISCUSSION

The proposed post-processing protocol allowed us to obtain highly reliable genotypes at two MHC class I and two MHC class II loci for 863 Alpine marmots, the largest set of individuals genotyped at MHC loci to date for a non-model species (see Table 1 and Lighten *et al.*, 2014b).

Our protocol maximizes the number of retained reads and amplicons due to the use of the three first bp instead of the complete primer sequences to identify the loci and to the assignment of reads with indels to either previously known alleles or correct length variants. Moreover, the genotype assignment error rate of 0.21% attests the reliability of our method. None of the previous studies have conducted such an extensive validation of an amplicon-based NGS procedure. Consequently, the expected coverage per amplicon can be reduced to improve cost-efficiency without lowering the reliability of the obtained genotypes.

To date, no previous studies have used only part of the primer sequence, and only two recent studies (Stutz and Bolnick, 2014; Lighten *et al.*, 2014b) used indels or repeatable base-mismatch errors to optimize genotyping procedures. Interestingly, identifying a given locus based solely on the first three bps instead of the complete sequence of the primer as well as the assignment of indels allowed us to increase the number of retained amplicons. This protocol could be especially useful in future studies where users will choose a low mean coverage to optimize time and funding.

During post-processing, the differentiation of artifacts and PCR chimeras from true alleles has proven to be challenging (Babik *et al.*, 2009; Babik, 2010; Zagalska-Neubauer *et al.*, 2010; Huchard *et al.*,

Table 4 Allelic frequencies of four MHC loci for the Grande Sassière Nature Reserve population

Mama-UB *alleles*	Mama-UB *allelic freq.*	Mama-UD *alleles*	Mama-UD *allelic freq.*	Mama-DRB1 *alleles*	Mama-DRB1 *allelic freq.*	Mama-DRB2 *alleles*	Mama-DRB2 *allelic freq.*
*01	0.74	*01	0.74	*01	0.39	*01	0.45
*02	0.26	*02	0.08	*02	0.23	*02	0.39
		*03	0.19	*03	0.13	*03	0.16
				*04	0.02		
				*05	0.03		
				*06	0.07		
				*07	0.09		
				*08	0.04		
Nind	754		712		664		665

Abbreviations: freq., frequency; MHC, major histocompatibility complex; nind, number of individuals.

Table 5 Validation of 1363 assigned genotypes after 454 post-processing by checking mother–father–offspring triads among all MHC studied loci for the Grande Sassière Nature Reserve

Locus	*Number of checked triads*			*Total offspring genotypes*	*Total mother genotypes*	*Total father genotypes*
	Homozygous parents, same alleles ♀AA and ♂AA Offspring [a]AA	*Homozygous parents, different alleles ♀AA and ♂BB Offspring [a]AB*	*At least one heterozygous parent ♀AB and ♂CD Offspring [a]AC/AD/BC/BD*			
Mama-UB	106	10	195[b(1)]	311	48	55
Mama-UD	79	0	207[b(2)]	286	44	53
Mama-DRB1	2	0	210	212	33	37
Mama-DRB2	13	8	186	207	36	41
TOTAL	200	18	798	106	161	186

Abbreviation: MHC, major histocompatibility complex.
♀: An example of mother's genotype; ♂: an example of father's genotype.
[a]An example of pup's possible genotype.
[b(N)]Number of non-matching genotypes.

2012; Sommer *et al.*, 2013). The absence of artifacts and PCR chimeras after removing singletons in our study could be the result of multiples causes. First, the whole set of reads obtained during the 454 sequencing process was found to be of high quality (see Supplementary Data S2 for a quality diagnostic), such that substitutions were likely very rare. Second, because we amplified each locus separately, the opportunity for production of PCR chimeras was largely reduced (that is, there could have been at most two allelic variants/amplicon, and only one variant in homozygous amplicons). However, avoiding production of PCR chimeras is not always possible as, for instance, in the case of recently duplicated loci where the same primers amplify several loci and the loci cannot be separated (that is, Babik *et al.*, 2009; Castillo *et al.*, 2010; Cammen *et al.*, 2011; Oomen *et al.*, 2013). Additionally, the moderate number of PCR cycles we used should have further reduced production of PCR chimeras (Lenz and Becker, 2008). Finally, the low coverage per amplicon found in this study might have decreased the probability of finding both artifacts and chimeras after post-processing.

The extensive validation we conducted demonstrated the accuracy (genotype assignment error rate of 0.21%) of the proposed protocol for amplicon-based next generation post-processing. According to previous studies using amplicon-based NGS (Babik *et al.*, 2009; Galan *et al.*, 2010; Huchard *et al.*, 2012; Oomen *et al.*, 2013), the use of a stringent post-processing procedure has demonstrated its usefulness in disentangling sequencing artifacts and chimeras from true alleles. Moreover, the validation of genotype assignment by confrontation with known parentage relationships has proven to be useful since, among the 47% of assigned genotypes that could be confronted with parentage relationships, 99.79% were validated. Consequently, such validation should be recommended whenever parentage relationships are available.

High percentages of genotype assignment were obtained in most studies using amplicon-based NGS techniques for MHC typing (Table 1), with the exception of the study of Sepil *et al.* (2012) (56% of genotype assignment) and our study (between 63% and 72% of genotype assignment). Our study was specifically designed to improve its cost efficiency, and a cost to pay was the reduced expected coverage per amplicon. This low expected coverage combined with the high variability of the coverage per amplicon may have contributed to a higher loss of assigned genotypes compared with previous studies (Table 1). Even though care was taken to equalize the amount of DNA among amplicons within pools, and the mean coverage per amplicon seemed adequate to allow genotyping most amplicons (it exceeded the minimum requirement of twelve reads per amplicon by over an order of magnitude in this study), a relatively large proportion of amplicons could not be genotyped because they did not reach the minimal number of reads.

Variability of coverage per amplicon is a well-known issue when using NGS and can result from the use of suboptimal primers (Sommer *et al.*, 2013) or from the presence of homopolymer regions in the tags (Huse *et al.*, 2007). However, our marmot-specific primers were carefully designed (Kuduk *et al.*, 2012) and the tags we used (provided in the Supplementary Table S1) were designed to minimize this problem, which was confirmed after sequencing (the number of homopolymers in the tags was not found to affect the obtained coverage per amplicon).

The variability of genomic DNA quality among samples is more likely to have been a key factor for the high loss of assigned genotypes in our study. Indeed, the inflated number of individuals for which no genotype was assigned suggests a poor genomic DNA quality of some samples. Poor DNA quality may reduce the coverage per amplicon as well as the quality of the reads, resulting in the failure to obtain genotypes for some individuals. The existence of poor DNA quality samples may be due to the differences in sample collection and/or sample preservation, which are likely to occur in studies conducted in the field over extended period of time. In the present study, the wide period of time elapsed between sample collection and sample sequencing (ranging from 1 to 21 years) may have resulted in a poor genomic DNA quality of some samples and therefore to the high number of obtained individuals with no assigned genotype. If genotyping a whole set of loci for a given individual is needed, efforts to maximize DNA quality (collection, storage and DNA extraction) should be done.

If genotyping most of the individuals is a priority, one should consider a higher coverage per amplicon (for example, Zagalska-Neubauer *et al.*, 2010; Oomen *et al.*, 2013), despite the associated increase in the cost of genotyping. Fortunately, other sequencing platforms such as Illumina now allow for sequencing the full length of functionally important MHC exons (for example, MHC class IIb exon 2 in guppies (*Poecilia reticulata*; Lighten *et al.*, 2014b) at much lower cost than 454 sequencing. Similarly, higher minimum coverage per amplicon should be used to sequence recently duplicated genes (that is, Castillo *et al.*, 2010; Cammen *et al.*, 2011; Oomen *et al.*, 2013).

When several loci are amplified with the same set of primers but the number of loci is not known in advance, the post-processing procedure could be adapted to firstly obtain the number of sequenced loci. The minimal number of sequenced loci can be determined using the visualization of the distribution of the F_{ij} (see Results section and Figure 1 for details) and exploration of the F_j (distribution of the variants for a given amplicon; Babik *et al.,* 2009). Then, the parameters of the model proposed by Galan *et al.* (2010) could be changed according to the determined number of sequenced loci or using an extension of this model proposed by Sommer *et al.* (2013), that allows to take into account variation in amplification efficiency, a common pattern when amplifying several loci with a unique pair of primers. Additionally, a modification of the maximum-likelihood framework proposed by Hohenlohe *et al.* (2010) and Etter *et al.* (2011) is required to determine homozygous and heterozygous amplicons for recently duplicated genes.

The success of our genotyping protocol stems not only from a low number of PCR and sequencing artifacts, but also from careful optimization of the primers we used (see Kuduk *et al.,* 2012 for more details). Sub-optimal primers can result in low coverage of some true allelic variants (Sommer *et al.,* 2013), and this would make it impossible to assign genotypes based on the likelihood ratio test between the most likely homozygous and the most likely heterozygous genotypes. The fact that we found perfect, or nearly perfect, congruence in all three of our validation procedures confirms that poorly fitting primers were not an issue in our system. Although investment in the careful design of primers may often be substantial, it is a well spent effort that is more than compensated during large-scale genotyping given that methods similar to ours can be applied to any system for which a given pair of primers amplifies a given number of loci. However, these methods may not be suitable for species with extensive copy-number variation (that is, Babik *et al.,* 2009; Radwan *et al.,* 2012).

No new alleles were found for any of the four MHC loci despite our extensive sampling effort, indicating that the sample of 38 Alpine marmots studied by Kuduk *et al.* (2012) was enough to capture all MHC alleles in these populations (MHC class I loci: two alleles for *Mama-UB* and three alleles for *Mama-UD*; MHC class II loci: eight alleles for *Mama-DRB1* and three alleles for *Mama-DRB2*). Although similar levels of allelic MHC diversity were found in wolverine (two and eight alleles were found in two MHC class II loci (Oomen *et al.,* 2013), the low allelic MHC diversity in Alpine marmots contrasts with higher allelic diversity generally found in mammals. For example, 20 alleles at the MHC class I loci were found in woodchuck (*Marmota monax*; Yang *et al.,* 2000; Zhou *et al.,* 2003) and 7 in the European bison (*Bison bonasus*; Babik *et al.,* 2012). Concerning MHC class II loci, between 7 and 38 alleles were found in 10 rodent species (Goüy de Bellocq *et al.,* 2008), and 60 and 61 in two loci in Gray mouse lemur (*Microcebus murinus*; Huchard *et al.,* 2012). In addition to poor MHC diversity, low levels of genetic variability in Alpine marmots have been previously found using allozymes (Arnold, 1990; Preleuthner and Pinsker, 1993), minisatellites (Arnold, 1990; Rassmann *et al.,* 1994; Kruckenhauser *et al.,* 1997) and microsatellites (Cohas *et al.,* 2009, but see Goossens *et al.,* 2001) and this in different populations spread across the whole Alpine arc. Alpine marmots were widespread in Europe until the early Holocene (Couturier, 1955; Besson, 1971). However, several drastic reductions of their population size possibly due to the environmental variations (Preleuthner and Pinsker, 1993; Rassmann *et al.*, 1994) occurred during the early Holocene and reduced the distribution of this species to the Alpine arc and the Carpathian Mountains, its actual natural distribution. Such a low observed genetic diversity found at several genetic markers (MHC, allozymes, minisatellites and microsatellites) can be the result of bottlenecks occurring during Holocene. Moreover, small population size may have favored inbreeding that in turn has reduced genetic variability.

Our results emphasize the efficiency of amplicon-based NGS combined with adequate post-processing and validation procedures to obtain large numbers of accurately assigned genotypes from any species and genes in the absence of extensive copy-number variation. In conclusion, amplicon-based NGS and post-processing open new possibilities for research on the evolution of MHC and other highly polymorphic functional genes that require accurate large scale assigned genotypes at the individual level.

DATA ARCHIVING

DNA reads are accessible upon request on the Laboratoire de Biologie et Biométrie Evolutive data storage website (http://umr5558-bddeco.univ-lyon1.fr). All Python codes are freely available at https://github.com/tbigot/alFinder. Genotyping data are available from the Dryad Digital Repository: http://doi.org/10.5061/dryad.rp7n9.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Professor Jacek Radwan for helpful discussions and comments on former version of the manuscript. We are grateful to all students and Earthwatch volunteers involved in catching the marmots and the authorities of the Vanoise National Park for granting us permission to work in the Grande Sassière Nature Reserve. We thank Irene Figueroa, Bernat Claramunt-López, Bruno Bassano and Denis Réale for providing genetic samples from Catalan Pyrenees and Italian Alps. We thank Aurélie Johanet and Audrey Second for their helpful collaboration in 454 and Sanger sequencing, respectively. We thank Micki Harrington, Mark Hewison and Murray Patterson for English editing. This work has been greatly improved by three anonymous referees who made insightful suggestions on a previous version of the manuscript. Financial support was received from the Agence Nationale de la Recherche (ANR, project ANR-08-BLAN-0214-03 and project ANR-13-JSV7-0005), the 'FR41 BioEnvironnement et Santé de l'Université de Lyon', the Centre National de la Recherche Scientique (CNRS) and Earthwatch Institute. MFR is supported by two scholarships for postgraduate studies (Obra Social Fundació 'La Caixa' and VetAgro Sup).

Akira S, Takeda K, Kaisho T (2001). Toll-like receptors: critical proteins linking innate and acquired immunity. *Nat Immunol* **2**: 675–680.

Arnold W (1990). The evolution of marmot sociality. I. why disperse late. *Behav Ecol Sociobiol* **27**: 229–237.

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009). New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Mol Ecol Resour* **9**: 713–719.

Babik W (2010). Methods for MHC genotyping in non-model vertebrates. *Mol Ecol Resour* **10**: 237–251.

Babik W, Kawalko A, Wojcik JM, Radwan J (2012). Low major histocompatibility complex class I (MHC I) variation in the European bison (*Bison bonasus*). *J Hered* **103**: 349–359.

Besson JP (1971). Introduction de la marmotte dans les Pyrénées occidentales. *C.R. du 96ème Congrès des Sociétés Savantes, Toulouse* **vol 3**, 397–399.

Cammen K, Hoffman JI, Knapp LA, Harwood J, Amos W (2011). Geographic variation of the major histocompatibility complex in eastern atlantic grey seals (*Halichoerus grypus*). *Mol Ecol* **20**: 740–752.

Castillo S, Srithayakumar V, Meunier V, Kyle CJ (2010). Characterization of major histocompatibility complex (MHC) DRB exon 2 and DRA exon 3 fragments in a primary terrestrial rabies vector (*Procyon lotor*). *PLoS One* **5**: e12066.

Charlesworth D, Awadalla P (1998). Flowering plant self-incompatibility: the molecular population genetics of Brassica S-loci. *Heredity* **81**: 1–9.

Cohas A, Yoccoz NG, Da Silva A, Goossens B, Allainé D (2006). Extra-pair paternity in the monogamous alpine marmot (*Marmota marmota*): the roles of social setting and female mate choice. *Behav Ecol Sociobiol* **59**: 597–605.

Cohas A, Yoccoz NG, Bonenfant C, Goossens B, Genton C, Galan M *et al.* (2008). The genetic similarity between pair members influences the frequency of extrapair paternity in alpine marmots. *Anim Behav* **76**: 87–95.

Cohas A, Bonenfant C, Kempenaers B, Allainé D (2009). Age-specific effect of heterozygosity on survival in alpine marmots. *Marmota marmota. Mol Ecol* **18**: 1491–1503.

Couturier MAJ (1955). Acclimatation et acclimatement de la Marmotte des Alpes, *Marmota marmota marmota* (Linné 1758), dans les Pyrénées françaises. *Saugetierk Mitt* **3**: 105–107.

Da Silva A, Luikart G, Allainé D, Gautier P, Taberlet P, Pompanon F (2003). Isolation and characterization of microsatellites in European alpine marmots (*Marmota marmota*). *Mol Ecol Notes* **3**: 189–190.

Dunn PO, Bollmer JL, Freeman-Gallant CR, Corey R, Whittingham LA (2013). MHC variation is related to a sexually selected ornament, survival, and parasite resistance in common yellowthroats. *Evolution* **67**: 679–687.

Ellis JS, Turner LM, Knight ME (2012). Patterns of selection and polymorphism of innate immunity genes in bumblebees (*Hymenoptera: Apidae*). *Genetica* **140**: 205–217.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: Orgogozo V, Rockman MV (eds). *Molecular methods for evolutionary genetics*. Humana Press: TotowaVol 772, pp 157–178.

Feng DF, Doolittle RF (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **25**: 351–360.

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**: 296.

Glenn TC (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**: 759–769.

Goossens B, Chikhi L, Taberlet P, Waits LP, Allainé D (2001). Microsatellite analysis of genetic variation among and within Alpine marmot populations in the French Alps. *Mol Ecol* **10**: 41–52.

Goüy de Bellocq J, Charbonnel N, Morand S. (2008). Coevolutionary relationship between helminth diversity and MHC class II polymorphism in rodents. *J Evolution Biol* **21**: 1144–1150.

Grueber CE, Wallis GP, King TM, Jamieson IG (2012). Variation at innate immunity Toll-Like Receptor genes in a bottlenecked population of a New Zealand robin. *PLos One* **7**: e45011.

Hanslik S, Kruckenhauser L (2000). Microsatellite loci for two European sciurid species (*Marmota marmota, Spermophilus citellus*). *Mol Ecol* **9**: 2163–2165.

Hedrick P (1994). Evolutionary genetics of the major histocompatibility complex. *Am Nat* **143**: 945–964.

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

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**: e1000862.

Huchard E, Albrecht C, Schliehe-Diecks S, Baniel A, Roos C, Peter PM *et al.* (2012). Large-scale MHC class II genotyping of a wild lemur population by next generation sequencing. *Immunogenetics* **64**: 895–913.

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

Jones J (2001). Putting knowledge of plant disease resistance genes to work. *Curr Opin Plant Biol* **4**: 281–287.

Jørgensen MH, Lagesen K, Mable BK, Brysting AK (2012). Using high-throughput sequencing to investigate the evolution of self-incompatibility genes in the *Brassicaceae*: strategies and challenges. *Plant Ecol Divers* **5**: 473–484.

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

Klinkicht M (1993). *Untersuchungen zum Paarungssystem des Alpenmurmeltiers, Marmota marmota mittels DNA Fingerprinting*. Ph.D. thesis University of Munich.

Kruckenhauser L, Miller W, Preleuthner M, Pinsker W (1997). Differentiation of Alpine marmot populations traced by DNA fingerprinting. *J Zool Syst Evol Res* **35**: 143–149.

Kuduk K, Johanet A, Allainé D, Cohas A, Radwan J (2012). Contrasting patterns of selection acting on MHC class I and class II DRB genes in the Alpine marmot (*Marmota marmota*). *J Evolution Biol* **25**: 1686–1693.

Lighten J, van Oosterhout C, Bentzen P (2014a). Critical review of NGS analyses for de novo genotyping multigene families. *Mol Ecol* **23**: 3957–3972.

Lighten J, van Oosterhout C, Paterson IG, Mcmullan M, Bentzen P (2014b). Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Mol Ecol Resour* **14**: 753–767.

Lenz TL, 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**: 117–123.

Oomen RA, Gillett RM, Kyle CJ (2013). Comparison of 454 pyrosequencing methods for characterizing the major histocompatibility complex of nonmodel species and the advantages of ultra deep coverage. *Mol Ecol Resour* **13**: 103–116.

Preleuthner M, Pinsker W (1993). Depauperate gene pools in *Marmota m. marmota* are caused by an ancient bottleneck: electrophoretic analysis of wild populations from Austria and Switzerland. *Acta theriol* **38**: 121–139.

Promerová M, Babik W, Bryja J, Poláková R, Schnitzer J, Munclinger P *et al.* (2012). Evaluation of two approaches to genotyping major histocompatibility complex class I in a passerine-CE-SSCP and 454 pyrosequencing. *Mol Ecol Resour* **12**: 285–292.

Radwan J, Zagalska-Neubauer M, Cichon M, Sendecka J, Kulma K, Gustafsson L *et al.* (2012). MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Mol Ecol* **21**: 2469–2479.

Rassmann K, Arnold W, Tautz D (1994). Low genetic-variability in a natural alpine marmot population (*Marmota marmota, Sciuridae*) revealed by dna-fingerprinting. *Mol Ecol* **3**: 347–353.

Schlipf NA, Schule R, Klimpe S, Karle KN, Synofzik M, Schicks J *et al.* (2011). Amplicon-based high-throughput pooled sequencing identifies mutations in CYP7B1 and SPG7 in sporadic spastic paraplegia patients. *Clin Genet* **80**: 148–160.

Sepil I, Moghadam HK, Huchard E, Sheldon BC (2012). Characterization and 454 pyrosequencing of major histocompatibility complex class I genes in the great tit reveal complexity in a passerine system. *BMC Evol Biol* **12**: 68.

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012). Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* **21**: 1794–1805.

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

Stutz WE, Bolnick DI (2014). Stepwise threshold clustering: a new method for genotyping MHC loci using next-generation sequencing technology. *PLoS One* **9**: e100587.

Vasemägi A, Primmer CR (2005). Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Mol Ecol* **14**: 3623–3642.

Wu J, Saupe S, Glass N (1998). Evidence for balancing selection operating at the het-c heterokaryon incompatibility locus in a group of filamentous fungi. *P Natl Acad Sci USA* **95**: 12398–12403.

Yang D, Lu M, Hao L, Roggendorf M (2000). Molecular cloning and characterization of major histocompatibility complex class I cDNAs from woodchuck (*Marmota monax*). *Tissue Antigens* **55**: 548–557.

Zagalska-Neubauer M, Babik W, Stuglik M, Gustafsson L, Cichon M, Radwan J (2010). 454 sequencing reveals extreme complexity of the class II major histocompatibility complex in the collared flycatcher. *BMC Evol Biol* **10**: 395.

Zhou J, Ferencik S, Rebmann V, Yang DL, Lu M, Roggendorf M *et al.* (2003). Molecular genetic and biochemical analysis of Woodchuck (*Marmota monax*) MHC class I polymorphism. *Tissue Antigens* **61**: 240–248.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)