Received: 2 December 2019 | Accepted: 18 April 2021

DOI: 10.1111/2041-210X.13643

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

An approach for estimating haplotype diversity from sequences with unequal lengths

Ping Fan[1,2] | Jon Fjeldså[3] | Xuan Liu[4] | Yafei Dong[5] | Yongbin Chang[1,2] | Yanhua Qu[1] | Gang Song[1] | Fumin Lei[1,2,6]

[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China; [2]University of Chinese Academy of Sciences, Beijing, China; [3]Center for Macroecology, Evolution and Climate, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark; [4]Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China; [5]College of Life Sciences, Shaanxi Normal University, Xi'an, China and [6]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

Correspondence
Fumin Lei and Gang Song
Email: leifm@ioz.ac.cn; songgang@ioz.ac.cn

Funding information
The Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/Award Number: XDA19050202; National Natural Science Foundation of China, Grant/Award Number: 31572291, 31630069 and 32070434; The National Science and Technology Basic Resources Survey Program of China, Grant/Award Number: 2019FY100204; China Scholarship Council, Grant/Award Number: [2017]7011; The Second Tibetan Plateau Scientific Expedition and Research (STEP) Program, Grant/Award Number: 2019QZKK0304

Handling Editor: Antonino Malacrinò

Abstract

1. Genetic diversity is an essential component of biodiversity. Developing robust quantification methods is critically important in depicting the genetic diversity underlying the geographical distributions of species, especially for the sequence data with unequal lengths.
2. Traditional calculation of genetic diversity depends on sequences of equal length. However, many homologous sequences downloaded from online repositories vary in length, posing a significant challenge to quantify the genetic diversity, especially haplotype diversity. We developed a new approach independent of sequence length by applying the same parameters used in calculating nucleotide diversity to estimate haplotype diversity. We compared this novel approach with the calculations by the program DNAsp, and we used simulation data from terrestrial vertebrates (birds, mammals and amphibians) and *Homo sapiens* to validate the method's performance. We further applied this approach to explore the global latitudinal gradients of haplotype diversity in amphibians, mammals and birds, and compared the results by traditional methods.
3. The haplotype diversity calculated by our novel approach is consistent with the results from DNAsp. The simulations showed that our approach is robust and has a good estimating performance for sequence data with unequal lengths.
4. For the datasets of terrestrial vertebrates and *H. sapiens*, our approach is capable of estimating haplotype diversity with unequal intraspecific sequence lengths. In contrast to patterns based on traditional methods, we observed different latitudinal patterns of haplotype diversity between the northern and southern hemispheres for terrestrial vertebrates, which is consistent with the updated pattern of nucleotide diversity for mammals. The present work contributes to the development of more precise quantification methods, which may be broadly applied to assessing biogeographical patterns of genetic diversity.

KEYWORDS
genetic diversity, haplotype diversity, haplotype–nucleotide diversity relationship, nucleotide diversity, unequal length sequences

1 | INTRODUCTION

Quantifying genetic diversity and its spatial–temporal patterns is crucial for understanding a species' evolutionary history and population dynamics. The most widely used approaches for describing genetic diversity from mitochondrial DNA sequences are nucleotide diversity (π) and haplotype diversity (h, Goodall-Copestake et al., 2012; Miraldo et al., 2016). The calculation of genetic diversity traditionally depends on intraspecific sequences with equal lengths (hereafter, equal intraspecies sequence lengths). However, most sequences in public databases such as GenBank (https://www.ncbi.nlm.nih.gov/) and BOLD (http://www.boldsystems.org/) vary in length, which poses a significant challenge to quantify genetic diversity for the traditional methods. Recent developments in calculating genetic diversity have allowed the investigation of π for the data with unequal intraspecific sequence lengths (Miraldo et al., 2016). However, there is still no method or strategy to calculate h based on the unequal length sequences (Miraldo et al., 2016), which is another important parameter sculpting the property of genetic diversity.

Previous studies calculate π based on the pairwise comparison of nucleotide differences between distinct haplotypes (Goodall-Copestake et al., 2012; Nei & Li, 1979; Nei & Tajima, 1981). Since that h is mathematically related to the nucleotide diversity, it can be inferred by estimating nucleotide diversity. For instance, a previous study using *Cytochrome oxidase subunit I* (*COI*) genes from 23 animal species revealed a quantitative relationship between nucleotide diversity and haplotype diversity as $\pi = 0.008h^2$ (Goodall-Copestake et al., 2012). This suggested that h might be estimated from π by constructing a model that combines π and h. Obtaining accurate values of π and h is a prerequisite for building a model incorporating the two parameters. However, since in data with unequal intraspecies sequence lengths, π can only be approximately estimated, the existing methods for estimating h cannot handle such data. To the best of our knowledge, such a model-dependent approach is not available.

Tajima has uncovered the relationship between segregating sites and π (Tajima, 1989, 1993), which provides a new idea to estimate h. Segregating sites between two randomly chosen sequences is the criteria to judge whether those two sequences are different haplotypes. Based on this, we developed a new approach using the same parameters as is used for π to estimate h from mitochondrial DNA sequences. The approach is capable of estimating h without summarizing population-level haplotypes and their relative frequencies. The advantage of this method is that we can now analyse h for sequence data of variable lengths, thus dealing with the main shortcoming of the existing methods. We therefore used this novel method to explore the theoretical relationship between haplotype and nucleotide diversity. We evaluated the performance of our method in terrestrial vertebrates (amphibians, birds and mammals) and *Homo sapiens* for data with unequal intraspecies sequence lengths. Finally, we applied the method to analyse of latitudinal patterns of h based on *Cytochrome* b (*CYTB*) and *COI* genes of terrestrial vertebrates. Our work may promote further quantification and understanding of global genetic diversity using multiple metrics (e.g. π and h) regardless of sequence length given the increasing availability of genetic data.

2 | MATERIALS AND METHODS

2.1 | Conception of the algorithm for genetic diversity measures

Nucleotide diversity is the average number of nucleotide differences per site between two randomly chosen sequences (following Nei & Li, 1979), and is defined as.

$$\pi = \sum_{ij} x_i x_j k_{ij}, \tag{1}$$

where x_i is the frequency of the ith sequence in the population, and k_{ij} is the number of nucleotide differences per site between the ith and jth sequences. In practice, the sample size (n) is often very small (Nei & Li, 1979; Nei & Tajima, 1981). In this case, the nucleotide diversity should be estimated by

$$\pi = \frac{n}{n-1} \sum_{ij} x_i x_j k_{ij}. \tag{2}$$

Haplotype diversity is the probability that two randomly chosen haplotypes are different (de Jong et al., 2011; Harris & DeGiorgio, 2017; Nei & Roychoudhury, 1974; Nei & Tajima, 1981). The estimator of expected h (de Jong et al., 2011; Harris & DeGiorgio, 2017; Nei & Roychoudhury, 1974; Nei & Tajima, 1981) is.

$$h = \frac{n}{n-1} \left(1 - \sum_i p_i^2\right), \tag{3}$$

where p_i is the (relative) haplotype frequency of each haplotype in the sample and n is the sample size.

In Equation 2, the ith sequence is equivalent to the ith haplotype in the population; x_i in Equation 2 also represents the (relative) frequency of each haplotype in the sample. Hence, Equation 2 can be written as.

$$\pi = \frac{n}{n-1} \sum_{ij} p_i p_j k_{ij}, \tag{4}$$

where p_i is the (relative) frequency of each haplotype in the sample, and k_{ij} is the number of nucleotide differences per site between the ith and jth haplotypes.

Consider the comparison of two randomly chosen sequences, which would produce two situations:

$$k_{ij} = \begin{cases} 0 \text{ sequence } i \text{ and sequence } j \text{ are the same haplotype} \\ !0 \text{ sequence } i \text{ and sequence } j \text{ are different haplotypes} \end{cases}.$$

As $k_{ii} = 0$, Equation 4 can be simplified to

$$\pi = \frac{n}{n-1}\sum_{i \neq j} p_i p_j k_{ij}. \tag{5}$$

Thus, the estimated value of π is based on the (relative) frequency of each haplotype in the sample, which means that the haplotype information has significant influence on the estimated value of π. In practice, the number of sampled haplotypes is often very small, so the expectation of $\sum_{i \neq j} P_{ij} k_{ij}$ becomes $\frac{n}{n-1}\sum_{i \neq j} p_i p_j k_{ij}$, where P_{ij} denotes the situation of two randomly chosen sequences belonging to haplotype i and haplotype j (Nei & Tajima, 1981). Hence Equation 5 can be written as.

$$\pi = \sum_{i \neq j} P_{ij} k_{ij}. \tag{6}$$

Interestingly, the non-simple algorithm for h is.

$$h = \frac{n}{n-1}\sum_{i \neq j} p_i p_j, \tag{7}$$

or

$$h = \sum_{i \neq j} P_{ij}. \tag{8}$$

It may be informative to compare Equations 5 and 7 or Equations 6 and 8, as the comparisons demonstrate that π and h are both influenced by the (relative) frequency of each haplotype in the sample.

In this section, we construct an estimator of h using the value of k_{ij} in the sample. As Equations 5 and 7 (or Equations 6 and 8) show, the difference between the equations for π and h lies in the definition of k_{ij}. If we offset the k_{ij} in Equations 7 and 8, we can obtain a new method for calculating h. This minor trick can help us realize that in Equations 5 and 6, $i \neq j$, which means that $k_{ij} \neq 0$. Here we focus on counting the number of $k_{ij} \neq 0$ in the sample rather than the exact value of k_{ij}. For this we consider the situation of two random sequences belonging to different haplotypes. Let $\widehat{D_{ij}}$ denotes the characteristic value of $k_{ij}(k_{ij} = 0 \text{ or } k_{ij} \neq 0)$ and is expressed as.

$$\widehat{D_{ij}} = \begin{cases} 1 \; k_{ij} \neq 0 \; i \neq j \\ 0 \; k_{ij} = 0 \; i = j \end{cases}.$$

Here we use the $\widehat{D_{ij}}$ to replace the k_{ij}, and thus we can transform the equation from π to h.

$$h = \frac{n}{n-1}\sum_{i \neq j} p_i p_j \widehat{D_{ij}}, \tag{9}$$

or

$$h = \sum_{i \neq j} P_{ij} \widehat{D_{ij}}. \tag{10}$$

In the previous process, if we calculate h, we need to count the number of haplotypes and their relative frequencies in the sample. If two different sequences represent different haplotypes, the nucleotide differences between these two sequences will be greater than zero (Nei & Tajima, 1981). Hence, the two randomly chosen haplotypes can be seen as two randomly chosen sequences. We can obtain a single new equation,

$$h = \frac{M_{k_{ij}>0}}{\begin{pmatrix} n \\ 2 \end{pmatrix}}, \tag{11}$$

where k_{ij} is the number of different nucleotides per site between sequence i and sequence j, $M_{k_{ij}>0}$ is the total number of $k_{ij} > 0$ in the set of pairwise comparisons, and $\begin{pmatrix} n \\ 2 \end{pmatrix}$ is the number of pairwise comparisons made.

2.2 | Nucleotide–haplotype relationship analysis

We found that h and π can be estimated by the same parameters, thus providing a chance of quantifying the relationship between h and π. It has been reported that π and h are positively related (Goodall-Copestake et al., 2012). Hence, we can define $\pi = kh$. According to Equations 2 and 11 , it can be inferred that (Supporting Information):

$$k = \frac{\mathrm{Sum}_{k_{ij}>0}}{M_{k_{ij}>0}}, \tag{12}$$

where k_{ij} is the number of different nucleotides per site between sequence i and sequence j, $\mathrm{Sum}_{k_{ij}>0}$ is the sum of $k_{ij} > 0$ in the pairwise comparisons, $M_{k_{ij}>0}$ is the total number of $k_{ij} > 0$ in the pairwise comparisons, and $\begin{pmatrix} n \\ 2 \end{pmatrix}$ is the number of pairwise comparisons.

2.3 | Case study #1: Performance of the method with equal sequence length data

To investigate whether our method would achieve the computational performance of previous methods (equal sequence length data), we used 951 species (97 species for *Amphibians COI*, 87 species for *Amphibians CYTB*, 242 species for *Birds COI*, 79 species for *Birds CYTB*, 244 species for *Mammals COI* and 201 species for *Mammals CYTB*) and artificial DNA sequences with a length of 20 nucleotides (Tajima, 1993), based on an R script (Fan et al., 2021a) developed to calculate *h*. Besides, we also calculated the *h* using the 'DNA Polymorphism' function in DNAsp V5 (Librado & Rozas, 2009). Before calculations, DNA sequences for each species were aligned using MUSCLE (Edgar, 2004) with the default setting. We assessed the accuracy of our approach by comparing the *h* estimated by DNAsp V5 (Librado & Rozas, 2009) using the Mann–Whitney *U* test (wilcox.test in R).

2.4 | Case study #2: The estimation performance with unequal sequence length data

The k_{ij} values are calculated using the statistics for the nucleotide differences in the overlapping range of sequence *i* and sequence *j*. Most data in online public repositories have different intraspecific sequence lengths, so that the overlapping regions in different pairwise comparisons (two randomly chosen sequences) will be different (Fan et al., 2021b). For this type of data, Miraldo et al. (2016) adjusted the k_{ij} in Equation 2 to K_{ij}/m_{ij} to calculate the nucleotide diversity. Here we use the same strategy for Equation 11 to obtain.

$$h = \frac{M_{K_{ij/m_{ij}>0}}}{\binom{n}{2}}, \tag{13}$$

where K_{ij} is the number of nucleotide differences between sequence *i* and sequence *j*, and m_{ij} is the overlap length between sequence *i* and sequence *j*. $M_{K_{ij/m_{ij}>0}}$ is the total number of $K_{ij}/m_{ij} > 0$ in the pairwise comparisons, and $\binom{n}{2}$ is the number of pairwise comparisons.

To evaluate the performance of Equation 13 for data with unequal intraspecies sequence lengths, we first calculated *h* using data (case study #1) with equal sequence lengths. We then performed a 'Random Length analysis' using the same sequence data to simulate different degrees of overlap in the sequences by randomly choosing the sequence length used to calculate each pairwise result for the species, using a threshold of 50% of the original length (Figure 1). This procedure was repeated 100 times for each species. To validate the accuracy and stability of Equation 13, we estimated the accuracy of each repeat (R_i) for each species by the relative error:

$$R_i = \frac{|e_i - e_0|}{e_0}, \tag{14}$$

where e_i is the estimated value for the *i*th repeat for different simulated lengths, and e_0 is the value for equal length sequences.

We used the mean value of R_i for each species to determine the accuracy and the standard deviation (*SD*) to measure the stability. The same data were used to evaluate the performance of the new method for quantifying π for the unequal intraspecific sequence length data, and the accuracy and stability of *h* and π were compared using Mann–Whitney *U* tests (wilcox.test in R). Additionally, we reported the average accuracy and stability values for three terrestrial vertebrate groups (*Birds*, *Mammals* and *Amphibians*), a popular method used to represent a specific region's diversity (areas being

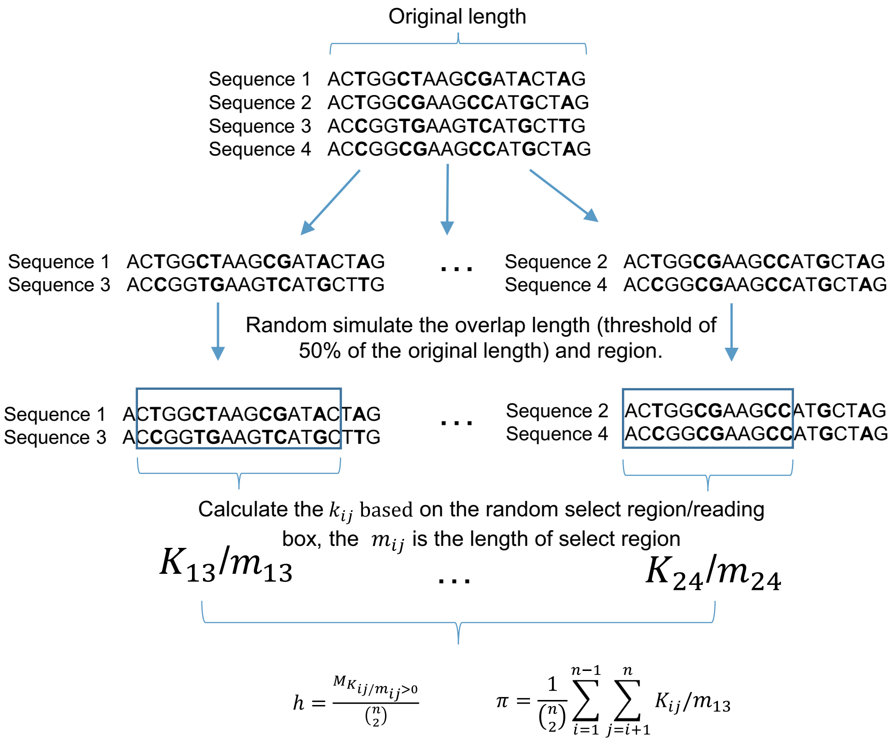
Original length
Sequence 1 ACTGGCTAAGCGATACTAG
Sequence 2 ACTGGCGAAGCCATGCTAG
Sequence 3 ACCGGTGAAGTCATGCTTG
Sequence 4 ACCGGCGAAGCCATGCTAG
Sequence 1 ACTGGCTAAGCGATACTAG
Sequence 3 ACCGGTGAAGTCATGCTTG
. . .
Sequence 2 ACTGGCGAAGCCATGCTAG
Sequence 4 ACCGGCGAAGCCATGCTAG
Random simulate the overlap length (threshold of 50% of the original length) and region.
Sequence 1 ACTGGCTAAGCGATACTAG
Sequence 3 ACCGGTGAAGTCATGCTTG
. . .
Sequence 2 ACTGGCGAAGCCATGCTAG
Sequence 4 ACCGGCGAAGCCATGCTAG
Calculate the k_{ij} based on the random select region/reading box, the m_{ij} is the length of select region
K_{13}/m_{13}
. . .
K_{24}/m_{24}
$h = \frac{M_{K_{ij}/m_{ij}>0}}{\binom{n}{2}}$
$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K_{ij}/m_{13}$

FIGURE 1 Demonstration of the random length analysis for genetic diversity using equal length sequences. The blue box indicates simulation of the overlap region of each pairwise comparison of sequences

grid cells or latitudinal bands) in recent studies (Gratton, Marta, Bocksberger, Winter, Keil, et al., 2017; Millette et al., 2019; Miraldo et al., 2016).

Additionally, we used gene fragments (*CYTB*, *COX3*, *D-loop* and *HVR-1*) and the complete mitochondrial genome of *H. sapiens* to run the 'Random Length analysis' to test of our method with a broader range of sequence datasets.

2.5 | Case study: #3 An application to the latitudinal gradient of haplotype diversity in terrestrial vertebrates

We used our method to explore the latitudinal gradient of *h* across three terrestrial vertebrate groups. The mitochondrial DNA sequences were downloaded following the procedures used by Miraldo et al. (2016) and Gratton, Marta, Bocksberger, Winter, Trucchi, et al. (2017). For the sequences only marked by geographical information, we first quoted the coordinate information for mammals and amphibians provided by Miraldo et al. (2016). After that, we used the GEOCODE package in R (https://github.com/scottcame/geocode) with the Google map to assign geographical coordinates to the sequences for birds and the remaining sequences for mammals and amphibians (51.06%). Finally, we acquired 96,530 records annotated with geographical coordinates (3,911 species for birds, 2,263 species for mammals and 1,628 species for amphibians), representing approximately 46.3% of the sequences obtained from GenBank and BOLD. We used the species' ranges to select the sequences (Miraldo et al., 2016), but only screened those from breeding populations for migratory birds. The IUCN species range was used for mammals and amphibians (IUCN Red List of Threatened Species, 2018), and the species range map for birds was downloaded from BirdLife International and Handbook of the Birds of the World (2017). We discarded the species range polygons annotated by 'possibly extant', 'presence uncertain', 'introduced' and 'vagrant'. Sequence alignment for each species was performed in MUSCLE (Edgar, 2004) with a default setting. To reduce the over- or under-estimation of haplotype diversity, we excluded species with sample sizes fewer than five (Goodall-Copestake et al., 2012; Miraldo et al., 2016). The *h* and π were calculated by the 10° latitude band (Miraldo et al., 2016). Since not all sequences were completely overlapping, we only counted pairwise comparisons where sequences overlapped by at least 50% of the longer sequence (Miraldo et al., 2016). We first estimated the GD values of each species. Then the *h* value of each latitudinal band (LB_h) was calculated by the average value of *h* for multiple species that falling in each latitudinal band (Equation 15). We also calculated the value of π for each latitude band by Equation 16 (Miraldo et al., 2016). These equations are given by.

$$\mathrm{LB}_h = \frac{1}{s}\sum_{i=1}^{s} h_i, \tag{15}$$

$$\mathrm{LB}_\pi = \frac{1}{s}\sum_{i=1}^{s} \pi_i \tag{16}$$

where *s* is the number of species falling in the latitudinal band, π_i is the value of π for species *i* in that latitudinal band and h_i is the value of *h* for species *i* in that latitudinal band.

We applied a beta regression linear model (Ferrari & Cribari-Neto, 2004; Miraldo et al., 2016) to investigate the relationship between genetic diversity and latitude. Considering that previous research has demonstrated a poleward decreasing trend in π, for simplicity, we evaluated whether the model (formula: nucleotide diversity ~ latitude + latitude^2; weights: D) used to account for the π along with the latitude also applies for *h*. The linear and quadratic terms of latitude were introduced to better fit the relationship between latitude and genetic diversity because the peak value of genetic diversity is not always at the equator and latitude values are opposite between the northern and southern hemispheres. As the similarity of two gene copies tends to decay with their geographical distance (Gratton, Marta, Bocksberger, Winter, Keil, et al., 2017), we also calculated the geographical distance between pairs of conspecific sequences (D) using the 'distm()' function with the method of 'distVincentyEllipsoid' provided by GEOSPHERE packages (Hijmans et al., 2019) and introduced it as a weight in the models to eliminate the influence of geographical distance from our results. We ran independent models for amphibians, mammals and birds, with *COI* and *CYTB* as examples for testing.

3 | RESULTS

3.1 | Case study #1

We obtained the same estimated values of *h* from our method and DNAsp V5 (Figure S1). With the equal intraspecies sequence length data for 951 species, the Mann–Whitney *U* test revealed no differences between DNAsp V5 and our method in *h* across *Birds*, *Mammals* and *Amphibians* (*Birds COI*, $W = 20{,}402$, $p = 1$; *Birds CYTB*, $W = 2{,}245$, $p = 1$; *Mammals COI*, $W = 22{,}898$, $p = 1$; *Mammals CYTB*, $W = 13{,}612$, $p = 1$; *Amphibians COI*, $W = 2{,}888$, $p = 1$; *Amphibians CYTB*, $W = 3{,}362.5$, $p = 1$), indicating that our new method is robust in performing analyses with equal intraspecies sequence lengths for *h* (Figure S1A).

3.2 | Case study #2

The results of our method on random length analysis for 951 species showed that the accuracy of the *h* estimation was significantly higher than that for π (Figure 2a) in all terrestrial vertebrate groups except for *Birds* at *CYTB* (*Birds COI*, $W = 33{,}649$, $p < 0.05$; *Birds CYTB*, $W = 3{,}377$, $p = 0.372$; *Mammals COI*, $W = 36{,}598$, $p < 0.001$; *Mammals CYTB*, $W = 29{,}711$, $p < 0.001$; *Amphibians COI*, $W = 5{,}674$, $p < 0.05$; *Amphibians CYTB*, $W = 4{,}871$, $p < 0.01$; Mann–Whitney *U* tests). The results indicate that *h* may be more suitable for unequal sequence length data. The stability of the *h* estimates was significantly higher than the π (Figure 2b) in *Mammals* and *Birds* (except for *Birds* at *CYTB*)

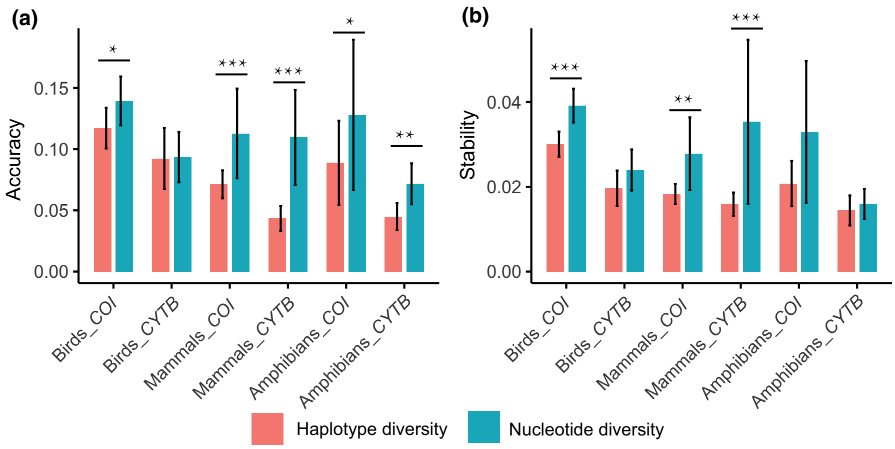

FIGURE 2 The accuracy (a) and stability (b) of h and π based on unequal sequence lengths. Stability of each class, presented as a column diagram, was estimated by the average value of the standard deviation of the relative error for each species; accuracy, presented as a column diagram, was estimated by the average value of the mean of the relative error of each species. Error bar denotes 95% confidence interval. Because the accuracy and stability were estimated from the relative error, a lower value indicates a better performance

but not in *Amphibians* (*Birds COI*, $W = 34{,}824$, $p < 0.001$; *Birds CYTB*, $W = 3{,}572$, $p = 0.116$; *Mammals COI*, $W = 34{,}356$, $p < 0.01$; *Mammals CYTB*, $W = 26{,}307$, $p < 0.001$; *Amphibians COI*, $W = 5{,}215$, $p = 0.19$; *Amphibians CYTB*, $W = 4{,}317$, $p = 0.289$; Mann–Whitney *U* tests), suggesting that variation in sequence length influences the estimates of h as well as π. Additionally, when we used the average mean of the genetic diversity to represent a specific taxon, the stability and accuracy values of h were smaller than those of π (Figure 2).

Consistent with our result for mammals and birds, we found that the stability and accuracy values of h were smaller than those of π in both gene fragments (*CYTB*, *COX3*, *D-loop* and *HVR-1*) and the complete mitochondrial genome of *H. sapiens* (Table S1). These results suggest that our method is reliable for calculating h from unequal intraspecies sequence length data in broader-resourced sequence datasets.

3.3 | Case study #3

Relying on the new method, we found a significant latitudinal gradient in h. The beta regression (formula: haplotype diversity/nucleotide diversity ~ latitude + latitude^2; weights: D) had a significant negative quadratic term (Figures 3 and 4, Table 1). The results showed that the latitudinal gradient of h was similar between the *COI* and *CYTB* genes for each group. We observed a poleward decrease of π for birds (Figures 3b and 4b), mammals (Figures 3d and 4d) and amphibians (Figure 4f). However, the h followed different poleward patterns between the northern and southern hemispheres (Figures 3a,c,e and 4a,c,e). In particular, there was a decreasing trend of h in the southern hemisphere and a weaker trend in the northern hemisphere for birds (Figures 3a and 4a), but the trends in both mammals and amphibians were opposite (Figures 3c,e and 4c,e).

4 | DISCUSSION

Genetic diversity, consisting of π and h, is the most fundamental measure of biodiversity (Robert, 1994). Over the last few years, measuring genetic diversity has become an attractive topic in biodiversity and conservation research because of the rapid increase in the availability of large DNA sequences (Gratton, Marta, Bocksberger, Winter, Keil, et al., 2017; Millette et al., 2019; Miraldo et al., 2016). However, current methods (e.g. DNAsp V5) are not available for calculating haplotype diversity from unequal length datasets, underlining the necessity for a new method. As early as Tajima (1983 and 1989), Tajima discussed the relationship between segregating sites and nucleotide diversity, and described how π can be estimated from the pairwise heterozygosity (segregating sites). Thanks to Tajima's contribution and this inspiration, we propose a method to estimate h from the pairwise segregating sites. It should be pointed out that the theoretical basis of our work (the segregating sites between two randomly chosen sequences) is similar to Tajima's, but the advantage is that the new approach for h presented here does not require summarizing the number of haplotypes and their relative frequencies. As a result, this method is that it can deal with sequences of varied lengths, thereby augment the existing methods that cannot employ data with unequal intraspecies sequence lengths. We propose that this study would provide a new tool to explore the spatial patterns of h.

The performance of our method was consistent with the other methods for h estimation (Figure S1). The new method performs well with unequal intraspecies sequence length data (Figure 2), and the h deriving from the sequences with few nucleotides missing is associated with high accuracy (Figure S2). Although the variability in sequence lengths may influence both π and h, we found that the accuracy of h is higher than that of π, indicating that the h may be an ideal metric genetic diversity parameter for estimating genetic diversity from unequal sequence length data. However, when the number of DNA sequences is large, the calculation of genetic diversity based on k_{ij} is time-consuming (Tajima, 1993), this is also a limitation of our method. Hence, we would suggest that future studies should explore a time-saving approach in calculating genetic diversity from big data sequences of unequal lengths.

The method presented in this paper uses the same parameters as for π to estimate h from mitochondrial DNA sequences. Moreover, spatial analyses across taxa provide robust evidence for

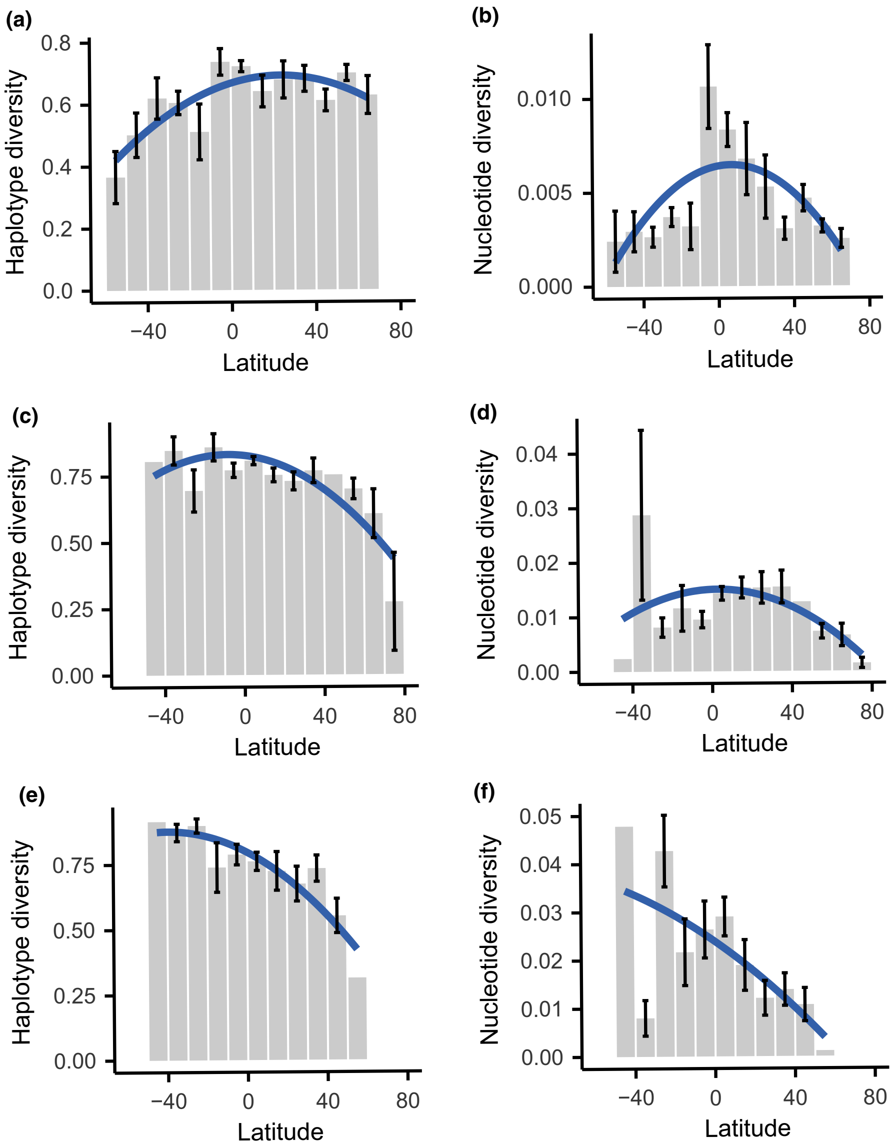

FIGURE 3 Distribution of genetic diversity for *COI* across latitudes for birds (*h*: a, π: b), mammals (*h*: c, π: d) and amphibians (*h*: e, π: f). The plots show genetic diversity (*y*-axis) against latitude (*x*-axis), and the regression lines (blue lines) represent the predictions of the beta regression model. The results illustrate the different decreasing trends of *h* and π towards the poles. The patterns of haplotype diversity are different between the northern and southern hemispheres

a context-dependent relationship between π and *h*, a relationship that previously has been suggested as strictly positive. Based on our mathematical model ($\pi = kh$), *k* is not a constant value (defined in Equation 12), which depends on the degree of difference between samples and the sample size (k_{ij}), causing a variable correlation of nucleotide–haplotype diversity. This might be one reason why it is difficult to observe consistent directional patterns (negative or positive; Bird et al., 2007; Song et al., 2013; Wang et al., 2013; Zhang et al., 2012). The different patterns for both π and *h* between the northern and southern hemispheres across taxa further validated the variable relationship between π and *h*. A previous study showing high *h* and low π might be the signature of recently diverged populations (Garg & Mishra, 2018; Song et al., 2014). The parameter *k* here could thus be applied to indicate the potential population divergence; a small *k* value suggests recent divergence of populations.

An example application of this method, testing the genetic diversity of terrestrial vertebrates, showed a higher *h* in the tropics, with different decreasing trends towards the poles between the northern and southern hemispheres for each taxon (Figures 3 and 4). This is consistent with the updated π pattern for mammals (Gratton, Marta, Bocksberger, Winter, Keil, et al., 2017), suggesting that our method is robust to assessment of the pattern of genetic diversity. Furthermore, the average overlap percentage in our database (94.37%) is larger than those in our simulation (74.90%), suggesting the *h* derived from the real database may be more accurate than we expected (Figure S2). The difference in patterns between *h* and π, at least for birds and mammals in our analyses, suggests the necessity of multi-metrics analyses for both π and *h* for fully understanding the global pattern of genetic diversity. Here, we briefly discuss the potential explanations for different declining trends of *h* in the northern and southern hemispheres. This might reflect variation in the strength of natural selection for different haplotypes (Camus et al., 2017) and in the migratory ability of species (Gratton, Marta, Bocksberger, Winter, Keil, et al., 2017). For example, there are differences in the configuration of land and ocean between the northern and southern hemispheres, and sea surface temperatures are more

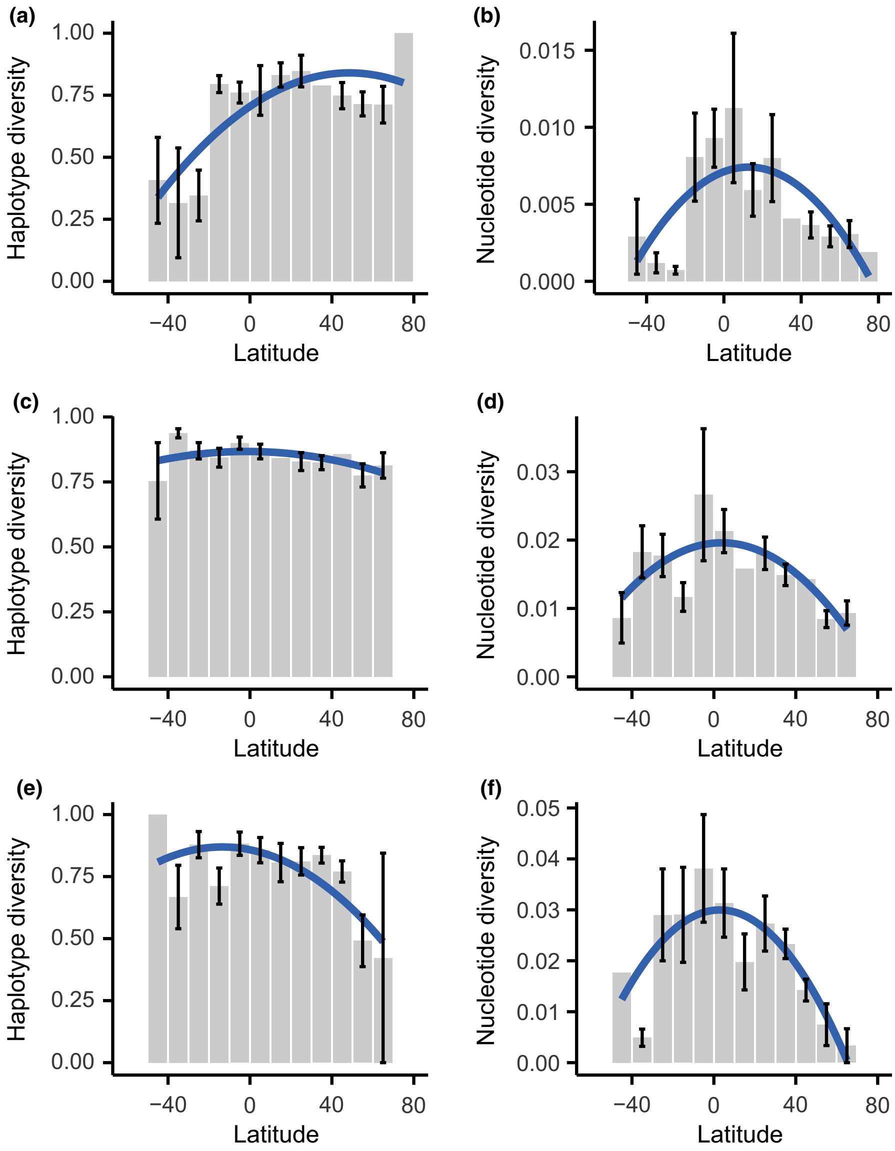

FIGURE 4 Distribution of genetic diversity for *CYTB* across latitudes for birds (*h*: a, π: b), mammals (*h*: c, π: d) and amphibians (*h*: e, π: f). The plots show genetic diversity (*y*-axis) against latitude (*x*-axis), and the regression lines (blue lines) represent the predictions of the beta regression model. The results illustrate the different decreasing trends of genetic diversity (*h* and π) towards the poles in mammals, birds and amphibians. The patterns in haplotype diversity are different between the northern and southern hemispheres

TABLE 1 Beta regression model for genetic diversity and latitude. Results from beta regression had significant negative quadratic terms, indicating significant latitudinal gradients of haplotype diversity and nucleotide diversity

		Haplotype diversity			Nucleotide diversity		
	Locus	Pseudo R^2	β_1	β_2	Pseudo R^2	β_1	β_2
Birds	*COI*	0.7073	−3.51e-03***	−1.81e-04***	0.4743	2.48e-04***	−2.88e-05***
Mammals	*COI*	0.6767	8.58e-03***	−1.79e-04***	0.3397	4.48e-04***	−5.38e-05***
Amphibians	*COI*	0.8417	−2.05e-02***	−1.85e-04***	0.4906	−6.31e-03***	−5.19e-06***
Birds	*CYTB*	0.2271	2.26e-02***	−4.65e-04***	0.4863	9.64e-04***	−5.36e-05***
Mammals	*CYTB*	0.159	1.99e-03***	−1.58e-04***	0.5599	7.25e-04***	−6.83e-05***
Amphibians	*CYTB*	0.4143	−9.07e-03***	−3.11e-04***	0.6814	−7.66e-05***	−1.21e-04***

Abbreviations: β_1, Coefficient of primary term; β_2, Coefficient of quadratic term.

$^{***}p < 0.001$.

stable owing to the smaller land area in the south than in the north (Fordham et al., 2017; Hong et al., 2019). This may be one potential explanation for the greater diversity in the southern hemisphere, at least for mammals and amphibians. Previous studies have shown that there is a higher bird species turnover in the northern hemisphere in response to global climate change (Virkkala & Lehikoinen, 2017), indicating that the haplotype assemblage of the northern species would be more sensitive to climate change, and there may be relatively higher *h* in the northern hemisphere for birds.

In summary, we present a novel method of estimating *h* that can use unequal intraspecies sequence length data. The three case studies confirmed that our method had high estimation reliability. Despite the fact that measuring genetic diversity based on varied sequence lengths is only an approximation, our method lays a solid foundation for more precise quantification of genetic diversity patterns that can accommodate the accelerated availability of genetic data and thereby lead to a more fruitful application of multi-metrics (e.g. π and *h*) in the era of big data.

ACKNOWLEDGEMENTS

We thank Xiaolu Jiao and Xin Yu for assistance with data analysis, Weiwei Zhai, Liang Ma and Hechuan Yang for discussion and Huijie Qiao for his generous help during our revision. This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19050202 to F.L.), the Second Tibetan Plateau Scientific Expedition and Research (STEP) Program (2019QZKK0304 to F.L. and G.S.), National Science Foundation of China (32070434 & 31572291 to G.S.; 31630069 to F.L.), the National Science and Technology Basic Resources Survey Program of China (2019FY100204 to F.L. and P.F.) and China Scholarship Council, Grant/Award Number: [2017]7011 to P.F.

AUTHORS' CONTRIBUTIONS

F.L., P.F. and J.F. conceived the idea and designed the methodology. P.F. collected the data. P.F., X.L. and Y.D. analysed the data. P.F. wrote the original draft. P.F., J.F., G.S., X.L., Y.C., Y.Q. and F.L. reviewed and edited the text. All authors contributed constructive comments and approved the submitted version of this manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13643.

DATA AVAILABILITY STATEMENT

The dataset and R codes used in this study can be found in the repository: https://github.com/PingFan6/estimating-haplotype-diversity. These files have been archived on Zendo (https://doi.org/10.5281/zenodo.4722108). The summary of the sequence lengths for each species in online datasets used in this study deposited in the Dryad Repository (https://doi.org/10.5061/dryad.ghx3ffbnt).

ORCID

Ping Fan https://orcid.org/0000-0002-3794-4676
Xuan Liu https://orcid.org/0000-0003-1572-1268
Yanhua Qu https://orcid.org/0000-0002-4590-7787

REFERENCES

Bird, C. E., Holland, B. S., Bowen, B. W., & Toonen, R. J. (2007). Contrasting phylogeography in three endemic Hawaiian limpets (*Cellana* spp.) with similar life histories. *Molecular Ecology*, *16*, 3173–3186. https://doi.org/10.1111/j.1365-294X.2007.03385.x

BirdLife International and Handbook of the Birds of the World. (2017). *Bird species distribution maps of the world*. Version 7.0. Retrieved from http://datazone.birdlife.org/species/requestdis

Camus, M. F., Wolff, J. N., Sgrò, C. M., & Dowling, D. K. (2017). Experimental support that natural selection has shaped the latitudinal distribution of mitochondrial haplotypes in Australian drosophila melanogaster. *Molecular Biology and Evolution*, *34*, 2600–2612. https://doi.org/10.1093/molbev/msx184

de Jong, M. A., Wahlberg, N., van Eijk, M., Brakefield, P. M., & Zwaan, B. J. (2011). Mitochondrial DNA signature for range-wide populations of *Bicyclus anynana* suggests a rapid expansion from recent refugia. *PLoS ONE*, *6*, e21385. https://doi.org/10.1371/journal.pone.0021385

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340

Fan, P., Fjeldså, J., Liu, X., Dong, Y. F., Chang, Y. B., Qu, Y. H., Song, G., & Lei, F. M. (2021a). Data from: PingFan6/estimating-haplotype-diversity: Estimating haplotype diversity. *Zenodo*, https://doi.org/10.5281/zenodo.4722108

Fan, P., Fjeldså, J., Liu, X., Dong, Y. F., Chang, Y. B., Qu, Y. H., Song, G., & Lei, F. M. (2021b). Data from: An approach for estimating haplotype diversity from sequences with unequal lengths. *Methods in Ecology and Evolution*. https://doi.org/10.5061/dryad.ghx3ffbnt

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*, 799–815. https://doi.org/10.1080/0266476042000214501

Fordham, D. A., Saltré, F., Haythorne, S., Wigley, T. M. L., Otto-Bliesner, B. L., Chan, K. C., & Brook, B. W. (2017). PaleoView: A tool for generating continuous climate projections spanning the last 21 000 years at regional and global scales. *Ecography*, *40*, 1348–1358. https://doi.org/10.1111/ecog.03031

Garg, R. K., & Mishra, V. (2018). Molecular insights into the genetic and haplotype diversity among four populations of *Catla catla* from Madhya Pradesh revealed through mtDNA cyto b gene sequences. *Joural of Genetic Engineering and Biotechnology*, *16*, 169–174. https://doi.org/10.1016/j.jgeb.2017.11.003

Goodall-Copestake, W. P., Tarling, G. A., & Murphy, E. J. (2012). On the comparison of population-level estimates of haplotype and nucleotide diversity: A case study using the gene cox1 in animals. *Heredity*, *109*, 50–56. https://doi.org/10.1038/hdy.2012.12

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Keil, P., Trucchi, E., & Kuhl, H. (2017). Which latitudinal gradients for genetic diversity? *Trends in Ecology & Evolution*, *32*, 724–726. https://doi.org/10.1016/j.tree.2017.07.007

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E., & Kühl, H. (2017). A world of sequences: Can we use georeferenced nucleotide databases for a robust automated phylogeography? *Journal of Biogeography*, *44*, 475–486. https://doi.org/10.1111/jbi.12786

Harris, A. M., & DeGiorgio, M. (2017). An unbiased estimator of gene diversity with improved variance for samples containing related and inbred individuals of any ploidy. *G3 Genes|genomes|genetics*, *7*, 671–691. https://doi.org/10.1534/g3.116.037168

Hijmans, R. J., Williams, E., & Vennes, C. (2019). *geosphere: Spherical trigonometry*. R package version 1.5-10. Retrieved from https://CRAN.R-project.org/package=geosphere

Hong, B., Rabassa, J., Uchida, M., Hong, Y., Peng, H., Ding, H., Guo, Q., & Yao, H. (2019). Response and feedback of the Indian summer monsoon and the Southern Westerly Winds to a temperature contrast between the hemispheres during the last glacial–interglacial transitional period. *Earth-Science Reviews*, *197*, 102917. https://doi.org/10.1016/j.earscirev.2019.102917

IUCN Red List of Threatened Species. (2018). *The IUCN Red List of Threatened Species*. Version 6.1. Retrieved from http://www.iucnredlist.org

Librado, P., & Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*, 1451–1452. https://doi.org/10.1093/bioinformatics/btp187

Millette, K. L., Fugere, V., Debyser, C., Greiner, A., Chain, F. J. J., & Gonzalez, A. (2019). No consistent effects of humans on animal genetic diversity worldwide. *Ecology Letters*, *23*(1), 55–67. https://doi.org/10.1111/ele.13394

Miraldo, A., Li, S., Borregaard, M. K., Florez-Rodriguez, A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K. A., & Nogues-Bravo, D. (2016). An Anthropocene map of genetic diversity. *Science*, *353*, 1532–1535. https://doi.org/10.1126/science.aaf4381

Nei, M., & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*, 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Nei, M., & Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics*, *76*, 379–390. https://doi.org/10.1093/genetics/76.2.379

Nei, M., & Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, *97*, 145–163. https://doi.org/10.1093/genetics/97.1.145

Robert, M. M. (1994). Biological diversity: Differences between land and sea. *Philosophical Transactions: Biological Sciences*, *343*, 105–111.

Song, G., Yu, L. J., Gao, B., Zhang, R. Y., Qu, Y. H., Lambert, D. M., Li, S., Zhou, T. L., & Lei, F. M. (2013). Gene flow maintains genetic diversity and colonization potential in recently range-expanded populations of an Oriental bird, the Light-vented Bulbul (*Pycnonotus sinensis*, Aves: Pycnonotidae). *Diversity and Distributions*, *19*, 1248–1262. https://doi.org/10.1111/Ddi.12067

Song, J., Hou, F., Zhang, X., Yue, B., & Song, Z. (2014). Mitochondrial genetic diversity and population structure of a vulnerable freshwater fish, rock carp (*Procypris rabaudi*) in upper Yangtze River drainage. *Biochemical Systematics and Ecology*, *55*, 1–9. https://doi.org/10.1016/j.bse.2014.02.008

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, *105*, 437–460. https://doi.org/10.1093/genetics/105.2.437

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*, 585–595. https://doi.org/10.1093/genetics/123.3.585

Tajima, F. (1993). Statistical analysis of DNA polymorphism. *The Japanese Journal of Genetics*, *68*, 567–595. https://doi.org/10.1266/jjg.68.567

Virkkala, R., & Lehikoinen, A. (2017). Birds on the move in the face of climate change: High species turnover in northern Europe. *Ecology and Evolution*, *7*, 8201–8209. https://doi.org/10.1002/ece3.3328

Wang, W. J., McKay, B. D., Dai, C. Y., Zhao, N., Zhang, R. Y., Qu, Y. H., Song, G., Li, S. H., Liang, W., Yang, X. J., Pasquet, E., & Lei, F. M. (2013). Glacial expansion and diversification of an East Asian montane bird, the green-backed tit (*Parus monticolus*). *Journal of Biogeography*, *40*, 1156–1169. https://doi.org/10.1111/Jbi.12055

Zhang, R. Y., Song, G., Qu, Y. H., Alstrom, P., Ramos, R., Xing, X. Y., Ericson, P. G. P., Fjeldsa, J., Wang, H. T., Yang, X. J., Kristin, A., Shestopalov, A. M., Choe, J. C., & Lei, F. M. (2012). Comparative phylogeography of two widespread magpies: Importance of habitat preference and breeding behavior on genetic structure in China. *Molecular Phylogenetics and Evolution*, *65*, 562–572. https://doi.org/10.1016/j.ympev.2012.07.011

SUPPORTING INFORMATION

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

How to cite this article: Fan, P., Fjeldså, J., Liu, X., Dong, Y., Chang, Y., Qu, Y., Song, G., & Lei, F. (2021). An approach for estimating haplotype diversity from sequences with unequal lengths. *Methods in Ecology and Evolution*, 12, 1658–1667. https://doi.org/10.1111/2041-210X.13643