Received: 13 September 2017 | Revised: 30 April 2019 | Accepted: 6 May 2019

DOI: 10.1111/1755-0998.13036

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Assigning alleles to different loci in amplifications of duplicated loci

Kang Huang[1] | Pei Zhang[1] | Derek W. Dunn[1] | Tongcheng Wang[1] | Rui Mi[1] | Baoguo Li[1,2]

[1]Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences, Northwest University, Xi'an, China

[2]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

Correspondence
Baoguo Li, Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences, Northwest University, Xi'an 710,069, China.
Email: baoguoli@nwu.edu.cn

Funding information
Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/Award Number: XDB310200000; National Natural Science Foundation of China, Grant/Award Number: 31730104, 31770411, 31572278 and 31770425; Young Elite Scientists Sponsorship Program by CAST, Grant/Award Number: 2017QNRC001; National Key Programme of Research and Development, the Ministry of Science and Technology of China, Grant/Award Number: 2016YFC0503202; Shaanxi Province Talents 100 Fellowship; Natural Science Basic Research Plan in Shaanxi Province of China, Grant/Award Number: 2018JM3024 and 2019JM258

Abstract

Duplicated loci, for example those associated with major histocompatibility complex (MHC) genes, often have similar DNA sequences that can be coamplified with a pair of primers. This results in genotyping difficulties and inaccurate analyses. Here, we present a method to assign alleles to different loci in amplifications of duplicated loci. This method simultaneously considers several factors that may each affect correct allele assignment. These are the sharing of identical alleles among loci, null alleles, copy number variation, negative amplification, heterozygote excess or heterozygote deficiency, and linkage disequilibrium. The possible multilocus genotypes are extracted from the alleles for each individual and weighted to estimate the allele frequencies. The likelihood of an allele configuration is calculated and is optimized with a heuristic algorithm. Monte-Carlo simulations and three empirical MHC data sets are used as examples to evaluate the efficacy of our method under different conditions. Our new software, MHC-TYPER V1.1, is freely available at https://github.com/huangkang1987/mhc-typer.

KEYWORDS
allele assignment, duplicated loci, likelihood estimation, major histocompatibility complex, simulate annealing algorithm

1 | INTRODUCTION

Gene duplication is defined as any duplication of a region of DNA that contains a gene and plays a vital role in the evolution of the genome and in the creation of new functional genes (Maruyama & Takahata, 1981). Due to the similarity in DNA sequences of duplicated loci, a pair of primers can sometimes coamplify several loci. Examples of gene duplication include microsatellites (Detwiler & Criscione, 2011), aldolase-B genes (Quattro, Jones, & Oswald, 2001), olfactory receptor (OR) genes (Ziegler et al., 2000), killer immunoglobulin-like receptor (KIR) genes (Murdoch, Seoud, Kircheisen, Mazhar, & Slim, 2006), homeobox genes (Bauer et al., 2004) and major histocompatibility complex (MHC) genes (Gaigher et al., 2018). Unless locus-specific primers can be redesigned (e.g., Shilling et al., 2002; Quattro et al., 2001), this may result in difficulties in genotyping and inaccurate analyses.

The MHC gene family is a well-known gene family that has arisen via gene duplication from a single ancestral gene and is one of the most polymorphic gene regions yet found in any vertebrate genome (Eizaguirre, Lenz, Kalbe, & Milinski, 2012). The MHC gene family plays a central role in the immune systems of vertebrates. MHC molecules first recognize foreign antigens, then bind and finally present them to T cells, thus triggering an appropriate immune response (Klein, 1986).

Huang and Zhang contributed equally to this work.

There are four major features of MHC genes that may make genotyping difficult. First, extreme allelic polymorphism prevents the designing of efficient primers to detect genetic variation within the MHC (except when extensive information of primer binding regions of studied populations are available) (Babik, 2009). Second, similarities among different loci within the MHC makes it difficult to design locus-specific primers (Biedrzycka, Sebastian, Migalska, Westerdahl, & Radwan, 2016). Third, frequent gene duplications and variation in the number of loci within and among species can result in ambiguous genotyping (Piertney & Oliver, 2005). Finally, the presence of pseudogenes makes it difficult to identify any functional variation (Zimmerman, Carrington, & Nutman, 1993).

Several studies have used the number of alleles that individuals possess to estimate variation of MHC loci within individuals, as well as the proportion of shared alleles to estimate pairwise similarity between individuals when alleles cannot be assigned to a particular locus (e.g., Gaigher et al., 2018; Huchard, Knapp, Wang, Raymond, & Cowlishaw, 2010; Miller, Moore, Nelson, & Daugherty, 2009; Santos, Michler, & Sommer, 2017; Strandh et al., 2012). However, these methods are generally inaccurate and are largely influenced by the polymorphism of loci or genetic diversity within a population. For example, at a monomorphic locus or in a population with an allele that has been fixed, any two individuals may share all alleles. In the opposite scenario, at a highly polymorphic locus or in a population with a high genetic diversity, two individuals will only share a small proportion of alleles. Therefore, different loci and populations with different genetic structure will have different distributions of similarity indices, and these indices cannot be compared among different loci and populations.

If alleles are assigned to the correct loci, the genotypes at each locus become available, several parameters with clear biological meanings are then available to evaluate the genetic background among/within individuals, for example by calculating the relatedness coefficient (Huang et al., 2016; Lynch & Ritland, 1999; Wang, 2002), kinship coefficient (Ritland, 1996), individual inbreeding coefficient (Hardy & Vekemans, 2002), and genetic distance (Nei, 1972). Some statistical tests can also be applied to test specific hypotheses, e.g., balancing selection can be tested via a heterozygote excess test (Rousett, 2008), and the presence of nonrandom mating can be tested by evaluating whether the genotypes between mates are more dissimilar than average (permutation; Guo, Huang, Ji, Garber, & Li, 2015). Classical population genetics methods can also be applied to particular scientific problems, e.g., analysis of molecular variance (Excoffier, Smouse, & Quattro, 1992); Bayesian clustering (Pritchard, Stephens, & Donnelly, 2000), and *F*-statistics (Weir & Cockerham, 1984).

Previous methods of allele assignment have been performed by haplotype phasing based on next generation sequencing (NGS) data (Selvaraj, Schmitt, Dixon, & Ren, 2015; Stuglik, Radwan, & Babik, 2011), which requires high-quality tissue samples. This method thus has limited potential in its application to many studies of wild populations, especially endangered animals. Moreover, haplotype phasing ignores any recombination among loci. Therefore, if any amplified loci are located on different chromosomes or are distant to each other (e.g., >4 Mb), this method cannot be used.

The phenotype of an individual ($\boldsymbol{P}$), which is defined as a set of alleles detected within an individual, contains two types of genetic information, which can both be used to assign alleles to different loci in multilocus amplification. The two genetic information types are: (a) the co-occurrence of alleles and (b) the frequencies of genotypes. By using both types of genetic information, we present a novel method for assigning the alleles in multilocus amplification into different loci directly from the phenotypes. Such an assignment can be determined by a set **A**, called the allele configuration. Each member in **A** is a set at a putative locus, whose elements are the alleles assigned to this locus.

In diploid organisms, each individual should have at least one and at most two alleles at each locus to form a phenotype. Therefore, based on the co-occurrence of alleles, some incorrect allele configurations can be indicated by either one or both of two errors: the missing error (no alleles are present within a genotype) and the mismatch error (more than two alleles are present within a genotype). Although the genotypes cannot be obtained directly, the possible multilocus genotypes can be obtained by the allele configurations, and their probabilities can be calculated by estimating the allele frequencies. Based on the genotypic frequencies, the likelihood of an allele configuration can be calculated. Information from both types of genetic information are thus combined in order to evaluate the allele configuration by applying the penalties of both missing and mismatch errors, and then by calculating the Bayesian Information Criterion (BIC) (Schwarz, 1978). Our new method then uses a heuristic algorithm to find the optimal allele configuration.

There are many challenges in obtaining the correct allele configuration. These include the following five potential problems: (a) the sharing of identical alleles among loci. This will largely increase the level of complexity if included into a model, because the number of possible allele configurations is approximately equal to the square of those without the sharing of identical alleles (see Discussion). (b) Deviation from the Hardy-Weinberg Equilibrium (HWE). This can cause the genotypic frequencies to deviate from expected values and bias the likelihood estimates. The MHC loci are under balancing selection and usually exhibit an excess of heterozygotes (Aguilar et al., 2004; Zhang et al., 2018). Additional reasons for deviation from the HWE include small sample sizes, inbreeding, population subdivision, and selection (e.g., heterozygous advantage or hybrid weakness) (Hedrick & Parker, 2017). (c) The presence of null alleles or copy number variation (CNV). A null allele is an allele that cannot be amplified because of mutation, often within the primer binding site (Brookfield, 1996). Copy number variation is the difference in the numbers of loci between two different haplotypes (McCarroll & Altshuler, 2007). If we consider that any absent loci due to copy number variation as alleles with deletion mutations cover both the primer binding site and the target amplification region, then such alleles cannot be amplified and are thus equivalent to null alleles. The presence of either null alleles or CNV thus has identical consequences, and results in either missing errors or genotyping errors

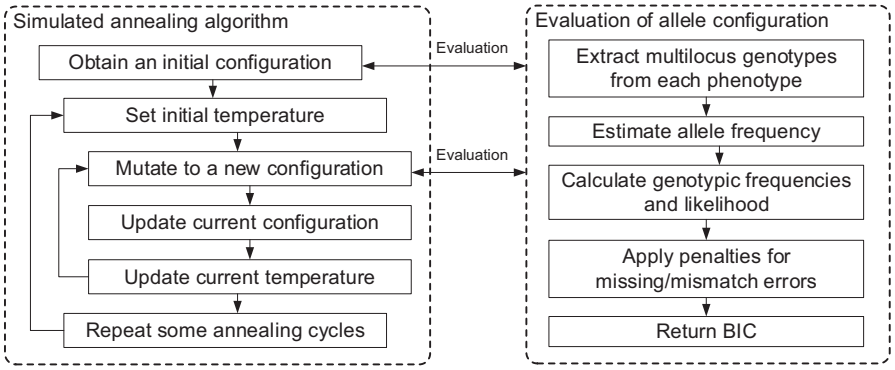

FIGURE 1 The workflow diagram of MHC-TYPER V1.1

(Wagner, Creel, & Kalinowski, 2006). (d) Linkage disequilibrium. This is a nonrandom association of alleles at different loci (Slatkin, 2008). In this case, the genotype at a single locus is correlated with that at the other locus, which also biases the likelihood estimation. Moreover, if any loci are under complete linkage, correct assignment cannot be obtained (see Discussion). (e) Negative amplification. This is failure of amplification at a locus for a reason other than the locus being homozygous for a null allele (e.g., due to poor quality DNA or experimental errors) (Kalinowski & Taper, 2006). Negative amplification can also cause missing errors, as well as null alleles and CNV.

We also performed Monte-Carlo simulations to calculate the correct assignment rate and evaluate our method using empirical data from natural populations from the North Atlantic Right Whale (*Eubalaena glacialis*) (Gillett, Murray, & White, 2013), the Wolf (*Canis lupus*) (Galaverni, Caniglia, Fabbri, Lapalombella, & Randi, 2013) and the Barn Owl (*Tyto alba*) (Gaigher et al., 2018). Our methods rely on a new software package, MHC-TYPER V1.1, that we have developed and made freely available to other researchers. The program is written in C++ and C#, includes a user-friendly graphical interface (GUI), and can be run on Microsoft Windows.

2 | MATERIALS AND METHODS

2.1 | Rationale

Here, we assume that *L* loci are coamplified with a pair of primers, *k*' amplifiable alleles are obtained from *n* different individuals, and all individuals are nonrelatives and are diploid. All loci are assumed to be under linkage equilibrium and the heterozygotes at a locus may exhibit either deficiency or excess. Copy number variation (CNV) may or may not be present. If present, CNV is treated as a null allele. Negative amplification is also considered. We also assume that a genotype is a set consisting of amplifiable alleles at a locus within an individual, and a multilocus genotype is a combination of genotypes at multiple loci.

Some impossible configurations can be directly excluded without evaluation. Given an allele configuration **A**, if each putative locus has at least one amplifiable allele and each amplifiable allele is assigned to at least one putative locus, then **A** is termed possible. If it is assumed that there are no identical alleles shared among loci, then the second condition should be revised as each amplifiable allele is assigned to exactly one putative locus. In this paper, we only consider these possible allele configurations.

A schematic diagram of the algorithm of MHC-TYPER V1.1 is shown in Figure 1. A simulated annealing algorithm (Bertsimas & Tsitsiklis, 1993) is used to find the optimal allele configuration. This algorithm is a probabilistic technique for approximating the global optimum.

2.2 | Searching algorithm

Assuming we can evaluate the allele configuration, the remaining problem is to search for the optimal allele configuration. This problem therefore becomes a hill-climbing problem, the goal of which is to find the highest peak in the solution space. We use a simulated annealing algorithm (Bertsimas & Tsitsiklis, 1993) to find the optimal allele configuration. This algorithm starts from a random initial allele configuration, and repeatedly generates a new allele configuration based on the current allele configuration. The initial allele configuration is not correlated with the final allele configuration and can be set arbitrarily.

Four mutation modes are used to generate a new allele configuration. These are: (a) swap (exchange a pair of amplifiable alleles between two putative loci), (b) move (move an amplifiable allele from one putative locus that contains at least two amplifiable alleles to another), (c) insert (insert a new amplifiable allele to a locus) and (d) delete (remove an amplifiable allele from a polymorphic locus). Some constraints should be made to ensure consistency: each amplifiable allele is present at more than one putative locus, and there is more than one amplifiable allele at each putative locus. The modes "swap" or "move" do not change the sum of amplifiable alleles across all putative loci, whilst the modes "insert" or "delete" will. The probability of each of the four modes is set at 0.4, 0.4, 0.1 and 0.1, respectively. These are empirical values and different probabilities can also work but may be slower or faster. The putative loci can also be assumed to not share identical alleles, and if so, each amplifiable allele can only be present at one putative locus in the initial allele configuration and only 'swap' and 'move' can be used. The number of putative loci remain unchanged during optimization.

The allele configuration is evaluated by BIC, which is a criterion for model selection among a finite set of models with the model having the lowest BIC being preferred. If the new allele configuration outperforms the current configuration (has a lower BIC), then it will

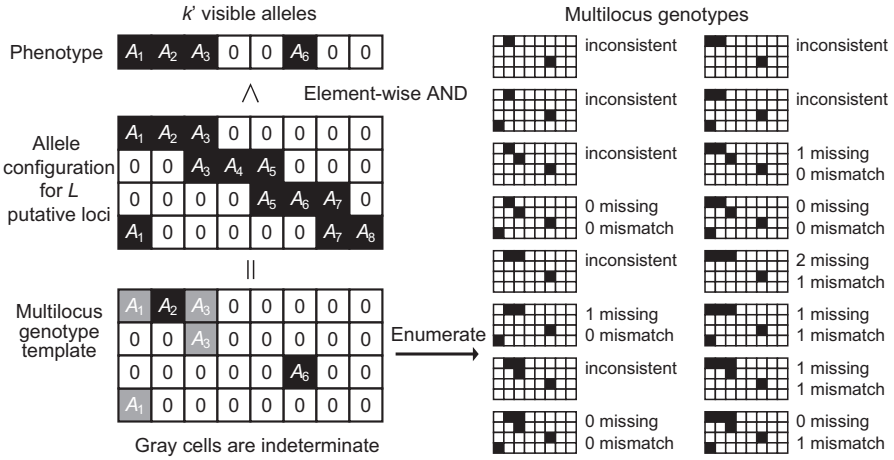

FIGURE 2 A schematic diagram of the extraction of multilocus genotypes. The symbol is 0 in white cells and A_m in black cells at the m^{th} column. The positions at grey cells are indeterminate

be directly adopted as the current configuration. Otherwise, it is adopted according to an acceptance rate given by.

$$P_{\text{accept}} = \exp\left(\frac{E_c - E_n}{E_b T}\right), \quad (1)$$

where E_n, E_c and E_b are the BICs of the new, current and current best allele configurations, respectively, and T is the current temperature. From Equation (1), it can be inferred that the acceptance rate depends on the difference in BIC between current and new allele configurations, as well as the current temperature. The temperature describes the activeness of current state. Under the same conditions, the same "worse" allele configuration will be accepted at a higher rate if the temperature is higher.

The searching algorithm randomly "walks" across different allele configurations (accepting a worse allele configuration at a high probability) with high temperatures and becomes increasingly "greedy" (accepting a worse allele configuration at a lower probability) as the temperature decreases. The simulated annealing algorithm takes advantage of this feature to avoid being confined to a suboptimal local maxima. It does this by simulating the annealing process of a metal, starting from a higher initial temperature, and then gradually decreasing by multiplying T by an annealing coefficient of less than one (e.g., 0.99). At each temperature, it repeatedly finds the new allele configuration hundreds of times. The initial temperature should be sufficiently high to allow worse allele configurations to be accepted, because the best allele configuration may be adjacent to any of these; the final temperature should be sufficiently low to allow the algorithm to become "greedy" enough to complete the cycle. The annealing cycle is repeated several times to ensure that the optimal allele configuration is found.

In order to ensure that the allele configurations are continuous, an invalid allele configuration (encounter mismatch and missing errors, see the subsection "Multilocus genotype" for the definition) can also be evaluated and adopted as the current allele configuration. Clearly, some penalties need to be applied to the evaluation of such an invalid configuration. This enables the searching algorithm to compare the evaluations between different allele configurations even if they are invalid, and to travel across invalid allele configurations to reach the optimal allele configuration.

Dynamic programming is applied to accelerate the searching algorithm, which stores the information of each allele configuration (e.g., likelihood, allele frequencies and BIC) into a hash table to prevent repeated calculations. When a new allele configuration has been generated, the searching algorithm will first look-up this configuration in the hash table. If it is outside the hash table, the searching algorithm will perform the evaluation procedures.

2.3 | Multilocus genotype

For the evaluation of an allele configuration **A**, the multilocus genotypes under **A** are first extracted from each phenotype. The extraction procedure is illustrated in Figure 2.

For calculation convenience, we let $A_1, A_2, \cdots, A_{k'}$ represent all k' amplifiable alleles, and regard a phenotype $\mathcal{P}$ as a $1 \times k'$ matrix $(\mathcal{P}_1, \mathcal{P}_2, \cdots \mathcal{P}_{k'})$, whose m^{th} element $\mathcal{P}_m$ is either A_m (if the m^{th} amplifiable allele is present) or 0 (otherwise), $m = 1,2, \cdots, k'$. The symbol 0 is treated as nothing, i.e., empty. Similarly, an allele configuration **A** consisting of L members will also be regarded as an $L \times k'$ matrix, whose l^{th} row $\mathbf{A}_l$ is the l^{th} member of **A** (i.e., the set consists of the amplifiable alleles assigned to the l^{th} putative locus), $l = 1,2, \cdots, L$. Similar to the phenotype $\boldsymbol{\mathcal{P}}$, the member $\mathbf{A}_l$ is regarded as a $1 \times k'$ matrix $(A_{l1}, A_{l2}, \cdots, A_{lk'})$, whose m^{th} element A_{lm} is either A_m or 0, $m = 1,2, \cdots, k'$.

We will first perform the element-wise AND operation for **A** and $\boldsymbol{\mathcal{P}}$, resulting in an $L \times k'$ matrix **C**, whose lm^{th} element C_{lm} is defined as

$$C_{lm} = A_{lm} \wedge \mathcal{P}_m$$
$$= \begin{cases} A_m \text{ if } A_{lm} = A_m \text{ and } \mathcal{P}_m = A_m, \\ 0 \text{ if } A_{lm} = 0 \text{ or } \mathcal{P}_m = 0, \end{cases}$$
$$l = 1,2, \cdots, L, m = 1,2, \cdots, k'.$$

The alleles in the l^{th} row of **C** denote the candidate alleles for the genotypes at the l^{th} putative locus. If neither null alleles nor negative amplification are taken into account, then the empty genotype cannot appear. Hence, under such conditions, when a row contains only one allele, this allele is definitively present within the genotype at the l^{th} putative locus.

It is clear that each element in the m^{th} column is either A_m or 0, $m = 1,2,\cdots,k'$. If A_m appears only once (i.e., it is located on the l^{th} row), then the allele at the lm^{th} position can be determined, namely A_m is definitively present within the genotype at the l^{th} putative locus (otherwise, this will cause an error that the set union of genotypes at the whole L putative loci cannot be the observed phenotype $\boldsymbol{\mathcal{P}}$, termed the *inconsistent error*). If A_m appears more than once (i.e. it is located on the l_1^{th} and l_2^{th} rows), then the allele at the l_1m^{th} or l_2m^{th} position cannot be determined. In other words, A_m can be either present or absent in the genotype at the l_1^{th} or l_2^{th} putative locus, because this does not cause an inconsistent error so long as A_m appears at least once. The matrix **C** is used as a template to generate the multilocus genotypes. Assuming that the number of all indeterminate positions in **C** is η, and that the inconsistent error is allowed to appear, then there are 2^{η} multilocus genotypes. Each can also be regarded as an $L \times k'$ matrix, denoted by **G**, and can be straightforwardly enumerated (refer to Figure 2).

For a multilocus genotype **G**, except when errors are inconsistent, there may be either one or both of two error types present: (a) more than two alleles are present within a genotype in **G** (mismatch error) or (b) no alleles are present within a genotype in **G** (missing error) if neither null alleles nor negative amplification are considered. The multilocus genotypes with mismatch or inconsistent errors will be directly excluded. If both null alleles and negative amplification are not considered, the multilocus genotypes with missing errors will also be excluded.

Assuming that there exists a multilocus genotype **G** without these three errors, such that **G** can be extracted from a phenotype $\boldsymbol{\mathcal{P}}$ under an allele configuration **A**, then **G** is called possible and $\boldsymbol{\mathcal{P}}$ is called valid. Moreover, if the whole phenotypes are valid under **A**, we call **A** valid. The possible multilocus genotypes extracted from $\boldsymbol{\mathcal{P}}$ will be recorded and used for allele frequency estimation and likelihood calculation.

The value v as a penalty for $\boldsymbol{\mathcal{P}}$ is defined as the minimum of the penalties for all consistent multilocus genotypes from $\boldsymbol{\mathcal{P}}$, i.e.

$$v = \min\left\{ v_j' \middle| v_j' = n_{j,\text{missing}} p_{\text{missing}} + n_{j,\text{mismatch}} p_{\text{mismatch}} \right\},$$

where v_j' is the penalty value for the j^{th} consistent multilocus genotype $\mathbf{G}_j$ from $\boldsymbol{\mathcal{P}}$, $n_{j,\text{missing}}$ is the number of missing errors (i.e. the number of empty genotypes in $\mathbf{G}_j$), $n_{j,\text{mismatch}}$ is the number of mismatch errors (i.e. the number of excess alleles in $\mathbf{G}_j$), and p_{missing} and p_{mismatch} are the corresponding values of penalties. It is clear that if $\boldsymbol{\mathcal{P}}$ is valid, then $v = 0$.

The numbers $n_{\mathcal{P},\text{missing}}$ and $n_{\mathcal{P},\text{mismatch}}$ for $\boldsymbol{\mathcal{P}}$ are those for the multilocus genotype with the minimal penalty, and the number of missing errors (or mismatch errors) for **A** is the sum of $n_{\mathcal{P},\text{missing}}$ (or $n_{\mathcal{P},\text{mismatch}}$) across all phenotypes $\boldsymbol{\mathcal{P}}$.

2.4 | Likelihood and BIC

The likelihood of an allele configuration **A** can be calculated after the allele frequencies at these L putative loci are estimated, which will be used to evaluate the configuration **A**. Given a hypothesis H, the likelihood $\mathcal{L}$ is defined as the probability of some observed data D given the hypothesis H, written as $\mathcal{L} = \Pr(D|H)$. Now, if the hypothesis H is that the correct allele configuration is **A**, and if the data are all phenotypes observed, then $\mathcal{L}$ can be expressed as:

$$\mathcal{L} = \prod_i \Pr\left(\boldsymbol{\mathcal{P}}_i|A\right) = \prod_i \sum_j \Pr\left(G_{ij}\right) = \prod_i \sum_j \prod_l \Pr\left(G_{ijl}\right),$$

where $\boldsymbol{\mathcal{P}}_i$ is the phenotype of the i^{th} individual, $\mathbf{G}_{ij}$ is the j^{th} possible multilocus genotype derived from $\boldsymbol{\mathcal{P}}_i$, G_{ijl} is the genotype of $\mathbf{G}_{ij}$ at the l^{th} locus, and $\Pr(G_{ijl})$ is the frequency of G_{ijl}. Because all n individuals are assumed to be nonrelatives, their phenotypes are mutually independent, and so the probability $\mathcal{L}$ is the product of the probabilities $\Pr(\mathcal{P}_i|\mathbf{A})$, $i = 1,2,\cdots,n$. Similarly, because all L loci are assumed to be under linkage equilibrium, the probability $\Pr(\mathbf{G}_{ij})$ is the product of the frequencies $\Pr(G_{ijl})$, $i = 1,2,\cdots,L$. Besides, because different multilocus genotypes are mutually exclusive, the probability $\Pr(\mathcal{P}_i|\mathbf{A})$ is the sum of the probabilities of all possible multilocus genotypes from $\boldsymbol{\mathcal{P}}_i$ under **A**.

In order to accommodate invalid allele configurations, the penalties for missing and mismatch errors are applied to our likelihood. Therefore, the searching algorithm is able to identify the allele configuration that is less "invalid" (e.g., with fewer missing or mismatch errors), i.e., more possible. Let $\mathcal{L}'$ be the value after any penalties have been applied to the likelihood $\mathcal{L}$ of **A**. Then the logarithm of $\mathcal{L}'$ is

$$\ln \mathcal{L}' = \ln \mathcal{L} + \sum_i v_i,$$

where v_i is the value of penalty for $\boldsymbol{\mathcal{P}}_i$.

We do not recommend $\mathcal{L}'$ to be used as an optimizing parameter, because this favors those allele configurations with a larger sum of the numbers of amplifiable alleles across putative loci. Such an allele configuration needs more parameters (i.e., allele frequencies) in order to describe the data. For example, if each amplifiable allele is assigned to all putative loci, then the number of allele frequencies is Lk'. In this situation, the likelihood is maximized. Alternatively, the allele configuration **A** is evaluated by the BIC. This parameter is calculated from the likelihood of **A**, but will be penalized according to the number of parameters. The appropriate formula is:

$$\text{BIC} = -2\ln \mathcal{L}' + d\left(\ln n - \ln 2\pi\right),$$

where n is the sample size, and d is the sum of the numbers of various parameters related to **A**, namely

$$d = \sum_l \left(k_l' - 1 + b_\beta + b_\gamma + b_{fh}\right).$$

Here, k_l' is the number of amplifiable alleles at the l^{th} locus; b_β, b_γ and b_{fh} are binary variables (i.e. the variables with 0 and 1 as their values), which are used to indicate the factors considered (b_β for negative amplification, b_γ for null alleles, and b_{fh} for heterozygous deficiency or excess). If a factor is considered, the corresponding binary variable is set to one; otherwise, it is set to zero.

If the number of indeterminate alleles is η, there are 2^{η} multilocus genotypes, where 2^{η} will increase exponentially as η increases. Here, a recursive algorithm will be used to generate all multilocus genotypes, and a pruning algorithm will be applied to exclude some impossible multilocus genotypes. To further simplify this problem, these L loci can be divided into several locus groups, such that any two loci between different groups will not share identical alleles. This is so the evaluation of an allele configuration will be divided into several less complex problems, with these L loci being evaluated group-by-group. In addition, the dynamic programming and the hash table are used to prevent repeat calculations.

Genotypic frequencies are used to estimate allele frequencies and to calculate the likelihood of an allele configuration. The derivation of genotypic frequencies under the factors considered is shown in Appendix A, and the algorithm of allele frequency estimation can be found in Appendix B.

2.5 | Evaluation

Here, we use three applications to evaluate the reliability of our model. Application 1 evaluates the correct assignment rate as a function of sample size. Application 2 evaluates the correct assignment rate as a function of one of the following five parameters: null allele frequency, negative amplification rate, heterozygote excess index, intensity of the sharing of identical alleles and the correlation coefficient of linkage disequilibrium. These two applications use Monte-Carlo simulations. Application 3 uses empirical data to validate our new algorithm.

For the Monte-Carlo simulations, we first simulate a population. We then randomly sample a number of individuals, whose genotypes are generated and then converted to phenotypes. Finally, allele assignment is performed to test whether this can reliably assign alleles to the correct loci.

Eight parameters are considered in each simulation: these are the sample size n, number L of loci, number k'_l of amplifiable alleles per locus, compression rate s (evaluating the intensity of sharing of identical alleles among loci), null allele frequency p_y, negative amplification rate β, heterozygote excess index h and the correlation coefficient r of linkage disequilibrium. Because genes within the MHC gene family are often under balancing selection (Aguilar et al., 2004), heterozygote deficiency is not considered. To simulate linkage disequilibrium, we assume that there are the same number of amplifiable alleles per locus, and that one allele at a locus is exclusively linked to another allele at another locus. The number of amplifiable alleles is compressed from Lk'_l to k' due to the sharing of identical alleles. Therefore, the compression rate s is calculated by $s=1-k'/Lk'_l$. Conversely, if the compression rate s is given, the total number k' of amplifiable alleles can be calculated by $k'=(1-s)\,Lk'_l$. Unfortunately, when s is given a particular value (e.g. = 0.1), the value of (1 −) Lk'_l is usually not an integer. In order to ensure the expected value of s is equal to our expectation, we first let $k'_{\text{temp}}=(1-S)Lk'_l$, $S_{\text{floor}}=1-\frac{\left[k'_{\text{temp}}\right]}{Lk'_l}$ and $S_{\text{ceiling}}=1-\frac{\left[k'_{\text{temp}}\right]}{Lk'_l}$, and we then set $k'=\left[k'_{\text{temp}}\right]$ at a probability of $\frac{S-S_{\text{floor}}}{S_{\text{ceiling}}-S_{\text{floor}}}$, and $k'=\left[k'_{\text{temp}}\right]$ at a probability of $\frac{S_{\text{ceiling}}-S}{S_{\text{ceiling}}-S_{\text{floor}}}$.

Because our algorithm can simultaneously account for multiple effects, it is impractical to present the results of all possible combinations in this paper. We combine five effects and define three basic benchmarks: IDEAL, GOOD and POOR. Our IDEAL benchmark assumes that there are no null alleles, no negative amplification, no heterozygote excess, no linkage disequilibrium and no sharing of identical alleles, i.e. $p_y = \beta = h = r = s = 0$. Our GOOD benchmark assumes that the values of these five parameters are all equal and low, and set as $p_y = \beta = h = r = s = 0.03$. Most values for our POOR benchmark are assumed to be about three times those of our GOOD benchmark, and are set as $p_y = \beta = h = r = s = 0.1$. Because MHC genes are usually under balancing selection, we consider heterozygote excess as the default in these benchmarks (i.e., $f = 0$) and the effect of heterozygote deficiency (inbreeding) is simulated in Application 2.

2.6 | Generation of phenotypes

The allele frequencies at each locus are randomly generated according to broken-stick numbers. For example, for a locus with k' amplifiable alleles, we generate $k'-1$ uniformly distributed random numbers: $x_i \sim U(0,1-p_y)$, and set $x_{k'}=0$, $x_{k'+1}=1-p_y$ and $x_{k'+2}=1$. We then sort these numbers into a sequence $X_1, X_2, \cdots, X_{k'+2}$ in ascending order, and then let the frequency of the ith allele be equal to $X_{i+1}-X_i$, with the last frequency being that of the null allele. The k'_l amplifiable alleles at each locus are randomly mapped to k' amplifiable alleles to simulate the sharing of identical alleles among loci (Figure 3).

We then sample n individuals and generate their genotypes with the effects of linkage disequilibrium and heterozygote excess. These L loci are assumed to be located on the same chromosome, with r as the correlation coefficient between two neighbouring loci. Each locus is assumed to have the same number of alleles, with each allele being linked to a corresponding allele at each neighbouring locus (Figure 3). The genotype G_1 at the first locus is randomly generated by Equation (A2). The genotype G_l at the l^{th} locus ($l > 1$) is generated conditional on G_{l-1} at the previous locus. Both alleles in G_l are linked with the two alleles in G_{l-1} at a probability of r. If both alleles are unlinked, then G_l is randomly generated according to Equation (A2). If one allele (say A_{l1}) is linked while the other (say A_{l2}) is not, then A_{l2} is a randomly sampled allele distinct from A_{l1} at a probability of h, and is a randomly sampled allele at a probability of $1 - h$.

The genotype of each individual at each locus is converted to an observed genotype by removing any null alleles, and then by randomly setting to $\varnothing$ at a probability of β to simulate the effects of both null alleles and negative amplification. Finally, by taking the set union of the observed genotypes at L loci, the phenotype of an individual is obtained.

2.7 | Application 1

For this application, we evaluate the correct assignment rate (denoted by c) as a function of sample size n. The correct assignment is the probability that all sampled alleles are assigned to the correct loci.

Each of our three benchmark categories combines five parameters, with the remaining three parameters being n, L and k'_l. Each of both L and k'_l is set with three levels: $L \in \{2,3,4\}$ and $k'_l \in \{5,10,15\}$, which each represent different complexities. The sample size n is simulated from 50 to 1,000 at an interval of 50.

A total of 27 graphs of the function c under different conditions (IDEAL, GOOD, POOR, $L \in \{2,3,4\}$ and $k'_l \in \{5,10,15\}$) are presented in Figure 4. Each graph is drawn by 20 points obtained using simulation. For each of these points, 2,000 simulations are performed and the probability of obtaining a correct result is calculated. The initial values of p_y, β and h of the likelihood estimator are each set to 0.05, the threshold d to terminate the EM algorithm is set to 0.0001. A two-step method is used to accelerate the calculation (see Searching algorithm in Discussion): 10 cycles of primary annealing without considering any factor (NONE) are initially performed, with the initial temperature and the final temperature of the simulated annealing algorithm being 0.01 and 0.00001, respectively. The penalties of missing and mismatch errors are $p_{\text{missing}} = -2$ and $p_{\text{mismatch}} = -80$, respectively. This is followed by 100 cycles of secondary annealing with the three factors $(h|f + \beta + p_y)$ considered and with an initial temperature of 0.001, where the symbol "|" denotes "or". The annealing coefficient is set at 0.99, and at each temperature 500 new allele configurations are generated from the current allele configuration. The penalties of missing and mismatch errors are $p_{\text{missing}} = 0$ and $p_{\text{mismatch}} = -1{,}000$, respectively.

2.8 | Application 2

We also evaluate the correct assignment rate c as a function of one of the six parameters p_y, β, h, f, r and s combined for each benchmark (for these parameters, one is regarded as the variable whilst the others are set as constants). The null allele frequency p_y, negative amplification rate β, heterozygote excess index h, inbreeding

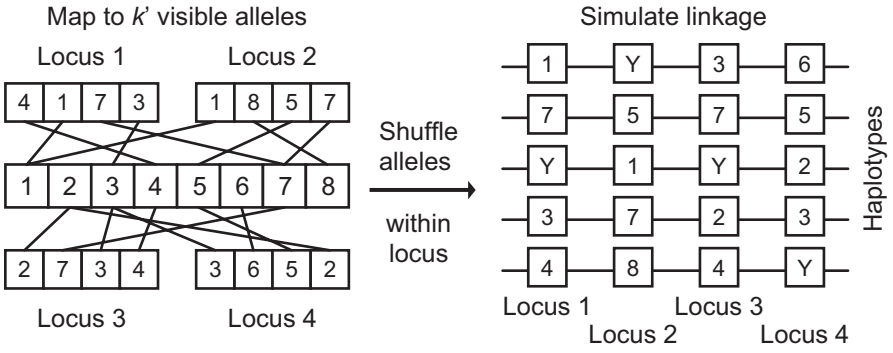

FIGURE 3 A schematic diagram of simulating sharing identical alleles among loci and linkage disequilibrium. Where Y is the null allele or deletion mutation (i.e., copy number variation)

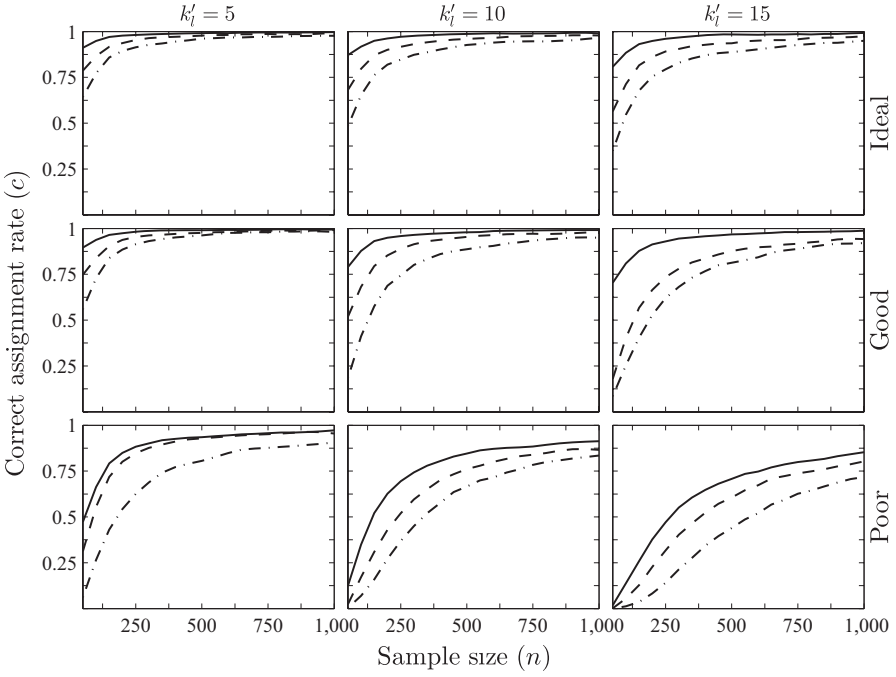

FIGURE 4 The graphs of the correct assignment rate c as a function of sample size n. The solid, dashed or dash-dotted lines represent the graphs when the number L of loci is 2, 3 or 4. Three levels (k'_l = 5, 10 and 15) of amplifiable alleles per locus were also included in separate simulations, and are shown in the three corresponding columns. Three benchmarks for genotyping and population genetic parameters were also included in separate simulations. The IDEAL benchmark with $p_y = \beta = h = r = s = 0$ is shown in the top row, the GOOD benchmark with $p_y = \beta = h = r = s = 0.03$ is shown in the middle row, and the POOR benchmark with $p_y = \beta = h = r = s = 0.1$ is shown in the bottom row. For each scenario, we performed 2,000 Monte-Carlo simulations

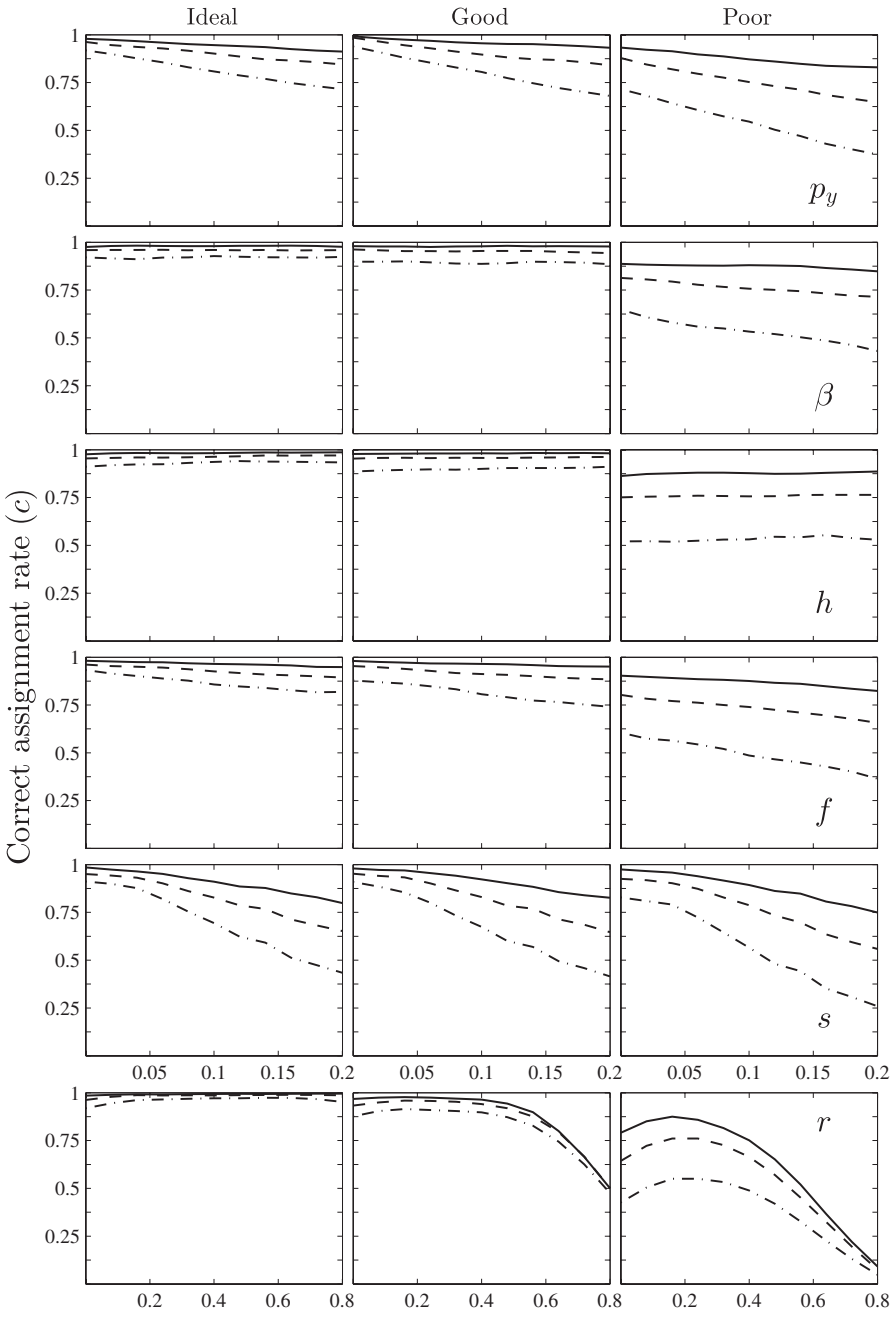

FIGURE 5 The graphs showing correct assignment rate c as a function of one of the following five parameters: null allele frequency p_y, negative amplification rate β, heterozygote excess index h, compression rate s, and the correlation coefficient r of linkage disequilibrium. Each row shows one independent variable. The numbers of loci and amplifiable alleles per locus were three and 10, respectively. The solid, dashed or dash-dotted lines represent the graphs when the total sample size n is 1,000, 500 or 250. Three benchmarks, IDEAL, GOOD and POOR, the same as shown in Figure 4, were each used for simulation and are shown in the left, middle and right columns. For each scenario, we performed 2,000 Monte-Carlo simulations

coefficient f and the compression rate s are all simulated from 0 to 0.2 at an interval of 0.02, whilst the correlation coefficient r for linkage disequilibrium is simulated at a range of values from 0 to 0.8 at an interval of 0.08. The remaining three parameters n, L and k'_l are set at some medium complexities: L = 3, k'_k = 10 and $n \in \{250,500,1,000\}$.

A total of 45 graphs of c as a function of p_y, β, h, f, s or r under different conditions (IDEAL, GOOD, POOR and $n \in \{250,500,1,000\}$) are presented in Figure 5. Each graph is drawn using 11 points obtained by simulation. For each point, we also perform 2,000 simulations. Various values of the parameters for the likelihood estimator and for the searching algorithm are the same as those in Application 1.

2.9 | Application 3

Here, we use three empirical data sets from the North Atlantic Right Whale (*Eubalaena glacialis*) (Gillett et al., 2013), the Wolf (*Canis lupus*) (Galaverni et al., 2013) and the Barn Owl (*Tyto alba*) (Gaigher et al., 2018) to validate our model. All three data sets used haplotype reconstruction to phase the haplotypes (Stephens & Donnelly, 2003).

The Right Whale data set of Gillett et al. (2013) contains data from two loci and 10 alleles (n = 30), and represents a low level of complexity (5.90×10^4 possible allele configurations). The Wolf data set of Galaverni et al. (2013) consists of data from three loci and 23 alleles (n = 74) and thus represents a medium level of complexity (2.74 × 1,019 possible allele configurations).

TABLE 1 The results of allele assignment of empirical data

Data source	*n*	*L*	*k′*	Allele configuration					
				Missing	Mismatch	BIC	ln $\mathcal{L}$	k^*	Diagnose
Gillett et al. (2013)	30	2	10	0	0	259.709	−93.181	0	Correct
Galaverni et al. (2013)	74	3	22	0	0	1,244.858	−536.442	4	Incorrect
				0	0	1,245.079	−536.552	0	True
Gaigher et al. (2018) Class I	937	2	69	67	0	18,907.711	−9,137.015	*N/A*	Consistent[a]
Gaigher et al. (2018) Class II	937	2	42	352	0	13,014.29	−6,307.492	0	Correct

Note: Where we denote *n* for the total sample size, *L* for the number of loci within data sets, *k′* for the number of amplifiable alleles, "Missing" and "Mismatch" for the numbers of missing error and mismatch error for an allele configuration, respectively, and k^* for the number of alleles assigned to the incorrect loci. If an allele configuration is incorrect, the true allele configuration will be also provided as a comparison.

[a]Allele configuration obtained from the phenotypes: (1,3,8,11–12,16–17,19–21,23–25,27,29,36–38,41,43,47,50,56–57,60,62,65,69–72,78,96,106,116,123)(2,4–7,9–10,13–15,18,22,26,28,30–32,35,39,44,46,48,52–54,58–59,67,83–85,93). This is consistent with the incomplete allele configuration inferred from the reconstructed haplotypes: (1,3,8,11–12,16–17,19–21,23–25,27,29,34,36–38,41,43,47,50,56–57,60,62,65,69–72,78,106,116,123)(2,4–7,9,10,13–16,18,22,26,28,30–31,35,39,44,46,48,52–54,58–59,83,85,93) 32? 67? 84? 96?, in which an allele followed by a question mark denotes the allele cannot be assigned by haplotype reconstruction.

The Barn Owl data set of Gaigher et al. (2018) contains data from the amplification of two classes of MHC genes (*n* = 74). The first, two MHC class I loci, consists of 69 amplifiable alleles and represents a high level of complexity (8.34 × 1,032 possible allele configurations); the second, two MHC class II loci (*DAB1* and *DAB2*), consists of 42 amplifiable alleles and represents a medium level of complexity (1.09 × 1,020 possible allele configurations). The MHC class II loci data were obtained by single-locus amplifications, and we converted the genotypes to the phenotypes using the same methods as described for Application 1. Some MHC class I alleles appeared to be under complete linkage, so their true locus origin could not be inferred from the haplotype data.

Various values of the parameters for the likelihood estimator and the searching algorithm are as for Application 1.

3 | RESULTS

3.1 | Application 1

The graphs for correct assignment rate *c* as a function of the sample size *n* under different benchmarks and different values of both *k′* and *L* are presented in Figure 4. It can be seen that for each curve in this figure, *c* increases as *n* increases, and the curve reaches close to an asymptote when *n* is large enough.

The curves of *c* when *k′* = 5 are similar to those when *k′* = 10, but a higher value of *k′* requires more samples to reach a similar asymptote. Similarly, for a higher value of *L* or in worse conditions, our method performs worse and the convergence speed is slower than that for a lower value of *L* or in better conditions.

The performance of our model is slightly reduced by using the GOOD benchmark compared with the IDEAL benchmark, but it is adversely affected by the use of the POOR benchmark. Under the POOR benchmark, even the use of one thousand samples fails to enable the model to reach a correct assignment rate of 95%.

3.2 | Application 2

The graphs of the correct assignment rate *c* as a function of $p_{y'}$ β, *h*, *f*, *s* or *r* are shown in Figure 5. It is clear that the performance of our model is similar for all parameters under both IDEAL and GOOD conditions, except for *s*. The correct assignment rate *c* decreases linearly with an increase of both $p_{y'}$ *f* and *s*, the slope for *s* is nearly twice as steep as that for $p_{y'}$.

Under both IDEAL and GOOD benchmarks, our method is largely unaffected by both β and *h*. Under the POOR benchmark, *c* slightly increases as *h* increases but decreases as β increases. In addition, *c* increases as *r* increases when *r* is small, then peaks when *r* = 0.4 and returns to the initial value when *r* = 0.4.

3.3 | Application 3

The results of the allele assignments of the empirical data are presented in Table 1, which show the number of missing errors, the number of mismatch errors, the BIC, the ln $\mathcal{L}'$, and the number of alleles assigned to incorrect loci. If an allele configuration is incorrect, the true allele configuration is provided for comparison.

Due to a small sample size and a medium level of complexity, the results of allele assignment presented in Galaverni et al. (2013) are likely to be incorrect. Comparison with the true allele configuration shows that both the BIC and the ln $\mathcal{L}'$ are suboptimal to those of our allele configuration. For the other two data sets, the results for allele assignment are correct or consistent with the results obtained by haplotype reconstruction.

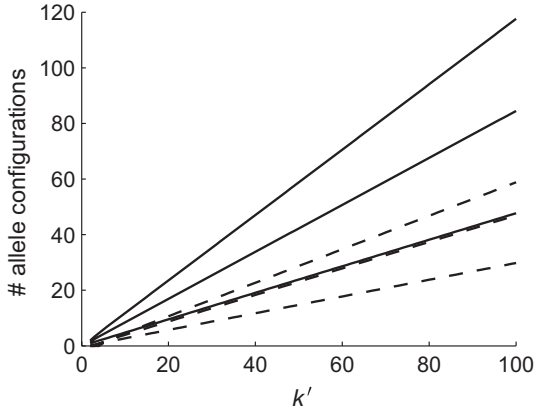

FIGURE 6 The common logarithm of the number of possible allele configurations to assign k' alleles into L loci. The results under the assumption that different loci share identical alleles are denoted by the solid lines, while the dashed lines denote no identical alleles to be shared among loci. From bottom to top, there are three lines of each line style, each of which denotes the graph of common logarithm of the number of possible allele configurations as a function of k', when L = 2, 3 or 4

4 | DISCUSSION

4.1 | Factors

We here develop a method to assign alleles in multilocus amplification to different loci. We extract possible multilocus genotypes from each phenotype under an allele configuration, and then use the posterior probability of each multilocus genotype as the weight to estimate the allele frequencies and four additional parameters (null allele frequency p_y, negative amplification rate β, heterozygote excess index h and inbreeding coefficient f). Based on the estimated allele frequencies, the genotypic frequencies and the likelihood for each allele configuration are calculated. BIC is used to evaluate the allele configuration and a simulated annealing algorithm is used to find the optimal allele configuration.

Our model allows different loci to share identical alleles, and the complexity (the number of possible allele configurations) is relatively high. If this is not considered and loci do not share identical alleles, then the complexity of the model is greatly reduced. The correct assignment rate will then be improved and the requirements can be relaxed (e.g., fewer samples are required, Figures S1 and S2).

We considered several factors in our model, including the presence of null alleles, negative amplification, and two factors that may allow the genotypic frequencies to deviate from the HWE: heterozygote excess and inbreeding. We also considered the CNV in our model and used null alleles as an equivalence. Due to the inclusion of such complexity, we did not include linkage disequilibrium into the model. However, the robustness of our method enabled us to account for this and to evaluate the effects of linkage disequilibrium by using the same method as for the other factors.

If any linkage is complete, then the problem of assigning an allele is unsolvable. For example, if two alleles (A and B) are only present in a same haplotype, then they are completely linked to each other and will always be concurrent within the same phenotype. Therefore, these two alleles are exchangeable, and the likelihood of assigning A to one locus (e.g., l_1) and B to another (e.g., l_2) is equal to that of assigning B to l_1 and A to l_2. Our algorithm cannot distinguish which is best and can only chose either randomly. Unfortunately, in this case, even if pedigree information is known and a haplotype reconstruction technique is used, the correct allele configuration will still be unavailable (e.g. Gaigher et al., 2018).

4.2 | Solution space

There are many possible allele configurations to assign k' amplifiable alleles into L loci, and the number $N_A(k', L)$ of possible allele configurations can be calculated with a recursive function as:

$$N_A(k',L) = \left(2^L - 1\right)^{k'} - \sum_{l=1}^{L-1} \binom{L}{l} N_A(k',L) \approx \left(2^L - 1\right)^{k'},$$

where $N_A(k', 1) = 1$. If it is assumed that there are no identical alleles shared among loci, then the number $N_A^*(k',L)$ of possible allele configurations can be obtained by the Stirling numbers of the second kind:

$$N_A^*(k',L) = \frac{1}{k'!} \sum_{i=0}^{k'} (-1)^{k'-i} i^L \binom{k'}{i}.$$

The common logarithm $\log_{10} N_A(k', L)$ or $\log_{10} N_A^*(k',L)$ can be regarded as a function of k' where L is set at 2, 3 or 4 (Figure 6). From this figure, $\log_{10} N_A(k', L)$ and $\log_{10} N_A^*(k',L)$ increase almost linearly as k' increases, and $\log_{10} N_A^*(k',L) \approx \frac{1}{2} \log_{10} N_A(k',L)$ for each L, then $N_A(k', L)$ and $N_A^*(k',L)$ increase almost exponentially as k' increases, and $N_A^*(k',L) \approx \sqrt{N_A(k',L)}$ for each L. When the value of k' is high, the number of possible allele configurations is also high, and there are too many allele configurations to enumerate each one sequentially. Therefore, heuristic algorithms are required to solve this problem.

4.3 | Limitations

The enormous number of possible allele configurations will cause an additional problem: if the sample size is limited, an invalid allele configuration may have a better BIC than the true solution (i.e., allele configuration). Even if we take advantage of the first kind of genetic information (i.e., the co-occurrence of alleles) and exclude some invalid allele configurations that contain either or both missing and mismatched errors, the number of remaining allele configurations may still be high.

In the data set of Galaverni et al. (2013), the optimal allele configuration is not the correct allele configuration (Table 1). Although our model provides a method to evaluate the likelihood among different allele configurations, when the complexity is high (Figure 6) and the sample size is insufficient, some incorrect allele

configurations may be "better" than the correct allele configurations (have a lower BIC).

Although our searching algorithm is able to find the optimal allele configuration, this may not be the correct allele configuration. In such a case, thousands of samples are required to ensure that the optimal allele configuration is correct. Under natural conditions, due to the limitation of many factors, such a condition cannot be attained. Therefore, our method can only be applied to scenarios of low to medium complexity (e.g., <60 alleles and five loci, Table S1), methods identical to those for Application 1.

Constraints due to sample size can be solved by collecting samples from multiple populations and performing the assignment with all samples. A potential problem of this approach is that the variation in allele frequencies among populations (quantified by F_{ST}) may affect the allele assignment. In this case, the genotypic frequency under heterozygote deficiency can still be calculated using Equation (A4), but the inbreeding coefficient f in Equation (A4) should be revised from F_{IS} to F_{IT} and the value of f is increased by F_{ST} $(1 - F_{IS})$. A low level of differentiation (F_{ST} < 0.1) will not greatly reduce the accuracy, and any loss can be offset by increased sample size (e.g., from 100 to 250) which will increase the value of f (e.g., from 0 to 0.1) (Figure 5). In addition, the heterozygote excess h can neutralize the heterozygote deficiency caused by inbreeding or differentiation.

Based on a survey of previous studies on MHC genes, we found in most cases there are 2–4 loci that are amplified simultaneously (26 out of 38 data sets, Table S2), with hundreds of samples often being available (32 out of 45 data sets, Table S2). Under real conditions, on one hand researchers are also attempting to design locus-specific primers, but sometimes loci are too similar to distinguish. On the other hand, some applications enable the use of an alternative strategy, and attempt to use a pair of universal primers to amplify as many alleles as possible. Therefore, researchers are, to some extent, able to control the number of loci that are amplified through the design of locus-specific primers. Under this scenario our new method enables researchers to attain a degree of accuracy that was previously impossible to achieve.

Our likelihood estimator is asymptotically unbiased under the factors we consider, and the correct assignment rate can reach one when there are an infinite number of samples. However, because we do not model any linkage disequilibrium, our likelihood estimator is biased under linkage disequilibrium. In such a case, even if there are an infinite number of samples, our method cannot ensure a 100% correct assignment rate.

4.4 | Likelihood estimator

We extended the maximum likelihood estimator of Kalinowski and Taper (2006) the likelihood of each allele configuration. This modified estimator is able to simultaneously estimate the following parameters: allele frequencies, negative amplification rate, inbreeding coefficient and heterozygote excess index. Additionally, these three factors (h|$f + \beta + p_y$) can be freely combined, with other factors being excluded as and when required. If any factors are to be excluded, then the corresponding parameters will be set as zero at the beginning and at the end of each iteration. For example, if inbreeding and heterozygote excess are both not to be considered, then this estimator is reduced to the original estimator of Kalinowski and Taper (2006).

Because the likelihood as a function of various factors being simultaneously considered is not unimodal, multiple solutions exist that are approximately equal to the maximized likelihood. For example, for a triallelic locus, if the allele frequencies at this locus are uniformly distributed ($p_A = p_B = p_y = 0.3333$) and $\beta = h = 0.1$, the estimated and true likelihoods are all −136.5659, approximately equal to the maximized likelihood. By adjusting the initial solution of allele frequency estimation, the true solution can also be obtained.

In the process of simulation and estimation, we generated 100 genotypes with 30 {*A*}, 30 {*B*}, 21 {*A*,*B*} and 19 ∅. For our likelihood estimator, the estimates of allele frequencies and both parameters β and h are $\hat{p}_A = \hat{p}_B = 0.3404$, $\hat{p}_y = 0.3192$, $\hat{\beta} = 0.0991$ and $\hat{h} = 0.0103$. According to Equations (A3) and (A5), the true and estimated (including observed and actual) genotypic frequencies are shown in Table 2. Here, it is shown that the estimated genotypic frequencies are equal to the true genotypic frequencies. Therefore, our estimator can only be used to estimate the likelihood of an allele configuration. The allele frequencies and the other four parameters are only intermediate variables used to calculate likelihood and BIC, and cannot be used for other applications.

4.5 | Searching algorithm

The number of loci should be assigned before performing the allele assignment. If loci are assumed to share identical alleles, then the number L of loci should satisfy the criterion that $L \geq \lceil k'_{max}/2 \rceil$, where k'_{max} is the maximum of the numbers of amplifiable alleles in each individual. The user can try different values of L and use BIC to identify the best allele configuration by comparison.

In practice, the penalties for missing or mismatch errors can be adjusted according to the performance of the searching algorithm. For example, if the current best allele configuration is stopped at some allele configurations with many missing errors but only few mismatched errors (e.g., $n_{missing} = 300$ and $n_{mismatch} = 6$), then it can be further optimized by aggravating the penalty for missing errors (e.g., $p_{missing} = -8$ or -10).

We tested the efficiency of the searching algorithm under different factor combinations and found that efficiency is high for the three combinations: NONE, β and p_y. The algorithm for these combinations can obtain the optimal allele configuration within the first two annealing cycles in the empirical examples. Whilst its efficiency is low for the remaining combinations ($h|f$, $h|f + \beta$, $h|f + p_y$, $\beta + p_y$ and $h|f + \beta + p_y$), in such a case, even 1,000 annealing cycles cannot ensure that the optimal allele configuration is found. Furthermore, if the factor combinations containing at least one factor are considered, then the allele frequency estimation using the EM algorithm usually requires many iterations (e.g., 1,000 iterations), which reduces the efficiency of our estimation. In contrast, the allele

TABLE 2 The true and estimated genotypic frequencies at a tri-allelic locus

Type	G	Frequency	Likelihood	G_A	Frequency
True solution	{A}	0.3000	−1.2040	{A}	0.0900
				{A,Y}	0.2100
	{B}	0.3000	−1.2040	{B}	0.0900
				{B,Y}	0.2100
	∅	0.1900	−1.6607	{Y}	0.0900
				∅	0.1000
	{A,B}	0.2100	−1.5606	{A,B}	0.2100
Estimated solution	{A}	0.3001	−1.2037	{A}	0.1033
				{A,Y}	0.1968
	{B}	0.3001	−1.2037	{B}	0.1033
				{B,Y}	0.1968
	∅	0.1900	−1.6607	{Y}	0.0909
				∅	0.0991
	{A,B}	0.2099	−1.5613	{A,B}	0.2099

Note: Where the sample size is 100; the true parameters are $p_A = p_B = p_y = 0.3333$, $\beta = h = 0.1$, and the likelihood is − 136.5659; the estimated parameters are $\hat{p}_A = \hat{p}_B = 0.3404$, $\hat{p}_y = 0.3192$, $\hat{\beta} = 0.0991$, $\hat{h} = 0.0103$, and the likelihood is also − 136.5659.

frequency estimation without considering any factor (NONE) requires only two iterations.

We conclude that using a two-step method to accelerate the calculation is optimal. The first step uses the factor combination without considering any factor (NONE) to perform a primary annealing with 10 cycles and to obtain a preoptimal allele configuration. The second step then uses the target factor combination (e.g., $h|f + \beta + p_y$) and a reduced initial temperature (e.g., 0.001) to optimize the allele configuration obtained from the previous step.

5 | CONCLUSION

In eukaryotes, a large proportion of genes (such as 38% in *Homo sapiens* and 65% in *Arabidopsis thaliana*) have evolved via gene duplication (Zhang, 2003). Such genes are involved in many important ecological and evolutionary processes. For example, olfactory receptor (OR) genes provide a molecular basis for understanding how some animals adapt to their chemical environment (Krieger & Breer, 1999), and RNase genes are used to study the evolution of digestive function of mammals (e.g., Yu & Zhang, 2006; Zhang, 2006).

Although the MHC gene family was used as a primary example to demonstrate our model, our model can also be applied to other duplicated genes. This expands both the scope and scale of future studies of systems involving duplicated genes. Our new method will improve studies of other duplicated genes in three distinct ways. First, our model enables more accurate parameters to measure the genetic diversity of loci/populations (e.g., Hardy & Vekemans, 2002) or the relationships between individuals (e.g., Lynch & Ritland, 1999; Wang, 2002; Wang, 2002). Second, more powerful statistical tests and classical population genetics methods can be performed to resolve specific scientific problems (e.g., Guo et al., 2015; Rousett, 2008). Finally, our new method enables the study of more target species, because the necessity of pedigree information and the quality of tissue samples are both reduced.

ACKNOWLEDGEMENTS

We thank the five anonymous reviewers for their suggestions and comments. This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB310200000), the National Natural Science Foundation of China (31730104, 31770411, 31572278 and 31770425), the Young Elite Scientists Sponsorship Program by CAST (2017QNRC001), the National Key Programme of Research and Development, the Ministry of Science and Technology of China (2016YFC0503202), and the Natural Science Basic Research Plan in Shaanxi Province of China (2018JM3024, 2019JM258). DWD is supported by a Shaanxi Province Talents 100 Fellowship and KH is supported by a scholarship from China Scholarship Council.

AUTHOR CONTRIBUTIONS

K.H. and B.G.L. designed the project, K.H. and P.Z. established the model and wrote the draft, R.M., T.C.W. and Y.D. checked the model and performed the simulation, D.W.D. helped write the manuscript.

DATA AVAILABILITY STATEMENT

The software MHC-TYPER V1.1, user manual and example dataset of Table 1 are available on GitHub (https://github.com/huangkang1987/mhc-typer). Data deposited at Dryad: https://doi.org/10.5061/dryad.4d789, https://doi.org/10.5061/dryad.15r7f, https://doi.org/10.5061/dryad.745t0.

REFERENCES

Aguilar, A., Roemer, G., Debenham, S., Binns, M., Garcelon, D., & Wayne, R. K. (2004). High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 3490–3494. https://doi.org/10.1073/pnas.0306582101

Babik, W. (2009). Methods for MHC genotyping in non-model vertebrates. *Molecular Ecology Resources*, *10*, 237–251. https://doi.org/10.1111/j.1755-0998.2009.02788.x

Bauer, P., Lubkowitz, M., Tyers, R., Nemoto, K., Meeley, R. B., Goff, S. A., & Freeling, M. (2004). Regulation and a conserved intron sequence of *liguleless3/4 knox* class-I homeobox genes in grasses. *Planta*, *219*, 359–368. https://doi.org/10.1007/s00425-004-1233-6

Bertsimas, D., & Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science*, *8*, 10–15. https://doi.org/10.1214/ss/1177011077

Biedrzycka, A., Sebastian, A., Migalska, M., Westerdahl, H., & Radwan, J. (2016). Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. *Molecular Ecology Resources*, *17*, 642–655. https://doi.org/10.1111/1755-0998.12612

Brookfield, J. F. Y. (1996). A simple new method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology*, *5*, 453–455. https://doi.org/10.1111/j.1365-294X.1996.tb00336.x

Detwiler, J. T., & Criscione, C. D. (2011). Testing Mendelian inheritance from field-collected parasites: Revealing duplicated loci enables correct inference of reproductive mode and mating system. *International Journal for Parasitology*, *41*, 1185–1195. https://doi.org/10.1016/j.ijpara.2011.07.003

Eizaguirre, C., Lenz, T. L., Kalbe, M., & Milinski, M. (2012). Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nature Communications*, *3*, 621. https://doi.org/10.1038/ncomms1632

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, *131*, 479–491.

Gaigher, A., Roulin, A., Gharib, W. H., Taberlet, P., Burri, R., & Fumagalli, L. (2018). Lack of evidence for selection favouring MHC haplotypes that combine high functional diversity. *Heredity*, *120*(5), 396–406. https://doi.org/10.1038/s41437-017-0047-9

Galaverni, M., Caniglia, R., Fabbri, E., Lapalombella, S., & Randi, E. (2013). MHC variability in an isolated wolf population in Italy. *Journal of Heredity*, *104*, 601–612. https://doi.org/10.1093/jhered/est045

Gillett, R. M., Murray, B. W., & White, B. N. (2013). Characterization of Class I- and Class II-like major histocompatibility complex loci in pedigrees of North Atlantic right whales. *Journal of Heredity*, *105*, 188–202. https://doi.org/10.1093/jhered/est095

Guo, S. T., Huang, K., Ji, W. H., Garber, P. A., & Li, B. G. (2015). The role of kinship in the formation of a primate multilevel society. *American Journal of Physical Anthropology*, *156*, 606–613.

Hardy, O. J., & Vekemans, X. (2002). SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, *2*, 618–620. https://doi.org/10.1046/j.1471-8286.2002.00305.x

Hedrick, P. W., & Parker, K. M. (2017). MHC variation in the endangered Gila topminnow. *Evolution*, *52*, 194–199. https://doi.org/10.1111/j.1558-5646.1998.tb05152.x

Huang, K., Ritland, K., Dunn, D. W., Qi, X., Guo, S., & Li, B. (2016). Estimating relatedness in the presence of null alleles. *Genetics*, *202*, 247–260. https://doi.org/10.1534/genetics.114.163956

Huchard, E., Knapp, L. A., Wang, J. L., Raymond, L., & Cowlishaw, G. (2010). MHC, mate choice and heterozygote advantage in a wild social primate. *Molecular Ecology*, *19*, 2545–2561. https://doi.org/10.1111/j.1365-294X.2010.04644.x

Kalinowski, S. T., & Taper, M. L. (2006). Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics*, *7*, 991–995. https://doi.org/10.1007/s10592-006-9134-9

Klein, J. (1986). *Natural history of the major histocompatibility complex*. New York:John Wiley & Sons.

Krieger, J., & Breer, H. (1999). Olfactory reception in invertebrates. *Science*, *286*, 720–723. https://doi.org/10.1126/science.286.5440.720

Lynch, M., & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, *152*, 1753–1766.

Maruyama, T., & Takahata, N. (1981). Numerical studies of the frequency trajectories in the process of fixation of null genes at duplicated loci. *Heredity*, *46*, 49–57. https://doi.org/10.1038/hdy.1981.5

McCarroll, S. A., & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature Genetics*, *39*, S37–S42. https://doi.org/10.1038/ng2080

Miller, H. C., Moore, J. A., Nelson, N. J., & Daugherty, C. H. (2009). Inuence of major histocompatibility complex genotype on mating success in a free-ranging reptile population. *Proceedings of the Royal Society B: Biological Sciences*, *276*, 1695–1704.

Murdoch, S., Seoud, M., Kircheisen, R., Mazhar, B., & Slim, R. (2006). Detailed gene and allele content analysis of three homozygous KIR haplotypes. *Tissue Antigens*, *68*, 72–77. https://doi.org/10.1111/j.1399-0039.2006.00606.x

Nei, M. (1972). Genetic distance between populations. *American Naturalist*, *106*, 283–292. https://doi.org/10.1086/282771

Piertney, S. B., & Oliver, M. K. (2005). The evolutionary ecology of the major histocompatibility complex. *Heredity*, *96*, 7–21. https://doi.org/10.1038/sj.hdy.6800724

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Quattro, J. M., Jones, W. J., & Oswald, K. J. (2001). PCR primers for an aldolase-B intron in acanthopterygian fishes. *BMC Evolutionary Biology*, *1*, 9.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, *67*, 175–185. https://doi.org/10.1017/S0016672300033620

Rousett, F. (2008). GENEPOP'007: A complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*, 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Santos, P. S. C., Michler, F. U., & Sommer, S. (2017). Can MHC-assortative partner choice promote offspring diversity? A new combination of MHC-dependent behaviours among sexes in a highly successful invasive mammal. *Molecular Ecology*, *26*, 2392–2404. https://doi.org/10.1111/mec.14035

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Selvaraj, S., Schmitt, A. D., Dixon, J. R., & Ren, B. (2015). Complete haplotype phasing of the MHC and KIR loci with targeted HaploSeq. *BMC Genomics*, *16*, 900. https://doi.org/10.1186/s12864-015-1949-7

Shilling, H. G., Guethlein, L. A., Cheng, N. W., Gardiner, C. M., Rodriguez, R., Tyan, D., & Parham, P. (2002). Allelic polymorphism synergizes with variable gene content to individualize human *KIR* genotype. *The Journal of Immunology*, *168*, 2307–2315.

Slatkin, M. (2008). Linkage disequilibrium|understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*, 477–485. https://doi.org/10.1038/nrg2361

Stephens, M., & Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, *73*, 1162–1169. https://doi.org/10.1086/379378

Strandh, M., Westerdahl, H., Pontarp, M., Canback, B., Dubois, M.-P., Miquel, C., … Bonadonna, F. (2012). Major histocompatibility complex class II compatibility, but not class I, predicts mate choice in a bird with highly developed olfaction. *Proceedings of the Royal Society*

B: Biological Sciences, *279*, 4457–4463. https://doi.org/10.1098/rspb.2012.1562

Stuglik, M. T., Radwan, J., & Babik, W. (2011). jMHC: Software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Molecular Ecology Resources*, *11*, 739–742. https://doi.org/10.1111/j.1755-0998.2011.02997.x

Wagner, A. P., Creel, S., & Kalinowski, S. T. (2006). Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, *97*, 336–345. https://doi.org/10.1038/sj.hdy.6800865

Wang, J. L. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, *160*, 1203.

Weir, B. S., & Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution*, *38*, 1358–1370.

Yu, L., & Zhang, Y. P. (2006). The unusual adaptive expansion of pancreatic ribonuclease gene in carnivora. *Molecular Biology and Evolution*, *23*, 2326–2335. https://doi.org/10.1093/molbev/msl101

Zhang, J. (2003). Evolution by gene duplication: An update. *Trends in Ecology & Evolution*, *18*, 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, J. Z. (2006). Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nature Genetics*, *38*, 819. https://doi.org/10.1038/ng1812

Zhang, P., Huang, K., Zhang, B., Dunn, D. W., Chen, D., Li, F., … Li, B. (2018). High polymorphism in MHC-*DRB* genes in golden snub-nosed monkeys reveals balancing selection in small, isolated populations. *BMC Evolutionary Biology*, *18*, 29. https://doi.org/10.1186/s12862-018-1148-7

Ziegler, A., Ehlers, A., Forbes, S., Trowsdale, J., Volz, A., Younger, R., & Beck, S. (2000). Polymorphisms in olfactory receptor genes: A cautionary note. *Human Immunology*, *61*, 1281–1284. https://doi.org/10.1016/S0198-8859(00)00219-6

Zimmerman, P. A., Carrington, M. N., & Nutman, T. B. (1993). Exploiting structural differences among heteroduplex molecules to simplify genotyping the *DQA1* and *DQB1* alleles in human lymphocyte typing. *Nucleic Acids Research*, *21*, 4541–4547. https://doi.org/10.1093/nar/21.19.4541

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Huang K, Zhang P, Dunn DW, Wang T, Mi R, Li B. Assigning alleles to different loci in amplifications of duplicated loci. *Mol Ecol Resour.* 2019;19:1240–1253. https://doi.org/10.1111/1755-0998.13036