Title

VCFPOP: performing population genetics analyses for polyploids and anisoploids based on next-generation sequencing variant calling dataset

Authors

Kang Huang[1,2], Bing Yang[1], Jincuo Ao[1], Yuhang Li[1], Yunxia Cui[1], Yuchen Kong[1], Yifan Wu[1], Derek W. Dunn[1], Baoguo Li[1]

Addresses

[1] Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences, Northwest University, Xi'an 710069, China

[2] Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC V6T1Z4, Canada

Keywords

Polysomic inheritance, next-generation sequencing data, variant calling format, population genetics.

Corresponding author

Baoguo Li

Telephone: +8613572209390; Fax: +86 029 88303304; E-mail: baoguoli@nwu.edu.cn

Running title

Population genetics analysis for autopolyploids

Abstract

Polyploids are cells or organisms with a genome consisting of more than two sets of homologous chromosomes. Polyploid plants have important traits that facilitate speciation and are thus often model systems for evolutionary, molecular ecology and agricultural studies. However, due to their unusual mode of inheritance and double-reduction, diploid models of population genetic analysis cannot properly be applied to polyploids. To overcome this problem, we developed a software package entitled VCFPOP to perform a variety of population genetic analyses for autopolyploids, such as parentage analysis, analysis of molecular variance, principal coordinates analysis, hierarchical clustering analysis and Bayesian clustering. We make this software freely available, downloadable from http://github.com/huangkang1987/vcfpop.

Keywords: polysomic inheritance, next-generation sequencing data, population genetics, AMOVA, Bayesian clustering, *F*-statistics.

Introduction

Polyploids are cells or organisms with a genome consisting of more than two sets of homologous chromosomes. Polyploids represent a significant portion of all plant species, with from 30-80% of angiosperms showing polyploidy (Burow et al. 2001). Because of their propensity to facilitate speciation, polyploid species have often been used as model systems for evolutionary, molecular ecology and agricultural studies. More recently, polyploids have increasingly become the focus of theoretical and experimental work (Avni et al. 2017; Ling et al. 2018).

There are two distinct mechanisms of genome duplication that result in polyploidy: allopolyploidy and autopolyploidy. This paper focuses on autopolyploids that display polysomic inheritance. In autopolyploids, more than two homologous chromosomes can pair at meiosis, resulting in the formation of multivalents and polysomic inheritance. A typical feature of polysomic inheritance is the possibility that a gamete inherits a single gene copy twice, termed double-reduction (Butruille & Boiteux 2000). For example, an autotetraploid individual *ABCD* produces a gamete *AA*. The double-reduction will change the frequency of genotypes within a population (Huang et al. 2019), resulting in increased homozygosity and an inflated inbreeding coefficient (Hardy 2016).

Due to differences in data format and modes of inheritance between diploids and polyploids, population genetics software designed for diploid organisms such as GENEPOP (Rousset 2008) and ARLEQUIN (Excoffier & Lischer 2010) cannot be used for autopolyploids. Some software packages have been developed in order to accommodate polyploid

genotype datasets, e.g. POLYSAT (Clark & Jasieniuk 2011), SPAGEDI (Hardy & Vekemans 2002), POLYRELATEDNESS (Huang *et al.* 2014), GENODIVE (Meirmans & Tienderen 2004) and STRUCTURE (Pritchard *et al.* 2000), etc. However, such software includes the assumption of a disomic mode of inheritance and genotypic frequencies in accordance with the *Hardy-Weinberg Equilibrium* (HWE), whereas alleles of the same genotype are assumed to be independent.

Huang *et al.* (2019) derived the genotypic frequency for various double-reduction models: the *random chromosome segregation* (RCS) (Muller 1914), the *pure random chromatid segregation* (PRCS) (Haldane 1930), the *complete equational segregation* (CES) (Mather 1935) and the partial equational segregation (PES) (Huang *et al.* 2019). The software package POLYGENE is able to use all of these double-reduction models and can perform population genetics analyses for both allelic phenotypic and genotypic data (Huang *et al.* 2020). However, POLYGENE cannot accommodate the large datasets that result from the use of next-generation sequencing (NGS) methods. To solve this problem, we developed a new software package entitled VCFPOP.

The new software package VCFPOP

VCFPOP is a free software developed with C++. It works only via command-line mode and will run on Windows, Linux and Mac OS X. To ensure free copying, distribution and modifications of the software and its source code, VCFPOP is distributed under a GNU General Public License (GPL, version 3).

VCFPOP has been optimized for memory allocation and calculation speed to analyze

large genotype datasets. It can analyze the genotypes of haploids, diploids, polyploids and anisoploids. For polyploids, VCFPOP applies the RCS, PRCS, CES and PES double-reduction models and supports a maximum ploidy level of 10. VCFPOP also supports multi-level region definition, so as to analyze the variance component of different hierarchies via *analysis of molecular variance* (AMOVA) (Huang *et al.* 2021).

Input format

VCFPOP can handle a multiple genotype format, consisting of *variant call format* (VCF) V4.x (compressed or uncompressed) (Danecek *et al.* 2011), *binary call format* (BCF) V2.x, GENEPOP V4.3 (Rousset 2008), STRUCTURE V2.3 (Pritchard *et al.* 2000), CERVUS V3.0 (Kalinowski *et al.* 2007) , ARLEQUIN V3.6 (Excoffier & Lischer 2010), POLYGENE V1.4 (Huang et al. 2020), and POLYRELATEDNESS V1.7 (Huang *et al.* 2015a).

Multiple VCF/BCF files for the same samples sequenced at different variants (vertical concatenation, separated by '`|`' in `-g_input`), or different samples sequenced at the same variants (horizontal concatenation, separated by '`&`' in `-g_input`) can be analyzed together without additional concatenation.

Because both VCF and BCF formats do not contain population or regionally differentiation, populations and regions should be additionally defined using the arguments, in which '`-g_indfile`' reads the content from a file and '`-g_indtext`' reads from the command. Because the command-line mode does not allow linebreaks, '#n' or space is used as the escape character in `-g_indtext`. Multi-level region definition is supported in all corresponding analysis (e.g., AMOVA and population assignment) (Huang et al. AMOVA

paper). The example format of `-g_indfile` is shown as follows:

```
-g_indtext="pop1:ind1,ind2,ind3   pop2:#4,#5-#6   pop3:ind7,ind8,ind9
#REG A1:#1,#2 A2:pop3 #REG B1:#1-#2"
```

The identifier of a population is followed by a semi-colon, then the individual identifiers, ordinations or ordination ranges separated by commas. The separator '`#REG`' is used to separate regions or populations at different levels. Similarly, the identifier of a region is followed by a semi-colon, then the sub-region or population identifiers.

Usage

After installation, the user should open the terminal to launch VCFPOP.

To view the help for all functions, execute

```
./vcfpop -h
```

To view the detail help information of some specific functions, execute

```
./vcfpop -h -func1 -func2
```

To use specific functions, execute

```
./vcfpop –func1 -func1_parameters –func2 -func2_parameters ...
```

To use a parameter file in the same format as the program arguments but allow line breaks, execute

```
./vcfpop –p=parameter_file
```

After calculation, the results are saved as '`*.func.txt`'. The specific functions are as follows:

`-g`	General settings

`-f`	Filter for individual, locus or genotype
`-haplotype`	Haplotype extraction
`-convert`	File conversion
`-diversity`	Genetic diversity indices
`-indstat`	Individual statistics
`-fst`	Genetic differentiation
`-gdist`	Genetic distance
`-amova`	Analysis of molecular variance
`-popas`	Population assignment
`-relatedness`	Relatedness coefficient
`-kinship`	Kinship coefficient
`-pcoa`	Principal coordinate analysis
`-cluster`	Hierarchical clustering
`-structure`	Bayesian clustering

Functions

General settings: configuration of input and output files, output format (e.g., scientific notation, decimal places), temporary directory, sampling population for individuals, region definitions, number of threads, *single instruction multiple data* (SIMD) instructions (e.g., SSE, AVX, AVX512), and random number generator seed.

Filter: exclusion of some variants of low quality, genotyping ratio, or polymorphism, individuals with poor genotyping ratio, and genotypes of low quality or read depth before

file conversion and analyses. There are four types of filters: (i) variant information filters that exclude variants by their quality, type (e.g., single nucleotide polymorphism or indel) and original filter; (ii) genotype filters that exclude genotypes by their read depths, genotype qualities and ploidy levels. The excluded genotype will be set as the missing genotype; (iii) individual filters that exclude individuals by the number of variants typed and ploidy levels; and (iv) diversity filters that exclude variants based on the estimated genetic diversity indices (e.g., minor allele frequency, number of alleles, number of individuals genotyped, heterozygosity, significance of genotypic equilibrium test) in a specified reference population.

Haplotype extraction: several adjacent variants are combined into a highly polymorphic locus and the haplotypes are subsequently extracted. The haplotypes are used as alleles in subsequent analyses.

Conversion: genotypes can be converted into another genotype format, either GENEPOP (Rousset 2008), SPAGEDI (Hardy & Vekemans 2002), CERVUS (Kalinowski et al. 2007) , ARLEQUIN (Excoffier & Lischer 2010), STRUCTURE (Pritchard et al. 2000), POLYGENE (Huang et al. 2020), or POLYRELATEDNESS (Huang et al. 2015a).

Genetic diversity indices: this estimates the genetic diversity indices (e.g., observed and expected heterozygosity, effective number of alleles and polymorphic information content, Shannon's information index, inbreeding coefficient). A genotypic distribution test (i.e., the HWE test in polyploids) is performed using a Fisher's *G*-test, with the null hypothesis being the genotypic frequencies are in accordance with the prediction of a specific double-redouble model (e.g., RCS, PRCS, CES, PES).

Individual statistics: this enables the estimation of individual heterozygosity, the genotype likelihood, inbreeding coefficient and kinship coefficient. In polyploids, the heterozygosity of a genotype is the probability of randomly sampling two different *identical-by-state* (IBS) alleles without replacement (e.g., 2/3 for a tetraploid genotype *AABB*). The individual heterozygosity is the arithmetic average of the heterozygosity of genotypes of an individual across loci. The genotypic likelihood is the product of genotypic frequencies across loci of an individual. Three method-of-moment estimators (Loiselle *et al.* 1995; Ritland 1996; Weir 1996) are employed to estimate the individual kinship coefficient and inbreeding coefficient.

Genetic differentiation: estimates the differentiation index F_{ST} and tests for genetic differentiation between/among populations/regions. The F_{ST} estimators available are: Nei's (1973) G_{ST}, Weir & Cockerham's (1984)'s θ, Hudson *et al.*'s (1992), Slatkin's (1995) R_{ST}, Hedrick's (2005) G'_{ST} and Jost's (2008) D and Huang *et al.*'s (2021) variance decomposition. All estimators can be applied to both polyploids and anisoploids with the exception of Weir & Cockerham's (1984)'s θ. Differentiation is tested using Fisher's *G*-test for all loci or for each locus based on the distribution of genotypes or alleles.

Genetic distance: this enables the calculation of a variety of genetic distance indices between individuals, populations or regions: Nei's (1972) standard genetic distance, Cavalli-Sforza's (1967) chord distance, Reynolds *et al.*'s (1983) θ_w, Nei's (1983) D_A distance, Euclidean distance, Goldstein's (1995) distance, Nei's (1974) minimum genetic distance, Roger's (1972) distance, and two F_{ST} transformation-based genetic distances: (i) Reynolds *et al.*'s (1983) D and (ii) Slatkin's (1995) linearized F_{ST}. Based on the estimated

184 genetic distance matrices, the *principal coordinate analysis* (PCoA) and the *hierarchical*
185 *clustering analysis* (HCA) can be performed.

186 **Analysis of molecular variance**: Classical AMOVA only supports data for haploids
187 and diploids, and will only support from one to four hierarchies. Based on the generalized
188 framework (Huang *et al.* 2021), we extended AMOVA to accommodate any ploidy level
189 and any number of hierarchies. The generalized framework models the symbolic expres-
190 sion of the expected *Sum of Squares* (SS) $\mathbf{S}$ by variance components $\mathbf{\Sigma}$, which can be writ-
191 ten as $\mathbf{S} = \mathbf{C\Sigma}$. The method-of-moment estimate of variance components can be solved by
192 $\hat{\mathbf{\Sigma}} = \mathbf{C}^{-1}\hat{\mathbf{S}}$. Three methods are provided: (i) the homoploid method, (ii) the anisoploid
193 method and (iii) the likelihood method. The homoploid method uses the dummy haplo-
194 type method as in GENALEX (Peakall & Smouse 2006), and combines all loci into one
195 dummy locus, then calculate $\hat{\mathbf{S}}$ and $\mathbf{C}$ using the dummy haplotypes. This method can
196 only be applied to homoploids and is slightly biased when there are missing data. The
197 anisoploid model can be applied to both homoploids and anisoploids, which calculates $\hat{\mathbf{S}}$
198 and $\mathbf{C}$ using the alleles for each locus. The matrices $\hat{\mathbf{S}}$ and $\mathbf{C}$ are summed over loci, and
199 the variance component matrix $\hat{\mathbf{\Sigma}}$ is solved at once. Because the anisoploid method per-
200 mutes alleles at each locus to test the significance, it takes increased calculation time com-
201 pared with the homoploid method. VCFPOP uses a pseudo-permutation method to solve
202 this problem, which first perform a small number of permutations (e.g., 100) for each locus,
203 then subsamples one permutation at each locus to generate results for each pseudo-per-
204 mutation. For the likelihood method, the *F*-statistics are first estimated by maximizing gen-
205 otypic likelihood under differentiation or subdivision, and the variance components and

206 other statistics are subsequently solved.

207 **Population assignment**: enables the calculation of the likelihood for each individual
208 of being a member of a particular population (Paetkau *et al.* 2004). Each individual is as-
209 signed to the population with the maximum likelihood. Such assignment can help identify
210 the natal population for an individual.

211 **Kinship coefficient:** this calculates the kinship coefficient (θ) between two individu-
212 als, the probability that two randomly sampled alleles, each from one individual, are *iden-*
213 *tical-by-descent* (IBD). The same estimators (Loiselle *et al.* 1995; Ritland 1996; Weir 1996) are
214 employed to estimate the kinship coefficient between individuals.

215 **Relatedness coefficient:** this estimates the relatedness coefficient between individuals.
216 Two native polyploid relatedness estimators that supports a maximum ploidy level of
217 eight are provided: (i) the method-of-moment estimator (Huang *et al.* 2014), and (ii) the
218 maximum-likelihood estimator (Huang *et al.* 2015a). Moreover, the relatedness coefficient
219 can also be transformed from the kinship coefficient, with VCFPOP providing two transfor-
220 mations. The original transformation is used for outbred populations and is performed by

221 $$\hat{r}_{HL} = v_{\min}\hat{\theta}_{xy},$$

222 Where $\hat{r}_{HL}$ is the estimate of the relatedness coefficient from a higher ploidy individual to
223 a lower ploidy individual, $\hat{\theta}_{xy}$ is the kinship coefficient between these two individuals,
224 $v_{\min}$ is the ploidy level of the lower ploidy individual (Huang *et al.* 2015b). The modified
225 transformation accommodates inbreeding and is performed by

226 $$\hat{r}_{HL} = \frac{v_{\min}}{v_{\min} + v_{\max}}\hat{\theta}_{xy}\left(\frac{1}{\hat{\theta}_{xx}} + \frac{1}{\hat{\theta}_{yy}}\right),$$

227 where $v_{\max}$ is the ploidy level of the higher ploidy individual, and $\hat{\theta}_{xx}$ (or $\hat{\theta}_{yy}$) is the

kinship coefficient within the individual x (or y) (Huang *et al.* 2015a).

Bayesian clustering: This estimates ancestral proportions of each individual by the Markov Chain Monte Carlo (MCMC) method. VCFPOP follows the software STRUCTURE and implements three models of Bayesian clustering: (i) the ADMIXTURE model (Pritchard *et al*. 2000), (ii) the LOCPRIORI model (Hubisz *et al*. 2009) and (iii) the F model (Falush *et al*. 2003).

Optimization

NGS datasets are usually large, with a single VCF file often reaching hundreds of gigabytes in size. Workstations or computer clusters are usually required to analyze such data. VCFPOP uses various of methods to reduce the requirement and exploit the capacity of the computer:

(i) Optimized algorithm (to accelerate the calculation);

(ii) Advanced instruction set (e.g., SSE, LZCNT, POPCNT, AVX, FMA, AVX512);

(iii) Lock-free technology (to reduce access conflicts among threads);

(iv) Virtual memory allocation (to avoid re-allocation and memory move);

(v) Local memory management class (to allocate millions of small pieces of memory);

(vi) Variable length array (to place temporary array on stack memory);

(vii) Fast hash algorithm (to detect identical genotypes);

(viii) Fast hash table (to access genotypes by either hash value or index);

(ix) Indexing alleles and genotype (to share instances and reduce memory expense);

(x) Memory cache (to avoid frequent disk I/O).

VCFPOP is optimized for memory expense and calculation speed. For Intel ® processors

249 later than SkyLake (released in 2017) and CannonLake (released in 2018), the AVX-512
250 SIMD instructions can be used to accelerate the calculation speed, which enables the pro-
251 cessing of 512 bits simultaneously. For Intel ® processors later than Haswell hardware (re-
252 leased in 2013) and AMD ® processors after Excavator hardware (released in 2015), the
253 AVX instructions can be used, which enables the processing of 256 bits simultaneously.
254 These SIMD instruction-sets can be flexibly switched without additional compilation.

255 For memory usage, VCFPOP indexes the individual genotypes and uses bitwise storage
256 to save the genotype index. The memory usage for a genotype at a biallelic locus (e.g.,
257 single nucleotide polymorphism) is reduced from 4 bytes (e.g., '0/0 ') to 2 bits (4 possible
258 states: *AA*, *AB*, *BB* and missing data) for VCF format (16-folds compression). The detailed
259 information (e.g., ploidy level, alleles) of genotypes is saved in an additional table, with
260 each genotype using approximately 12 bytes. Because additional memory is required for
261 information at the locus, individual and population levels, the typical compression ratio is
262 14.5-fold evaluated by the Chr 22 phased 3 data of the 1000 genome project, which uses
263 734 Mib memory to load 10.4Gib data.

264 Although the compression ratio can reach 50-fold for some professional compression
265 algorithms (e.g., deflate, lzma, z-std), the random access of genotypes requires a longer
266 decompression time. Our method can compress data at a considerable compression ratio,
267 and can also access the genotypes without additional cost. In other words, there is a trade-
268 off between the accession speed and memory usage.

269 Therefore, a typical laptop with 16 GiB of memory can process 100 GiB VCF files with-
270 out considering the calculation speed. The subsequent analysis methods may also require

additional memory. For example, the homoploid AMOVA method saves the genetic distance between dummy haplotypes and requires an additional $8H^2$ bytes (about 745 MiB in a dataset with 5000 diploids), where H is the number of dummy haplotypes.

The loading speed of VCFPOP is also optimized, and uses a single thread to read the data from the disk and multiple threads to process the data. With a sample benchmark test of a 10.4 GiB uncompressed VCF file, VCFPOP can load data at 320 MiB/s and 560 MiB/s on a laptop (Intel i7-8750H CPU with 2.2GHz and 6 cores, 16 GiB memory, 256 Gib nvme SSD) and a workstation (Intel Xeon E5-2696 V4 CPU with 2.2GHz and 44 cores, 64 GiB memory and 1 TiB nvme SSD), respectively. Restricted by the additional decompression process, the loading speed is reduced to 255 and 460 MiB/s on the laptop and the workstation for the compressed format (`vcf.gz`), respectively.

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020302), the National Natural Science Foundation of China (31730104, 32170515, 31770411, 32070453), and the Innovation Capability Support Program of Shaanxi (2021KJXX-027). KH is supported by a scholarship from China Scholarship Council. KH would like to thank Prof. Kermit Ritland for providing a visiting professor position at the University of British Columbia.

References

Avni R, Nave M, Barad O, *et al.* (2017) Wild emmer genome architecture and diversity elucidate

wheat evolution and domestication. *Science* **357**, 93-97.
Burow MD, Simpson CE, Starr JL, Paterson AH (2001) Transmission genetics of chromatin from a synthetic amphidiploid to cultivated peanut (*Arachis hypogaea* L.): broadening the gene pool of a monophyletic polyploid species. *Genetics* **159**, 823-837.
Butruille DV, Boiteux LS (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.
Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics* **19**, 233.
Clark LV, Jasieniuk M (2011) POLYSAT: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources* **11**, 562-566.
Danecek P, Auton A, Abecasis G, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.
Excoffier L, Lischer HE (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.
Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.
Goldstein DB, Ruiz LA, Cavallisforza LL, Feldman MW (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.
Haldane JB (1930) Theoretical genetics of autopolyploids. *Journal of Genetics* **22**, 359-372.
Hardy OJ (2016) Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular Ecology Resources* **16**, 103-117.
Hardy OJ, 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.
Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* **59**, 1633-1638.
Huang K, Dunn DW, Ritland K, Li B (2020) polygene: Population genetics analyses for autopolyploids based on allelic phenotypes. *Methods in Ecology and Evolution* **11**, 448-456.
Huang K, Guo ST, Shattuck MR, *et al.* (2015a) A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity* **114**, 133-142.
Huang K, Ritland K, Guo S, Shattuck M, Li B (2014) A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources* **14**, 734-744.
Huang K, Ritland K, Guo ST, *et al.* (2015b) Estimating pairwise relatedness between individuals with different levels of ploidy. *Molecular Ecology Resources* **15**, 772–784.
Huang K, Wang T, Dunn DW, *et al.* (2021) A generalized framework for AMOVA with multiple hierarchies and ploidies. *Integrative Zoology* **16**, 33-52.
Huang K, Wang TC, Dunn DW, *et al.* (2019) Genotypic frequencies at equilibrium for polysomic inheritance under double-reduction. *G3: Genes, Genomes, Genetics* **9**, 1693-1706.
Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**, 1322-1332.
Hudson RR, 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.
Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* **16**, 1099-1106.
Ling HQ, Ma B, Shi XL, *et al.* (2018) Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature* **557**, 424.
Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Cardiology* **82**, 1420-1425.
Mather K (1935) Reductional and equational separation of the chromosomes in bivalents and multivalents. *Journal of Genetics* **30**, 53-78.
Meirmans PG, Tienderen PHV (2004) GENOTYPE and GENODIVE : two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* **4**, 792–794.
Muller HJ (1914) A new mode of segregation in Gregory's tetraploid primulas. *The American Naturalist* **48**, 508-512.
Nei M (1972) Genetic distance between populations. *American Naturalist* **106**, 283-292.
Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70**, 3321-3323.
Nei M, Roychoudhury AK (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.
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.
Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
Reynolds J, Weir BS, Cockerham CC (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.
Rogers JS (1972) Measures of similarity and genetic distance. In: *Studies in Genetics VII* (ed. Wheeler MR), pp. 145-153. University of Texas Publication, Austin.
Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.
Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457-462.
Weir BS (1996) *Genetic data analysis II: methods for discrete population genetic data* Sinauer Associates, Sunderland.
Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Data Accessibility and Benefit-Sharing Section

The source code, binary executables, user manual and example files (input files, parameters, and commands) are available from GitHub (http://github.com/huangkang1987/vcfpop).

Benefits from this research accrue from the sharing of our software and source code on public databases as described above.

Author Contributions

KH and BGL designed the project, KH and BY designed the software and wrote the draft, YFW and YXC reviewed the code, JCA, YHL and YCK performed simulations and tests, and DWD checked the model and helped write the manuscript.