WILEY

Author Query Form

Journal: MEE3

Article: 13338

Dear Author,

During the copyediting of your manuscript, the following queries arose.

Please refer to the query reference callout numbers in the page proofs and respond to each by marking the necessary comments using the PDF annotation tools.

Please remember illegible or unclear comments and corrections may delay publication.

Many thanks for your assistance.

AUTHOR: Please note that missing content in references have been updated where we have been able to match the missing elements without ambiguity against a standard citation database, to meet the reference style requirements of the journal. It is your responsibility to check and ensure that all listed references are complete and accurate.

Query reference	Query	Remarks
1	**AUTHOR: Please confirm that given names (blue) and surnames/family names (vermilion) have been identified correctly.**	
2	**AUTHOR: Please verify that the linked ORCID identifiers are correct for each author.**	
3	**AUTHOR: Please take this opportunity to review any supporting information files that you submitted via ScholarOne. If you find any errors, please upload the corrected file and add a note to the PDF to confirm which file should be replaced. Also, we ask that you check and confirm the file names and descriptions for all supporting information. This will be your last opportunity to make any changes to the Supporting Information files, these files cannot be amended after publication.**	
4	**AUTHOR: Please check all website addresses and the functionality of the underlying links and confirm that they are correct. (Please note that it is the responsibility of the author(s) to ensure that all URLs given in this article are correct and useable.)**	
5	**AUTHOR: Your figure has been relabeled for clarity. Please check carefully that relabeling has been done correctly.**	
6	**AUTHOR: Please mention part labels 'a-c' in Figure 2 caption.**	
7	**AUTHOR: Yang et al. (2017) has not been included in the Reference List, please supply full publication details.**	

8	**AUTHOR: Please provide reference citation for GitHub and Zenodo.**	
9	**AUTHOR: Please provide volume and page range for Huang et al. (2019b).**	

Funding Info Query Form

Please confirm that the funding sponsor list below was correctly extracted from your article: that it includes all funders and that the text has been matched to the correct FundRef Registry organization names. If a name was not found in the FundRef registry, it may not be the canonical name form, it may be a program name rather than an organization name, or it may be an organization not yet included in FundRef Registry. If you know of another name form or a parent organization name for a "not found" item on this list below, please share that information.

FundRef name	**FundRef Organization Name**
Young Elite Scientists Sponsorship Program by CAST	
Strategic Priority Research Program of the Chinese Academy of Sciences	
National Key Programme of Research and Development, Ministry of Science and Technology	
Shaanxi Science and Technology Innovation Team	
National Natural Science Foundation of China	National Natural Science Foundation of China
Shaanxi Province Talents 1000 Fellowship	

Received: 14 May 2019 | Accepted: 11 November 2019

DOI: 10.1111/2041-210X.13338

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

POLYGENE: Population genetics analyses for autopolyploids based on allelic phenotypes

Kang Huang[1,2] | Derek W. Dunn[1] | Kermit Ritland[2] | Baoguo Li[1,3]

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

[2]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada

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

Correspondence
Baoguo Li
Email: baoguoli@nwu.edu.cn

Funding information
Young Elite Scientists Sponsorship Program by CAST, Grant/Award Number: 2017QNRC001; Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/Award Number: XDB31020302; National Key Programme of Research and Development, Ministry of Science and Technology, Grant/Award Number: 2016YFC0503200; Shaanxi Science and Technology Innovation Team, Grant/Award Number: 2019TD-012; National Natural Science Foundation of China, Grant/Award Number: 31572278, 31730104 and 31770411; Shaanxi Province Talents 1000 Fellowship

Handling Editor: Oscar Gaggiotti

Abstract

1. Polyploidy has appeared in almost every ancestral plant lineage, and in extant species, occurs frequently. When present, polyploidy presents problems for genetic data analysis, which are caused by both genotypic ambiguities and double-reduction.
2. To address these problems, we developed a new software package, POLYGENE, which enables the estimation of genotypic frequencies for a number of polysomic inheritance models. Specifically, POLYGENE obtains posterior probabilities for genotypes hidden within allelic phenotypes.
3. Comprehensive modes of genetic analyses are provided by POLYGENE, which include genetic diversity analysis, tests for allelic phenotypic or genotypic distributions, linkage disequilibrium and genetic differentiation, genetic distance analysis, principal coordinates analysis, hierarchical clustering analysis, individual inbreeding coefficient estimation, individual heterozygosity index estimation, population assignment, pairwise relatedness estimation, parentage analysis, *analysis of molecular variance* and Bayesian clustering.
4. POLYGENE enables easy and convenient allelic phenotype- or genotype-based analysis for both autopolyploids and diploids. POLYGENE will thus facilitate molecular ecology research involving autopolyploids.

KEYWORDS
allelic phenotype, AMOVA, double-reduction, parentage analysis, polysomic inheritance, population genetics

1 | INTRODUCTION

Polyploids are cells or organisms having a genome with more than two sets of homologous chromosomes. Polyploidy has occurred in almost every ancestral plant lineage, and frequently occurs in extant species, especially in plants (Barker, Arrigo, Baniaga, Li, & Levin, 2016). Due to their often important role in plant speciation, polyploids are regularly the subject of theoretical and experimental studies for evolutionary biology, molecular ecology and agriculture (Ling et al., 2018).

There are two distinct mechanisms of genome duplication that result in polyploidy: allopolyploidy and autopolyploidy. In allopolyploidy, chromosomes originate from two species; in autopolyploidy, all chromosomes originate within a single species, often due to

unreduced gametes. This paper mainly focuses on autopolyploids and allopolyploids that display polysomic inheritance.

In autopolyploids, both bivalents and multivalents can be formed during meiosis, resulting in disomic and polysomic inheritance, respectively. These are two extremes and many autopolyploid taxa represent intermediate stages (Butruille & Boiteux, 2000). Allopolyploids generally display disomic inheritance (Luo et al., 2006), because chromosomes from different species are not completely homologous. However, multivalent pairing can also occur in allopolyploids, resulting in a mixed inheritance pattern across loci in the genome, termed segmental allopolyploidy (Stebbins, 1950).

Because of differences in data formats and inheritance models, computer software designed for diploid organisms, such as GENEPOP (Rousset, 2008) and ARLEQUIN (Excoffier & Lischer, 2010), cannot be used for autopolyploid species. In general, polyploid population genetics analyses present two major challenges: (a) genotyping ambiguities and (b) double-reduction.

For PCR-based markers, the dosage of alleles cannot be directly determined by electrophoresis. For example, for autotetraploids, the genotype *AABB* has the same electrophoresis band pattern as the genotype *AAAB*. Therefore, the true genotype of individuals cannot be determined by electrophoresis and only the allelic phenotype is available for these markers. Novel technologies such as *genotyping-by-sequencing* (GBS) also uses PCR and suffers from such problems. Some researchers have claimed that genotypes can be assigned from the band intensity (Esselink, Nybom, & Vosman, 2004) or allelic read depth (Voorrips, Gort, & Vosman, 2011) using such methods. However, due to varying amplification efficiency among alleles in PCR, the results of these methods are still unreliable. Hereafter, we denote the allelic phenotype as a set of alleles and the genotype as a multiset of alleles, with the allelic phenotype being abbreviated as the 'phenotype'.

A peculiarity of polysomic inheritance is double-reduction, which occurs from a combination of three major events during meiosis: (a) the crossing-over between non-sister chromatids, (b) an appropriated pattern of disjunction, and (c) the subsequent migration of the chromosomal segments carrying a pair of sister chromatids to the same gamete (Darlington, 1929). For example, assume a tetraploid individual *ABCD* produces a gamete *AA*. Double reduction will result in gametes carrying *identical-by-descent* (IBD) alleles and increased homozygosity. A brief description of polysomic inheritance models can be found in the Supporting Information S1.

To solve these two problems, we developed a new software package: POLYGENE. We note that there are several other current software packages for polysomic inheritance, such as POLYSAT (Clark & Jasieniuk, 2011), SPAGEDI (Hardy & Vekemans, 2002), POLYRELATEDNESS (Huang, Ritland, Guo, Shattuck, & Li, 2014), GENODIVE (Meirmans & Tienderen, 2004) and STRUCTURE (Pritchard, Stephens, & Donnelly, 2000). POLYSAT converts the phenotype into a binary array and uses methods for dominant markers for analysis. Because the models of dominant markers and codominant markers are different, such a method is biased. STRUCTURE cannot handle ambiguous genotypes. The remaining three software packages assume that genotypic frequencies are in accordance with the *Hardy–Weinberg equilibrium* (HWE). Because genotypic frequencies under double reduction or inbreeding will always deviate from those expected under the HWE, this will override any heterozygous genotypes and bias the results. Moreover, some applications are sensitive to double reduction. For example, in a parentage analysis, the true father of the genotype *ABBB* will be excluded when the offspring genotype is *AACC*.

2 | THE NEW SOFTWARE PACKAGE 'POLYGENE'

POLYGENE v1.0b is written in C++ and C# with a graphical user interface. The source code is available from the GitHub website (http://github.com/huangkang1987/polygene). The binary executables for three platforms (Windows, Linux and Mac OS X) are provided. For the remaining Linux platforms, users can compile their own source code. Instructions on how to do this can be found in the user manual. To ensure free copying, distribution and modifications of the software and its source code, POLYGENE is distributed under the terms of a GNU General Public License, version 3.

POLYGENE supports a maximum ploidy level of 10, with different ploidy levels among populations also supported. The ploidy levels of all individuals within the same population are assumed to be equal, so that tests for allelic phenotypic distribution, linkage disequilibrium and differentiations can all be conducted. POLYGENE also supports multi-level region definition (Huang, Li, Dunn, Zhang, & Li, 2019), so as to analyse the variance component in different hierarchies in *analysis of molecular variance* (AMOVA). Other analyses are also supported for each hierarchy (e.g. genetic diversity, genetic distance, *principal coordinate analysis* (PCoA), hierarchical clustering). The hardware requirements to run POLYGENE are not high; a computer with a 4 GiB memory is sufficient.

3 | INPUT AND OUTPUT FORMAT

An example for the format of inputting the allelic phenotypic and genotypic data is shown in Table 1. Here, the first row is the header with each subsequent row representing an individual. The first three columns are: (a) the individual, (b) the population and

TABLE 1 An example of inputting allelic phenotypic/genotypic data

ID	Pop	Ploidy	Loc1	Loc2
Ind1	pop1	4	2,3,4	1,2,4
Ind2	pop1	4	4	2,3,4
Ind3	pop2	4	2,3,4,4	2,3,4,3
Ind4	pop2	4	1,3,4,1	3,4,3,3

(c) the ploidy level. Each cell of the fourth column onwards represents either a phenotype (the first two individuals) or a genotype (the latter two individuals) at a specific locus, with these columns being separated by tabs to enable data to be pasted from a spreadsheet. The alleles within a phenotype are identified by positive integers, which are separated by commas. For missing values, the corresponding cells are left blank. POLYGENE supports multi-level region definition, where each region is defined as a collection of populations or regions (e.g. Level I region is a collection of populations, Level II region is a collection of level I regions). The region can be defined in the parameters of the *analysis of molecular variance* (AMOVA, Figure 2).

POLYGENE is able to handle both phenotypic and genotypic data. The function 'remove duplicated alleles' is provided so as to convert genotypes into phenotypes. If this function is not used, the candidate genotypes will be generated based on the dosage of the inputted alleles. For example, an incomplete tetraploid genotype *ABB* will generate two candidate genotypes: *ABBA* and *ABBB*. If null alleles are also considered, the genotype *ABBY* will also be added to the list, where *Y* is the null allele. The missing data are usually not analyzed, except for the dummy haplotype-based methods or when null alleles are considered.

Performing each analysis in POLYGENE is relatively straightforward. First, the phenotypic data are formatted and then entered into the box marked 'Phenotypes/Genotypes.' Second, the required analytical method(s) is selected and the corresponding parameters on the page marked 'Parameters' is configured (Figure 1). Finally, the button marked 'Calc' in the toolbox menu at the top of the window is 'clicked'. All analyses will then be performed simultaneously, and the progress bar will show the progress of the current analysis. After a short delay, the results will be displayed in plain text on the corresponding pages (Figure 2).

4 | MATERIALS AND METHODS

POLYGENE improves on existing packages and also includes four polysomic inheritance models, that is, *random chromosome segregation* (RCS) (Muller, 1914), *pure random chromatid segregation* (PRCS) (Haldane, 1930), *complete equational segregation* (Mather, 1935) and *partial equational segregation* (Huang, Wang, et al., 2019). A brief description of these models can be found in the Supporting Information S1 (section double reduction models). The disomic inheritance model can also be used, whose genotypic frequency is equivalent to the RCS model.

These inheritance models determine the a priori probabilities of the genotypic frequencies, based on the genotypic and phenotypic frequencies under each model (Huang, Wang, et al., 2019). We are able to calculate the posterior probabilities of the genotypes hidden behind the phenotypes and these genotypes are weighted by their posterior probabilities. The correct model can accurately obtain the posterior probabilities of hidden genotypes and improve the accuracy of the analyses. The optimal model can be evaluated by the *Bayesian information criterion* (BIC), and the BIC of an

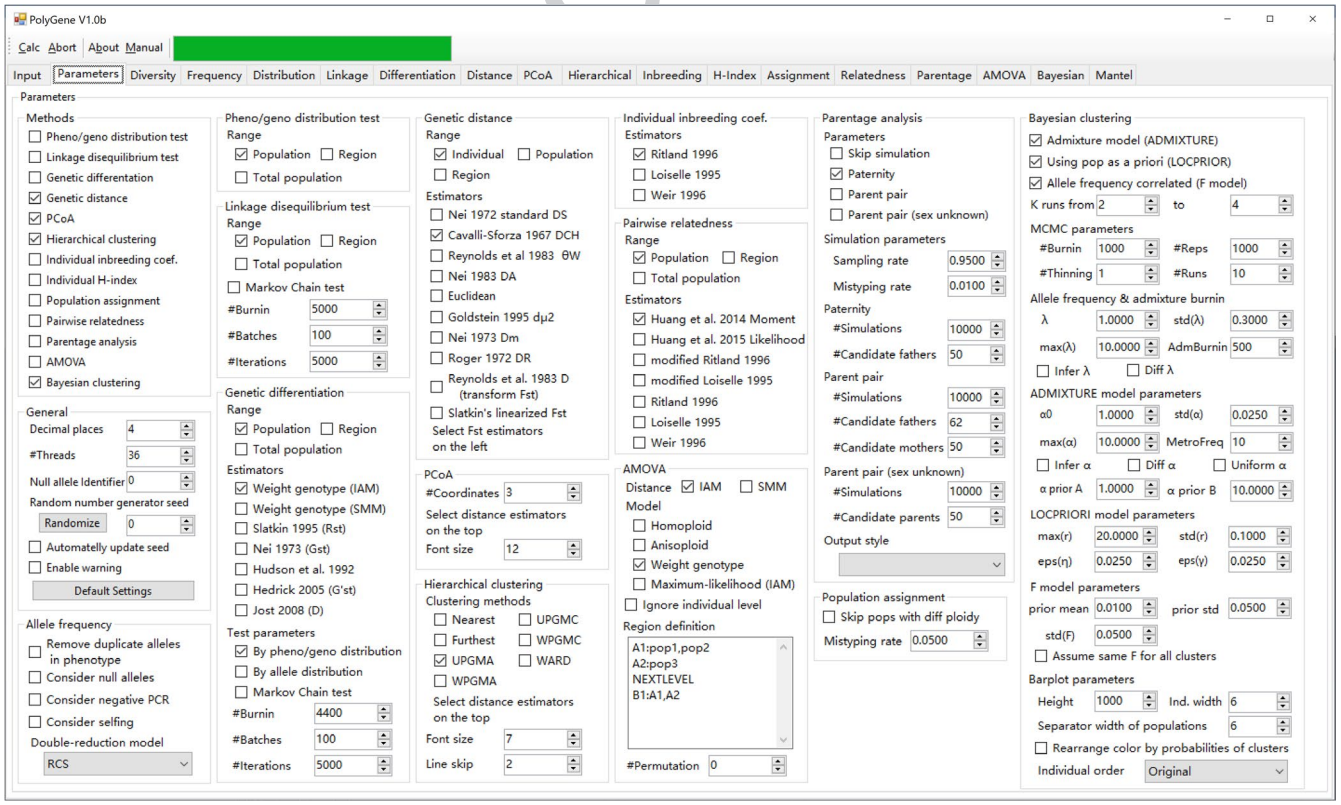

FIGURE 1 The graphical user interface of POLYGENE

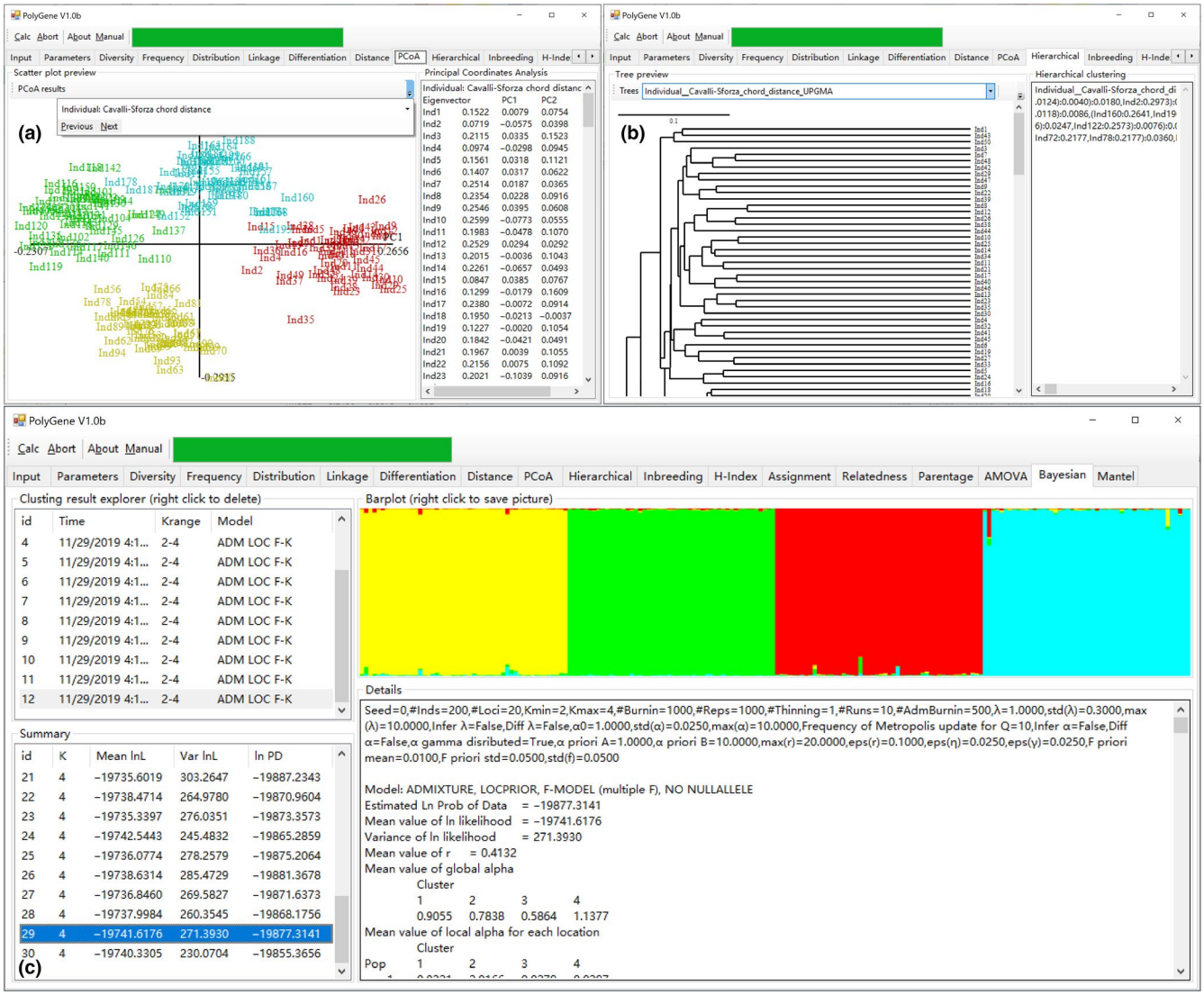

FIGURE 2 An example of the results output of PCoA, hierarchical clustering and Bayesian clustering. Simulated dataset D using the CES model is presented as an example. PCoA and hierarchical clustering are based on individual Cavalli-Sforza chord distance. The default settings are applied for Bayesian clustering

inheritance model is sum of all of the BIC values over all loci for all populations. The BIC is provided in the results of a phenotypic distribution test.

Multiple genetic analysis methods are incorporated into POLYGENE, and these can be classified into four categories: (a) weighted genotype, (b) allele frequency, (c) phenotype and (d) dummy allele/haplotype. The effects of double reduction, null alleles, negative amplification and self-fertilization can be freely taken into account in most methods provided by our new software. The analytical methods and their associated parameters are presented in Figure 1. An extensive description of the methods used in POLYGENE can be found in the Supporting Information S1. Because the genotypes are weighted at each locus independently, this assumes linkage equilibrium and some prior distribution of genotypes in all methods except for those that are phenotype-based. We briefly describe these methods below.

Weighted genotype-based methods are based on the weighting of extracted genotypes according to their posterior probabilities. These methods are allele frequency estimation based on an EM algorithm (De Silva, Hall, Rikkerink, McNeilage, & Fraser, 2005; Kalinowski & Taper, 2006), genetic diversity analysis, parentage analysis (Huang, Mi, Dunn, Wang, & Li, 2018; Kalinowski, Taper, & Marshall, 2007), kinship coefficient estimation (Loiselle, Sork, Nason, & Graham, 1995; Ritland, 1996; Weir, 1996), and pairwise relatedness coefficient estimation (Huang, Guo, et al., 2015; Huang, Ritland, et al., 2015; Huang et al., 2014). The effects of null alleles, negative amplification and self-fertilization can be freely taken into account in most methods, which use the genotypic and phenotypic frequencies. The likelihood equations in parentage analysis can incorporate self-fertilization.

Allele frequency-based methods are based on the estimated allele frequencies, which do not need specific modification for

polysomic inheritance. These are F_{ST} and its analogues (Hedrick, 2005; Hudson, Slatkin, & Maddison, 1992; Jost, 2008; Nei, 1973; Slatkin, 1995), genetic distance (Cavalli-Sforza & Edwards, 1967; Goldstein, Ruiz, Cavallisforza, & Feldman, 1995; Nei, 1972; Nei & Roychoudhury, 1974; Nei, Tajima, & Tateno, 1983; Reynolds, Weir, & Cockerham, 1983; Rogers, 1972; Slatkin, 1995), principal coordinate analysis, hierarchical clustering analysis and Mantel tests (Mantel, 1967; Smouse & Sokal, 1986). Genetic distance indices can be extended to that between two individuals. For instance, the frequencies of alleles within an individual can be obtained from the phenotypes by using the posterior probabilities of the candidate genotypes. The latter three methods are based on the estimated genetic distance matrix.

The phenotype-based methods are based on the observed or expected distribution of phenotypes. These are tests for linkage disequilibrium, genetic differentiation, phenotypic distribution and population assignment. The first two tests are respectively performed by Fisher's *G* test and the Markov Chain test (Raymond & Rousset, 1995) with the null hypotheses that two loci are under linkage equilibrium or the phenotypes in multiple populations are drawn from the same distribution. The phenotypic distribution test is performed by Fisher's *G* test, the null hypothesis of which is that the observed phenotypes accord with double-reduction. Population assignment assigns each individual to the population with the highest likelihood (Paetkau, Slade, Burden, & Estoup, 2004) to find the natal population of each individual.

Dummy allele or haplotype-based methods are based on the weighting of allele copies in the phenotypes. This technique is used in Weir and Cockerham's (1984) F_{ST} estimator, AMOVA (Cockerham, 1973; Excoffier, Smouse, & Quattro, 1992) and Bayesian clustering (Pritchard et al., 2000). AMOVA is generalized to any number of hierarchies, and the genetic distance between dummy alleles or haplotypes are calculated during variance decomposition. Bayesian clustering implements three models: the ADMIXTURE model

TABLE 2 RMSE values of estimated effective number of alleles, heterozygosity and *F*-statistics for our six simulated datasets

	Dataset	A_e	H_I	H_S	H_T	F_{IS}	F_{ST}
SPAGEDI	A	0.2784	0.3467	0.0227	0.0170	0.4485	0.0041
	B	0.2504	0.3467	0.0205	**0.0010**	0.4486	0.0134
	C	0.2671	0.4211	**0.0219**	0.0146	0.5598	0.0036
	D	0.2364	0.4211	0.0196	**0.0012**	0.5583	0.0139
	E	0.3029	0.3000	0.0245	0.0224	0.3783	**0.0040**
	F	0.2727	0.3000	0.0223	**0.0009**	0.3783	0.0159
	G	0.2714	0.3677	0.0222	0.0195	0.4791	**0.0041**
	H	0.2389	0.3677	0.0197	**0.0009**	0.4795	0.0129
GENODIVE	A	**0.2123**	0.3151	**0.0207**	0.0094	0.4433	**0.0028**
	B	0.1950	0.3144	**0.0164**	0.0014	0.4434	**0.0126**
	C	0.2410	0.3858	0.0231	0.0099	0.5515	**0.0033**
	D	0.2098	0.3862	0.0182	0.0013	0.5497	0.0138
	E	0.2139	0.2391	**0.0194**	0.0150	0.3270	0.0135
	F	0.1877	0.2364	**0.0155**	0.0010	0.3258	**0.0128**
	G	0.2354	0.2618	**0.0206**	0.0143	0.3659	0.0127
	H	0.1917	0.2580	**0.0156**	0.0010	0.3613	**0.0124**
POLYGENE[a]	A	0.2181	**0.0373**	0.0215	**0.0086**	**0.0414**	0.0066
	B	**0.1769**	**0.0370**	0.0168	0.0019	**0.0454**	0.0129
	C	**0.2387**	**0.0378**	0.0228	**0.0093**	**0.0478**	0.0079
	D	**0.1879**	**0.0386**	**0.0177**	0.0021	**0.0517**	**0.0136**
	E	**0.2120**	**0.0331**	0.0202	**0.0085**	**0.0351**	0.0060
	F	**0.1746**	**0.0325**	0.0166	0.0016	**0.0381**	0.0131
	G	**0.2150**	**0.0378**	0.0210	**0.0083**	**0.0455**	0.0068
	H	**0.1879**	**0.0638**	0.0177	0.0021	**0.0853**	0.0136

Note: Header row: A_e: effective number of alleles; H_I: observed heterozygosity; H_S: expected heterozygosity in a subpopulation; H_T: expected heterozygosity in the total population; H_I, H_S and F_{IS} is calculated for each subpopulation at each locus; while H_T and F_{ST} is calculated in the total population at each locus. The optimal RMSEs are marked in bold.

[a]The optimal model evaluated from BIC are: A: PRCS (ΔBIC = 10.36), B: PRCS (ΔBIC = 8.44), C: CES (ΔBIC = 81.25), D: CES (ΔBIC = 88.80), E: RCS (ΔBIC = 28.33), F: RCS (ΔBIC = 35.64), G: CES (ΔBIC = 12.94), H: CES (ΔBIC = 13.89).

(Pritchard et al., 2000), the LOCPRIORI model (Hubisz, Falush, Stephens, & Pritchard, 2009) and the F model (Falush, Stephens, & Pritchard, 2003) and calculates the probability that each dummy allele is drawn from each cluster.

5 | SIMULATIONS AND COMPARISONS

Here, we use both simulated and empirical datasets to evaluate the accuracy of estimated genetic diversity indices (including the effective number of alleles A_e, the observed heterozygosity H_I, the expected heterozygosity in a subpopulation H_S and in the total population H_T) and *F*-statistics (F_{IS} and F_{ST}) of the software packages GENODIVE V2.0b23 (Meirmans & Tienderen, 2004), SPAGEDI V1.5d (Hardy & Vekemans, 2002) and POLYGENE V1.0b.

Eight simulated datasets (denoted by Dataset A-H) were generated using the simulation function of POLYGENE. Specifically, Datasets A and E simulate no inbreeding and no differentiation; Datasets B and F simulate differentiation; Datasets C and G simulate inbreeding; and Datasets D and H simulate both inbreeding and differentiation. All datasets were generated with a random number generator seed equal to zero.

There were four populations in each dataset and each had 50 autotetraploids. Twenty tetra-allelic loci were simulated, whose allele frequencies were drawn from a triangular distribution (0.1, 0.2, 0.3 and 0.4), the allele frequencies being equal among populations in Datasets A, C, E and G. Frequencies in Datasets B, D, F and H were shifted accordingly to simulate differentiation. For Datasets C, D, G and H, we used a selfing rate of 0.3 to simulate inbreeding. The genotypes at 20 tetra-allelic loci for these individuals were generated under the PRCS model (Datasets A-D) or the RCS/disomic model (Dataset E-G) and then converted into phenotypes. The true values of the compared statistics were derived using the methods of Huang, Dunn, et al. (2019), Huang, Wang, et al. (2019), and are shown in Table S32.

For POLYGENE, we used the BIC in order to select the best inheritance model. Selfing was considered for Datasets C and D. We estimated statistics other than F_{ST} using the '`Genetic Diversity`' function, with the locus specific F_{ST} estimated using Nei's (1973) estimator. In SPAGEDI, the option '`2.3 global Gst and pairwise Gst`' was used with all additional options disabled except '`Report allele freq and diversity coef`'. In GENODIVE, all statistics were estimated using the '`Correct for unknown dosage of alleles`' option, in which A_e is estimated by the '`Genetic Diversity`' function for each population, F_{IS} is estimated

TABLE 3 General information for the four empirical datasets used for comparisons

Dataset	Species	#Individuals	#Populations	#Loci
Zwart et al. (2016)	Grey willow *Salix cinereal*	139	1	7
Yang et al. (2018)	*Lotus sessilifolius*	48	8	11
Hoeltgebaum, Londoño, Lando, and Reis (2017)	*Varronia curassavica*	681	4	8
Julier et al. (2010)	Lucerne *Medicago sativa*	463	16	7

TABLE 4 RMSE values of estimated effective number of alleles, heterozygosity and *F*-statistics for the four empirical datasets

	Dataset	A_e	H_I	H_S	H_T	F_{IS}	F_{ST}
SPAGEDI	Zwart et al. (2016)	2.6848	0.5608	0.3625		0.8837	
	Yang et al. (2017)	0.7930	0.4156	0.0443	0.0112	0.7857	0.0468
	Hoeltgebaum et al. (2017)	0.6031	0.5274	0.0517	0.0427	0.9028	0.0145
	Julier et al. (2010)	0.8240	0.3499	0.0558	0.0469	0.3794	0.0112
GENODIVE	Zwart et al. (2016)	2.4647	0.3447	0.3572		0.1900	
	Yang et al. (2017)	**0.1167**	0.1550	**0.0328**	**0.0083**	0.2048	0.0582
	Hoeltgebaum et al. (2017)	0.3499	0.2014	**0.0181**	**0.0122**	0.3040	0.0079
	Julier et al. (2010)	0.3754	0.1919	0.0287	0.0520	0.2342	0.0104
POLYGENE[a]	Zwart et al. (2016)	**0.1703**	**0.0073**	**0.0130**		**0.0221**	
	Yang et al. (2017)	0.3743	**0.0290**	0.0487	0.0221	**0.0774**	**0.0303**
	Hoeltgebaum et al. (2017)	**0.3247**	**0.0164**	0.0220	0.0180	**0.0424**	**0.0057**
	Julier et al. (2010)	**0.2770**	**0.0121**	**0.0196**	**0.0164**	**0.0216**	**0.0024**

Note: The optimal RMSEs are marked in bold.

[a]The optimal model evaluated from BIC are: Zwart et al. (2016): PES (r_s = 0.5) (ΔBIC = 1.95); Yang et al. (2017): PES (r_s = 0.25) (ΔBIC = 3.41); Hoeltgebaum et al.'s (2017): CES (ΔBIC = 1.40); Julier et al. (2010): CES (ΔBIC = 2.24).

by the 'Hardy-Weinberg' function, and H_S, H_T and F_{ST} are all estimated using the 'G-statistics' function.

The results for all simulations are presented in Table 2. We found that POLYGENE is the most accurate package for estimating A_e, H_I, H_T and F_{IS}. Especially for H_I and F_{IS}, the RMSE using POLYGENE is greatly improved. Previously available software packages thus appear to be unable to accurately estimate H_I and F_{IS}, because the estimates for H_I are close to one and those for F_{IS} are negative. However, POLYGENE is marginally less accurate than GENODIVE in estimating both H_S and F_{ST}. For H_T, SPAGEDI is most accurate for high differentiation with POLYGENE being most accurate for low differentiation.

In natural populations, the true values of the compared statistics are usually unknown. To evaluate accuracy, we adopted the statistics calculated from the genotypic datasets for each of several software packages as the true value, and those calculated from the phenotypic datasets as the observed values. We used four empirical genotypic datasets of autotetraploid species (details presented in Table 3).

The software configurations used were the same as for the simulated datasets and the results for the empirical datasets are presented in Table 4. POLYGENE performs best with the RMSE of all statistics being relatively low. In particular, the RMSE of all statistics are optimal for the datasets of Zwart, Elliott, Hopley, Lovell, and Young (2016) and Julier, Semiani, and Laouar (2010). Similar to the results obtained from the simulated datasets, both SPAGEDI and GENODIVE produce inaccurate estimates for F_{IS} and H_I, with each of these two software packages producing values that deviate from the values calculated from the genotypic data. For POLYGENE, the results from both the genotypic and phenotypic datasets are similar, with the RMSE reduced by ~90%.

6 | CONCLUSION

We show here that POLYGENE is a useful new tool for molecular ecology studies involving autopolyploids. This new software package utilizes genotypic frequencies under double reduction to perform subsequent analyses, and extends the genetic analyses methods that can be used for polysomic inheritance. POLYGENE also supports data from haploids, diploids and varying ploidy levels among populations, and can be run on Windows, Linux and Mac OS X. POLYGENE enables the analysis of autopolyploids based on allelic phenotypes or genotypes to be as easy and convenient as when using data from diploids. This new software will thus likely facilitate molecular ecology research involving autopolyploid species.

ACKNOWLEDGEMENTS

We thank Prof. Cronk Quentin (University of British Columbia, Canada), Dr. Marcia Patricia Hoeltgebaum (Federal University of Santa Catarina, Brazil) and Dr. Bernadette Julier (National Institute of Agricultural Research, France) for providing the empirical dataset. We also thank the associate editor, Grieves, Chris, and the four anonymous reviewers for their constructive comments that helped us improve the manuscript. This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020302), the Natural Science Foundation of China (31730104, 31770411 and 31572278), the Young Elite Scientists Sponsorship Program by CAST (2017QNRC001), the National Key Programme of Research and Development, Ministry of Science and Technology (2016YFC0503200), and the Shaanxi Science and Technology Innovation Team (2019TD-012). DWD is supported by a Shaanxi Province Talents 1000 Fellowship.

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

K.H. and B.L. conceived the ideas, K.R. designed methodology, D.W.D. checked the model, K.H. and D.W.D. wrote the draft and D.W.D. and K.R. edited the manuscript.

DATA AVAILABILITY STATEMENT

The binary executable, user manual and source code of POLYGENE, and the simulated and empirical results are available at GitHub (https://github.com/huangkang1987/polygene) and at Zenodo (https://doi.org/10.5281/zenodo.3541408).

ORCID

Kang Huang https://orcid.org/0000-0002-8357-117X
Derek W. Dunn https://orcid.org/0000-0001-5909-1224

REFERENCES

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., & Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytologist*, *210*, 391–398. https://doi.org/10.1111/nph.13698

Butruille, D. V., & Boiteux, L. S. (2000). Selection–mutation balance in polysomic tetraploids: Impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. *Proceedings of the National Academy of Sciences*, *97*, 6608–6613. https://doi.org/10.1073/pnas.100101097

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis: Models and estimation procedures. *Evolution*, *21*, 550–570. https://doi.org/10.1111/j.1558-5646.1967.tb03411.x

Clark, L. V., & Jasieniuk, M. (2011). POLYSAT: An R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, *11*, 562–566. https://doi.org/10.1111/j.1755-0998.2011.02985.x

Cockerham, C. C. (1973). Analyses of gene frequencies. *Genetics*, *74*, 679–700.

Darlington, C. D. (1929). Chromosome behaviour and structural hybridity in the Tradescantiae. *Journal of Genetics*, *21*, 207–286. https://doi.org/10.1007/BF02984208

De Silva, H., Hall, A., Rikkerink, E., McNeilage, M., & Fraser, L. (2005). Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity*, *95*, 327–334. https://doi.org/10.1038/sj.hdy.6800728

Esselink, G., Nybom, H., & Vosman, B. (2004). Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting-peak ratios) method. *Theoretical and Applied Genetics*, *109*, 402–408. https://doi.org/10.1007/s00122-004-1645-5

Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*, 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

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.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567–1587.

Goldstein, D. B., Ruiz, L. A., Cavallisforza, L. L., & Feldman, M. W. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 6723–6727. https://doi.org/10.1073/pnas.92.15.6723

Haldane, J. B. (1930). Theoretical genetics of autopolyploids. *Journal of Genetics*, *22*, 359–372. https://doi.org/10.1007/BF02984197

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. (2005). A standardized genetic differentiation measure. *Evolution*, *59*, 1633–1638. https://doi.org/10.1111/j.0014-3820.2005.tb01814.x

Hoeltgebaum, M. P., Londoño, D. M. M., Lando, A. P., & dos Reis, M. S. (2017). Reproductive strategy of the polyploid species *Varronia curassavica* Jacq. in restinga environment. *Journal of Heredity*, *108*, 424–430. https://doi.org/10.1093/jhered/esx024

Huang, K., Dunn, D. W., Li, Z. H., Zhang, P., Dai, Y., & Li, B. G. (2019). Inference of individual ploidy level using codominant markers. *Molecular Ecology Resources*, https://doi.org/10.1111/1755-0998.13032

Huang, K., Guo, S. T., Shattuck, M. R., Chen, S. T., Qi, X. G., Zhang, P., & Li, B. G. (2015). A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity*, *114*(2), 133–142. https://doi.org/10.1038/hdy.2014.88

Huang, K., Li, Y., Dunn, D. W., Zhang, P., & Li, B. (2019) A generalized framework of AMOVA with any number of hierarchies and any level of ploidies. *bioRxiv*, https://doi.org/10.1101/608117.

Huang, K., Mi, R., Dunn, D. W., Wang, T. C., & Li, B. G. (2018). Performing parentage analysis in the presence of inbreeding and null alleles. *Genetics*, *210*, 1467–1481. https://doi.org/10.1534/genetics.118.301592

Huang, K., Ritland, K., Guo, S. T., Dunn, D. W., Chen, D., Ren, Y., … Li, B. G. (2015). Estimating pairwise relatedness between individuals with different levels of ploidy. *Molecular Ecology Resources*, *15*, 772–784. https://doi.org/10.1111/1755-0998.12351

Huang, K., Ritland, K., Guo, S. T., Shattuck, M., & Li, B. G. (2014). A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources*, *14*, 734–744. https://doi.org/10.1111/1755-0998.12217

Huang, K., Wang, T. C., Dunn, D. W., Zhang, P., Liu, R. C., Cao, X. X., & Li, B. G. (2019). Genotypic frequencies at equilibrium for polysomic inheritance under double-reduction. *G3: Genes*, *Genomes*, *Genetics*, *9*, 1693–1706. https://doi.org/10.1534/g3.119.400132

Hubisz, M. J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, *9*, 1322–1332. https://doi.org/10.1111/j.1755-0998.2009.02591.x

Hudson, R. R., Slatkin, M., & Maddison, W. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*, 583–589.

Jost, L. (2008). G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*, *17*, 4015–4026.

Julier, B., Semiani, Y., & Laouar, M. (2010). Genetic diversity in a collection of lucerne populations from the Mediterranean basin evaluated by SSR markers. In C. Huyghe (Ed.), *Sustainable use of genetic diversity in forage and turf breeding* (pp. 107–112). Dordrecht, The Netherlands: Springer.

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

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, *16*, 1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Ling, H.-Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., … Liang, C. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature*, *557*, 424. https://doi.org/10.1038/s41586-018-0108-0

Loiselle, B. A., Sork, V. L., Nason, J., & Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Cardiology*, *82*, 1420–1425.

Luo, Z. W., Zhang, Z. E., Leach, L., Zhang, R. M., Bradshaw, J. E., & Kearsey, M. J. (2006). Constructing genetic linkage maps under a tetrasomic model. *Genetics*, *172*, 2635–2645. https://doi.org/10.1534/genetics.105.052449

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*, 209–220.

Mather, K. (1935). Reductional and equational separation of the chromosomes in bivalents and multivalents. *Journal of Genetics*, *30*, 53–78. https://doi.org/10.1007/BF02982205

Meirmans, P. G., & Tienderen, P. H. V. (2004). GENOTYPE and GENODIVE : Two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, *4*, 792–794. https://doi.org/10.1111/j.1471-8286.2004.00770.x

Muller, H. J. (1914). A new mode of segregation in Gregory's tetraploid primulas. *The American Naturalist*, *48*, 508–512. https://doi.org/10.1086/279426

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

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, *70*, 3321–3323. https://doi.org/10.1073/pnas.70.12.3321

Nei, M., & Roychoudhury, A. K. (1974). Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *American Journal of Human Genetics*, *26*, 421–443.

Nei, M., Tajima, F., & Tateno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular data II. Gene frequency data. *Journal of Molecular Evolution*, *19*, 153–170. https://doi.org/10.1007/BF02300753

Paetkau, D., Slade, R., Burden, M., & Estoup, A. (2004). Genetic assignment methods for the direct, real-time estimation of migration rate: A simulation-based exploration of accuracy and power. *Molecular Ecology*, *13*, 55–65. https://doi.org/10.1046/j.1365-294X.2004.02008.x

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

Raymond, M., & Rousset, F. (1995). An exact test for population differentiation. *Evolution*, *49*, 1280–1283. https://doi.org/10.1111/j.1558-5646.1995.tb04456.x

Reynolds, J., Weir, B. S., & Cockerham, C. C. (1983). Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*, *105*, 767–779.

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

Rogers, J. S. (1972). Measures of similarity and genetic distance. In M. R. Wheeler (Ed.), *Studies in Genetics VII* (pp. 145–153). Austin, TX: University of Texas Publication.

Rousset, 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

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, *139*, 457–462.

Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology*, *35*, 627–632. https://doi.org/10.2307/2413122

Stebbins, G. L. (1950). *Variation and evolution in plants*. New York, NY: Columbia University Press.

Voorrips, R. E., Gort, G., & Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*, *12*, 172. https://doi.org/10.1186/1471-2105-12-172

Weir, B. S. (1996). *Genetic data analysis II: Methods for discrete population genetic data*. Sunderland, MA: Sinauer Associates.

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

Yang, J. Y., Ojeda, D. I., Santos-Guerra, A., Molina, R. J., Caujapé-Castells, J., & Cronk, Q. (2018). Population differentiation in relation to conservation: Nuclear microsatellite variation in the Canary Island endemic *Lotus sessilifolius* (Fabaceae). *Conservation Genetics Resources*, *10*, 219–227. https://doi.org/10.1007/s12686-017-0778-1

Zwart, A. B., Elliott, C., Hopley, T., Lovell, D., & Young, A. (2016). Polypatex: An R package for paternity exclusion in autopolyploids. *Molecular Ecology Resources*, *16*, 694–700.

SUPPORTING INFORMATION

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

How to cite this article: Huang K, Dunn DW, Ritland K, Li B. POLYGENE: Population genetics analyses for autopolyploids based on allelic phenotypes. *Methods Ecol Evol*. 2019;00:1–9. https://doi.org/10.1111/2041-210X.13338