POLYGENE V1.5 User Manual

This software uses the genotypic frequencies under double-reduction to estimate the allele frequencies from allelic phenotypes for polyploids. The software performs a variety of population genetic analyses including genetic diversity analysis, allelic phenotypic or genotype distribution, effective population size estimation, linkage disequilibrium test, genetic differentiation tests, genetic distance estimation, principal coordinate analysis, hierarchical clustering analysis, individual inbreeding coefficient estimation, individual heterozygosity index estimation, pairwise relatedness estimation, population assignment, parentage analysis, AMOVA, Bayesian clustering and Mantel tests.

Developed by Kang Huang

PhD of Zoology, Associate Professor

College of Life Sciences, Northwest University

No. 229, Taibai North Avenue

Xi'an City, Shaanxi Province, China

Zip code: 710069

E-mail: huangkang@nwu.edu.cn

Comments and suggestions are welcome.

Table of contents

Table of contents

Table of contents

1 System Requirement & Limitations

- CPU: x64 compatible
- Operation system:
  Microsoft Windows 7 or later
  Ubuntu 13.04 or later
  Mac OS X 10.9 to 10.13   (Apple disabled 32 bits programs since 10.14)
- Runtime library for Windows:
  .Net Framework 4.0 or later version
- Runtime library for Linux or Mac OS X:
  Mono 6.4.0 or later version
  zlib
- Memory: 2 Gb, more memory maybe required when large or high ploidy dataset is processed.
- Hard drive: 1 Gb

- Maximum level of ploidy: 10
- Maximum number of loci: 16776960
- Maximum number of individuals: 65535
- Maximum number of populations: 65535
- Maximum number of alleles at a locus: 150

2 Setup, Uninstall, Launch and Compile

This software can be downloaded via https://github.com/huangkang1987/polygene. The example files can also be downloaded from this address.

To setup this software, please create a folder on your hard drive, and then extract the files to this folder. For Windows, create a short-cut on the desktop. For Ubuntu or Mac OS X, create a symbolic link:

1. Open the terminal and open the install path

2. Set the file mode to executable by

```
chmod 777 ./PolyGene.ubuntu   (for Ubuntu)

chmod 777 ./PolyGene.mac      (for Mac OS X)
```

3. Make a symbolic link so as to launch POLYGENE at any directory. Note that the source

   path must be full path. For example,

   ```
   sudo ln -s -b '/home/HuangKang/PolyGene/PolyGene.ubuntu' /usr/bin/Po

   lyGene         (for Ubuntu)

   sudo ln -s -b '/home/HuangKang/PolyGene/PolyGene.mac' /usr/bin/PolyG

   ene         (for Mac OS X)
   ```

To uninstall, just delete the install folder, the short-cut or the symbolic link:

```
sudo rm /usr/bin/PolyGene      (for Linux and Mac OS X)
```

To launch the software:

Windows: double-click 'PolyGene.exe' or its short-cut;

Linux/Mac OS X: execute the command 'PolyGene.ubuntu' or 'PolyGene.mac'.

For the other operation systems, the users can compile POLYGENE from the provided

source code with the following steps:

1. Setup the following two compilers:

   GCC (http://gcc.gnu.org/);

   Mono (https://www.mono-project.com);

2. Open the terminal and open the install path;

3. Set the file mode of 'makefile_linux.sh' or 'makefile_mac.sh' to executable by

   ```
   chmod 777 ./makefile_linux.sh

   chmod 777 ./makefile_mac.sh
   ```

4. Run 'makefile_linux.sh' or 'makefile_mac.sh' and wait for about 30 seconds;

5. The binary dynamic library 'dre.so' or 'dre.dylib' and executable

   'PolyGene.linux' or 'PolyGene.mac' are generated in the 'bin' folder.

6. Setup POLYGENE, see above.

3 Citation

Basic functions

Huang K, Dunn DW, Ritland K, Li BG (2020) POLYGENE: population genetics analyses for autopolyploids based on allelic phenotypes. *Methods in Ecology and Evolution*, 11, 448-456. doi:10.1111/2041-210X.13338.

Genotypic frequencies under polysomic inheritance

Huang K, Wang TC, Dunn DW, Zhang P, Cao XX, Liu RC, Li BG (2019) Genotypic frequencies at equilibrium for polysomic inheritance under double-reduction. *G3: Genes, Genomes, Genetics*, 9, 1693-1706. doi:10.1534/g3.119.400132.

AMOVA

Huang K, Wang TT, Dunn DW, Zhang P, Sun HJ, Li BG (2021) A generalized framework for AMOVA with multiple hierarchies and ploidies. *Integrat*ive *Zoology*, 16, 33-52, doi:10.1111/1749-4877.12460.

Parentage analyses

Huang K, Huber G, Ritland K, Dunn DW, Li BG (2021) Performing parentage analysis for polysomic inheritances based on allelic phenotypes. *G3: Genes, Genomes, Genetics*, 11, jkaa064. doi:10.1093/g3journal/jkaa064

Methods-of-moment relatedness estimator

Huang K, Ritland K, Guo ST, Shattuck M, Li BG (2014) A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources*, 14, 734-744. doi:10.1111/1755-0998.12217.

Maximum-likelihood relatedness estimator

Huang K, Guo ST, Shattuck MR, Chen ST, Qi XG, Zhang P, Li BG (2015) A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity*, 114, 133-142. doi:10.1038/hdy.2014.88.

Effective population size

To be added after publication.

Correlated sample correction for F_{ST} and relatedness

To be added after publication.

Heritability and Q_{ST} estimators

To be added after publication.

4 Usage

POLYGENE is designed to work for a single project, and the workspace is the directory of the executable files. If the user wants to change to another project, the executable file should be copied into another directory. To run POLYGENE, double-click the executable file 'PolyGene.exe'.

4.1 A brief introduction

Various functions of POLYGENE are each available on separate pages. These consist of 'Input', 'Parameters', 'Diversity', 'Frequency', 'Distribution', 'Linkage', 'Ne', 'Differentiation', 'Distance', 'Ordination', 'Clustering', 'Inbreeding', 'H-Index', 'Assignment', 'Spatial', 'Relationship', 'Heritability', 'Parentage', 'AMOVA', 'Bayesian' and 'Mantel'. Figures 1 and 2 show the first two pages: 'Input' and 'Parameters'.

4 Usage

Figure 1. The 'Input' page

Figure 2. The 'Parameters' page

A population simulator can be used to simulate population genetic drift (Figure 1) and to generate the input data. The simulation parameters, including the ploidy levels, population sizes, locus numbers and the allele frequency at each locus, can be configured in the 'Simulation: Population and allele frequencies' box on the page 'Input'. The bottom box 'Input data' is used to input the coordinates, and quantitative trait values and allelic phenotype/genotype data for all individuals.

All parameters used for analyses can be configured on the page 'Parameters' (Figure 2). The subsequent pages present the results of the corresponding analyses. Performing the analyses in POLYGENE is straightforward: first, format the allelic phenotype data and paste into the box marked 'Input data'; second, select the required analytical methods and configure the parameters on the page 'Parameters'; and finally click the button 'Calc' in the toolbox at the top of the window. The progress bar will show the current analysis and its progress. After a few minutes, the results will be presented on the corresponding pages. This software supports ploidy levels from 1 to 10, but odd levels of ploidies are only supported by the disomic/RCS model with or without selfing.

4.2 Input data format

The format of inputting data is shown in Boxes 1 and 2, where the first row is the header row. The subsequent rows are the individual information, coordinates, quantitative trait values and allelic phenotype/genotype data. The individual information includes the individual identifier, population and ploidy level. Note that the ploidy levels of individuals within the same population **should be equal**, so as to ensure that some analyses can be performed, e.g., allelic phenotype/genotype distribution test and parentage analysis. The following columns denote the coordinates and phenotypic values **(they are optional)**, the header of a coordinate is start with '[C]', while that for a quantitative trait is start with '[Q]' (Box 1). The columns are separated by tabs and the

4 Usage

alleles within an allelic phenotype/genotype are identified by **positive** integers and separated by commas (Box 1). The blank cells denote the missing data.

```
ID      Pop    Ploidy  [C]X    [Q]Q1    Loc1    Loc2
Ind1    pop1   4       -1.83   118.30   2,3,4   1,2,4
Ind2    pop1   4       -2.99   100.86   4       2,3,4
Ind3    pop1   4       -8.25   115.63   3,4     3,4
Ind4    pop1   4       3.24    110.76   2,3,4
Ind5    pop1   4       -2.07   107.23   1,4     1,2,3,4
Ind6    pop2   4       -4.14   116.37   2,3,4   3,4
Ind7    pop2   4       -4.42   111.88           2,3,4
Ind8    pop2   4       8.66    100.70   3,4     2,4
Ind9    pop2   4       2.50    113.38   1,3,4   3,4
Ind10   pop2   4       -1.17   114.08   2,3,4   2,3,4
```

Box 1. Example of allelic phenotype data with a coordinate and a phenotypic value of a quantitative trait

The difference between allelic phenotype and genotype is that the allelic phenotype is a set of allele copies (with 1 to v elements, v is the ploidy level), while genotype is a multiset of allele copies (with v elements). Therefore, there should not be duplicated alleles in the allelic phenotypes, and POLYGENE will remove the duplicated alleles in the allelic phenotypes and give a warning message.

```
ID      Pop    Ploidy L1   L2        ID      Pop    Ploidy L1,L2
Ind1    pop1   4      0    2         Ind1    pop1   4      02
Ind2    pop1   4      1    2         Ind2    pop1   4      12
Ind3    pop1   4      2    3         Ind3    pop1   4      23
Ind4    pop1   4      2    4         Ind4    pop1   4      24
Ind5    pop1   4      1    2         Ind5    pop1   4      12
Ind6    pop2   4      3              Ind6    pop2   4       4
Ind7    pop2   4      2    3         Ind7    pop2   4      2
Ind8    pop2   4      1    2         Ind8    pop2   4      12
Ind9    pop2   4      4    2         Ind9    pop2   4      42
Ind10   pop2   4      1    1         Ind10   pop2   4      11
```

Box 2. Example of one-digit genotype (a) with or (b) without deliminator

For large diallelic SNP genotype dataset, POLYGENE supports a one-digit genotype format, where the deliminator can be either present (Box 2a) or absent (Box 2b). In this format, a genotype is encoded as a hexadecimal digit, where 'A' or 'a' denotes 10 for decaploids. The missing data is identifier by blank cells for data with deliminator, and by spaces for

4 Usage

data without deliminator. The missing data are usually not analyzed, except for the dummy haplotype-based methods (Huang *et al.* 2020) including Bayesian clustering, AMOVA and Weir & Cockerham's (1984) F_{ST} estimator.

For huge inputting data that cannot be placed in the input textbox, POLYGENE will automatically read the inputting data from 'input.txt' when the textbox is empty, or from the file path in the textbox when it has only one line.

For genotype data, if the 'Remove duplicate alleles' option is used, then the genotypes will be converted into allelic phenotypes before analyses. Otherwise, all analyses will be performed based on genotypes. Note that the null allele and negative amplification options cannot be used for the genotype data, these options are used to estimate the allele frequencies from allelic phenotypes and calculate the expected genotypic or allelic phenotypic frequencies.

For allelic phenotype data, POLYGENE will use the posterior probabilities to weight the underlying genotypes in the weighted genotype-based methods (including allele frequency estimation, genetic diversity analysis, genetic distance, parentage analysis, spatial pattern analysis, heritability estimation and relatedness estimation) and dummy haplotype-based methods (including Bayesian clustering, AMOVA and Weir & Cockerham's (1984) F_{ST} estimator) (Huang *et al.* 2020). The inheritance model can influence the priori probability of genotypes, and the detail can be configured in the 'Allele frequency estimator' groupbox. A brief description of the inheritance models is given in section 5.1.

POLYGENE also provides the functions that allow for the importation of the genotype data from GENEPOP (Rousset 2008) and STRUCTURE (Pritchard *et al.* 2000), POLYRELATEDNESS (Huang *et al.* 2014), SPAGEDI (HARDY & VEKEMANS 2002) and VCF (variant calling format),

which can be performed by clicking the 'Genepop', 'Structure', 'PolyRelatedness', 'Spagedi' or 'VCF' buttons at the top in the page 'Input'.

POLYGENE supports multi-level regions, and the region of populations is defined in the 'Global setting' groupbox in the page 'Parameter' (See section 4.4).

4.3 Population simulator

Another function on the page 'Input' is the population simulator. The ploidy and allele frequencies of each population simulated can be configured at the top of this page, while the simulation parameters are at the right of this page (Figure 1).

Before the simulation, the initial allele frequencies, ploidy and size of each population need to be configured (see the example in Box 3). The columns in Box 3 are also separated by tabs, and the null alleles are identified by 'null' or '0'. Note that the sum of initial frequencies of all alleles at a locus should be equal to one. The null alleles will be removed from the outputting allelic phenotypes. If a genotype is a homozygote comprised of null alleles, its observed allelic phenotype will be empty (missing data).

```
ploidy  4       ninds   10
Loc1            Loc2
null    0.1     1       0.1
2       0.2     2       0.2
3       0.3     3       0.3
4       0.4     4       0.4
ploidy  4       ninds   10
Loc1            Loc2
null    0.1     1       0.1
2       0.2     2       0.2
3       0.3     3       0.3
4       0.4     4       0.4
```

Box 3. Example of inputting the relevant data to population simulator

The reproductive parameters can be configured in the right of the page 'Input', including the sex ratio, dioecy or monoecy, selfing rate, sampling rate, negative PCR rate,

distance from a locus to the corresponding centromere and output data (allelic phenotype or genotype).

The 'Distance from locus to centromere (cM)' box configures the inheritance model or the distance to centromere of a locus, and each line represent that of a locus. The inheritance model can be 'disomic' or 'RCS' (*random chromosome segregation*), 'PRCS' (*pure random chromatid segregation*), 'CES' (*complete equational segregation*). If it is a real number, then the PES (*partial equational segregation*) model is used. For this model, if a locus is far away from its corresponding centromere, then the double-reduction at this locus is more likely to occur.
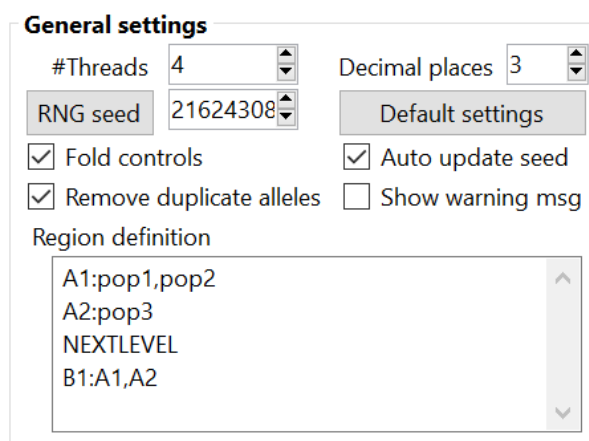
The random number generator seed can be configured in the 'RNG seed' box in the page 'Parameter'. After various parameters have been configured, click the button 'Founder' at the top of the page 'Input' to generate the founder populations. The genotypes of a founder population will then be generated according to the selected inheritance model and parameters. Next, click the button at the right of 'Reproduce' to reproduce the population for a number of generations to simulate the genetic drift. Alternatively, click the button 'Until Fst=' to reproduce the population until the differentiation coefficient F_{ST} reaches a specified value, where F_{ST} is estimated by Nei's (1973) G_{ST} estimator based on the true genotypes. After the reproduction has ended, the data for allelic phenotypes or genotypes are shown in the box 'Input data'.

4.4 General settings

General settings are shown in Figure 3. The '#Threads' specifies the number of threads used in performing the analyses, which can be equal to the number of CPU cores to accelerate the calculations, the 'Decimal places' is used to specify the decimal places of outputting real numbers, and 'RNG seed' is used to specify the seed used by the random

number generator. The same seed and parameters ensure the same results for the analyses. These analyses consist of the population simulation, Markov chain test (Raymond & Rousset's test), MCMC (Bayesian clustering), Monte-Carlo simulation (parentage analysis) and the permutation tests (AMOVA and Mantel's test), dummy genotype exportation. If more than one thread is used, the seed of the ith thread is the global seed XOR i. Therefore, the results can be changed when the numbers of threads are different in the same analysis process. Click the 'RNG seed' button can change a new seed. The 'Default settings' button reset the configurations.

Figure 3. General parameters

For the following checkboxes, 'Fold controls' hides unused analyses options and pages, 'Auto update seed' changes a new seed in each run, 'Remove duplicate alleles' converts genotype input into allelic phenotype input, and 'Show warning msg' prompts a warning dialog when there is something needs to be noticed.

POLYGENE supports multi-level regions, where a region is consisting of several populations or low-level regions. The regions are defined in the 'Region definition' textbox (Figure 3). The format is as follows: each row is used to define a region, and the text before the semi-colon is the region identifier, follow by the populations or low-level regions separated by commas. The populations (or regions) that have not been assigned to a region (or to a higher-level region) are classified into a default region. The regions

are sorted from the lowest to the highest levels, with 'NEXTLEVEL' used to separate the regions at different levels.

4.5 Allele frequency estimation

In POLYGENE, an EM algorithm is used to estimate the allele frequencies (see Section 5.1 for details). This algorithm was developed by Kalinowski *et al*. (2006) and can also be used to estimate the frequency of null alleles.

Figure 4. Parameters of estimating allele frequencies

The parameters of allele frequencies are shown in Figure 4, which include the options 'Null alleles', 'Negative PCR', 'Selfing', 'Selfing rate estimator' and 'Inheritance model'.

If the option 'Null alleles' is checked, the candidate genotypes with null alleles will be generated. For example, in tetraploids there are three additional candidate genotypes *ABYY*, *AABY* and *ABBY* determining the allelic phenotype *AB*, where *Y* is the null allele. Besides null allele, low DNA quality, low amplification efficiency or experimental errors can also lead to missing data. These factors may overestimate the null allele frequency. The option 'Negative PCR' can also be used to explain the missing data and prevent such overestimation.

The option 'Selfing' will use an approximated genotypic frequency to estimate the allele

4 Usage

frequency and calculate the posterior probabilities of genotypes hidden behind an allelic phenotype (See section 5.1). The selfing rate of each population is estimated independently. The option 'Selfing' is **incompatible** with the 'Polysomic (PES estimating rs)' inheritance model, because the selfing rate and the recombination fraction r_s will both influence the inbreeding coefficient. The estimated selfing rate together with the negative PCR rate is shown in the results of allele frequencies.

POLYGENE will use the selected inheritance model in the 'Inheritance model' box to estimate the allele frequencies, generate the candidate gametes, calculate the allelic phenotypic or genotypic frequencies, and will also use the Bayes formula to obtain the posterior probabilities of genotypes determining an allelic phenotype. For example, the genotype $ABCD \times ABCD$ can produce an offspring genotype $AAAA$ due to the double-reduction. However, this will be identified as a mismatch under the disomic, RCS model, or PES model with $r_s = 0$, while it is not a mismatch under other double-reduction models.

```
Loc1            5  Loc2            5
Total
null    0.149     null    0.138
1       0.055     1       0.046
2       0.110     2       0.181
3       0.218     3       0.245
4       0.468     4       0.389
pop1
null    0.000     null    0.311
1       0.056     1       0.105
2       0.112     2       0.164
3       0.190     3       0.174
4       0.641     4       0.246
pop2
null    0.335     null    0.000
1       0.052     1       0.000
2       0.108     2       0.195
3       0.252     3       0.302
4       0.252     4       0.503
```

Box 4. Example results of allele frequencies

An example of calculating results is shown in Box 4, where the first line shows the identifier of loci and the number of alleles (including null alleles). The frequency of an allele *A* at a given locus and in the total population (or in a region) is the weighted average of the frequencies of *A* at the same locus and in all populations (or in the populations included in this region), with the numbers of copies of *A* at the same locus and in the corresponding populations as the weights. Note that any missing data is not considered in the calculating process if neither null allele and negative PCR are considered.

4.6 Genetic diversity

An example of the genetic diversity indices is shown in Figure 5, where in order to show all data at a locus by one row, the decimal place is set to 1. The detail description of each genetic diversity index is shown in section 5.2.

Each genetic diversity index in the first four rows is calculated by taking the average of the corresponding indices at all loci for k, n, Ho, He, PIC, Ar or Fis, or by taking the product of the corresponding indices at all loci for NE-1P, NE-2P, NE-PP, NE-I and NE-SI. Beginning with the fifth row, each is the genetic diversity index for the total population, a region or a population at a locus.

4 Usage

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Pop	Locus	k	n	Ho	He	PIC	Ae	Ar	I	NP	NE-1P	NE-2P	NE-PP	NE-I	NE-SI	Fix	Fsx
2	Total	Avg	12	680	0.55	0.67	0.63	4.8	12	1.53	12	0.02	0	0	0	0	0.19	0.03
3	Laguna	Avg	8.63	100	0.5	0.59	0.56	3.84	10.24	1.34	0.38	0.04	0.01	0	0	0	0.15	
4	Campeche	Avg	8.75	99.88	0.53	0.62	0.58	3.68	10.53	1.35	0.25	0.05	0.01	0	0	0	0.15	
5	Palhoca	Avg	9.5	99.38	0.57	0.67	0.63	4.51	10.77	1.48	1	0.03	0	0	0	0	0.17	
6	Joaquina	Avg	10	380.75	0.56	0.65	0.62	4.28	10.58	1.45	0.88	0.03	0	0	0	0	0.17	
7	Total	Cv14	15	679	0.67	0.82	0.8	5.58	15	1.94	15	0.52	0.35	0.17	0.1	0.35	0.19	0.04
8	Total	Cv03	18	680	0.83	0.87	0.86	7.58	18	2.26	18	0.41	0.26	0.1	0.06	0.32	0.05	0.04
9	Total	Cv08	5	681	0.17	0.26	0.25	1.36	5	0.54	5	0.97	0.87	0.76	0.59	0.76	0.34	0.01
10	Total	Cv06	6	680	0.41	0.68	0.62	3.09	6	1.29	6	0.74	0.58	0.4	0.27	0.45	0.39	0.04
11	Total	Cv11	23	678	0.75	0.91	0.9	10.86	23	2.61	23	0.31	0.18	0.05	0.03	0.3	0.17	0.02
12	Total	Cv07	7	681	0.43	0.47	0.4	1.89	7	0.79	7	0.89	0.78	0.65	0.5	0.6	0.08	0.04
13	Total	Cv16	18	680	0.81	0.84	0.82	6.08	18	2.01	18	0.5	0.33	0.15	0.09	0.34	0.03	0.03
14	Total	Cv15	4	681	0.33	0.48	0.39	1.93	4	0.77	4	0.88	0.79	0.68	0.54	0.6	0.31	0.04
15	Laguna	Cv14	13	100	0.74	0.8	0.78	5.05	14.23	1.93	1	0.54	0.36	0.17	0.09	0.36	0.08	
16	Laguna	Cv03	12	100	0.76	0.77	0.75	4.39	14.9	1.85	0	0.59	0.41	0.21	0.12	0.38	0.02	
17	Laguna	Cv08	3	100	0.2	0.21	0.19	1.27	3.35	0.39	0	0.98	0.9	0.83	0.68	0.8	0.06	
18	Laguna	Cv06	5	100	0.28	0.57	0.52	2.35	5.8	1.08	0	0.82	0.67	0.5	0.34	0.52	0.51	
19	Laguna	Cv11	19	100	0.72	0.89	0.88	9.31	20.53	2.5	1	0.35	0.21	0.07	0.04	0.31	0.19	
20	Laguna	Cv07	3	100	0.21	0.24	0.22	1.31	5.84	0.44	0	0.97	0.89	0.81	0.65	0.78	0.1	
21	Laguna	Cv16	12	100	0.76	0.81	0.79	5.22	15.14	1.92	1	0.54	0.36	0.17	0.1	0.36	0.06	
22	Laguna	Cv15	2	100	0.35	0.44	0.34	1.79	2.12	0.63	0	0.9	0.83	0.74	0.6	0.63	0.21	
23	Campeche	Cv14	11	99	0.59	0.76	0.73	4.17	14.15	1.74	0	0.62	0.43	0.24	0.14	0.39	0.22	
24	Campeche	Cv03	13	100	0.75	0.76	0.74	4.23	15.81	1.84	0	0.6	0.42	0.22	0.12	0.39	0.02	
25	Campeche	Cv08	2	100	0.15	0.16	0.15	1.19	2.35	0.3	0	0.99	0.93	0.87	0.74	0.85	0.07	
26	Campeche	Cv06	5	100	0.37	0.62	0.55	2.6	5.8	1.08	0	0.8	0.66	0.5	0.35	0.5	0.39	
27	Campeche	Cv11	18	100	0.79	0.88	0.87	8.22	20.39	2.36	0	0.39	0.24	0.09	0.05	0.32	0.1	
28	Campeche	Cv07	5	100	0.44	0.44	0.39	1.78	6.73	0.8	1	0.9	0.78	0.65	0.48	0.62	0.01	
29	Campeche	Cv16	13	100	0.82	0.8	0.77	4.97	15.89	1.78	1	0.57	0.39	0.21	0.13	0.37	-0.03	
30	Campeche	Cv15	3	100	0.32	0.57	0.48	2.31	3.12	0.92	0	0.84	0.72	0.58	0.45	0.54	0.43	
31	Palhoca	Cv14	12	99	0.67	0.77	0.73	4.3	14.43	1.72	0	0.62	0.44	0.25	0.15	0.39	0.13	
32	Palhoca	Cv03	12	99	0.86	0.87	0.85	7.49	14.91	2.16	0	0.42	0.27	0.11	0.06	0.32	0.01	
33	Palhoca	Cv08	5	100	0.29	0.42	0.39	1.72	5	0.82	1	0.91	0.77	0.62	0.43	0.63	0.3	
34	Palhoca	Cv06	6	99	0.35	0.66	0.6	2.91	6	1.27	1	0.76	0.59	0.41	0.27	0.46	0.46	
35	Palhoca	Cv11	21	99	0.82	0.91	0.9	10.85	21.74	2.63	2	0.31	0.18	0.05	0.03	0.3	0.1	
36	Palhoca	Cv07	6	100	0.45	0.51	0.39	2.04	6.73	0.77	2	0.87	0.8	0.69	0.59	0.58	0.12	
37	Palhoca	Cv16	11	99	0.83	0.8	0.77	5.06	14.32	1.79	1	0.57	0.39	0.21	0.12	0.37	0.04	

Figure 5. Example results of genetic diversity indices

4.7 Allelic phenotype or genotype distribution test

In diploids, the Hardy-Weinberg equilibrium test is used to evaluate whether the distribution of genotypes accords with the HWE. For polyploids, POLYGENE uses Fisher's *G*-test to perform the genotypic or allelic phenotype distribution test. The null hypothesis is that the distribution of genotype or allelic phenotype accords with the expected frequencies under a specific double-reduction inheritance model (e.g., disomic/RCS, PRCS, CES, PES).

The parameters for the allelic phenotype or genotype distribution test are shown in Figure 6, whose range includes 'Population', 'Region' and 'Total population'. The expected allelic phenotypic or genotypic frequencies are obtained from the allele frequencies in the selected ranges.

Figure 6. Parameters of allelic phenotype or genotype distribution test

An example results from an allelic phenotype distribution test is shown in Box 5. If the ploidy levels of all populations in a region or in the total population are not all equal, POLYGENE will not perform this distribution test for this region or the total population.

The false discovery rate (FDR) correction (for each population, each region or the total population) is performed for multiple tests (Benjamini & Hochberg 1995). Here, the tests with zero degrees of freedom will be excluded from FDR correction. The data of #Par (numbers of parameters), lnL (natural logarithms of allelic phenotypic or genotypic likelihoods), AICc and BIC are also shown in the results.

```
Allelic phenotype distribution test
Pop    Locus  G      d.f.  P-val  FDR-Q  #Pheno  #Par  lnL      AICc    BIC
Total  Loc1   11.05  14    0.68   1.00   20      6     -41.34   101.14  100.65
Total  Loc2   13.08  14    0.52   0.52   20      6     -41.72   101.91  101.42
reg1   Loc1   11.05  14    0.68   1.00   20      6     -41.34   101.14  100.65
reg1   Loc2   13.08  14    0.52   0.52   20      6     -41.72   101.91  101.42
pop1   Loc1   9.36   14    0.81   0.81   10      3     -20.77   51.55   48.46
pop1   Loc2   9.03   14    0.83   1.00   10      3     -18.70   47.40   44.31
pop2   Loc1   6.16   14    0.96   1.00   10      3     -20.56   51.13   48.03
pop2   Loc2   9.36   14    0.81   0.81   10      3     -23.02   56.05   52.95
```
Box 5. Example results of allelic phenotype distribution test

4.8 Linkage disequilibrium test

A linkage disequilibrium test either based on Fisher's *G*-test or on the Raymond & Rousset's (1995) Markov Chain test is used by POLYGENE, whose parameters are shown in Figure 7. Here, 'range' includes the three options 'population', 'region' and 'total population', and the option 'Markov Chain test' includes '#Burnin', '#Batches' and '#Iterations,' which are used to specify the length of the burnin period, the number of batches and the number of iterations for each batch, respectively. For the Markov Chain test, each value of SE is the standard error of the corresponding *P* values among all

batches.

Figure 7. Parameters of linkage disequilibrium test

An example results of a linkage disequilibrium test is shown in Figure 8. The FDR correction (Benjamini & Hochberg 1995) is made for each population, each region or the total population. The results have been pasted into a spreadsheet because of the many columns

Linkage disequilibrium test based on phenotypic data

Pop	A	B	Fisher's G test				Raymond & Rousset's Markrov Chain test				Burrows's Delta test				Raymond & Rousset's Markrov Chain test				
			G	df	P	FDR-Q	switches	SE	P	FDR-Q	r2	G	df	P	FDR-Q	switches	SE	P	FDR-Q
Total	Loc1	Loc2	NaN	NaN	NaN	NaN	0	NaN	NaN	NaN	0.00201	8.88257	9	0.44818	0.44818	470289	0.01892	0.55178	0.55178
pop1	Loc1	Loc2	13.8228	20	0.83937	0.83937	353115	0.00341	0.86351	0.86351	0.00398	13.8493	9	0.12779	0.12779	435712	0.00654	0.13868	0.13868
pop2	Loc1	Loc2	22.5256	16	0.12702	0.12702	361351	0.00417	0.15202	0.15202	0.00295	2.96114	9	0.96582	0.96582	466024	0.00317	0.96935	0.96935

Figure 8. Example results of linkage disequilibrium test

4.9 Effective population size

The effective population size is the size of an ideal Wright-Fisher population that shows the same allele frequency change over time as an observed biological population regardless of its census population size (Hamilton 2009). POLYGENE provides three effective population size estimator: Pudovkin *et al.*'s (1996) estimator based on heterozygosity excess (HE), Nomura's (2008) estimator based on molecular coancestry (MC), and Waples & Do's (2010) estimator based on linkage disequilibrium (LD). Note that, the ploidy of Wright-Fisher population is the same as target population. Therefore, for those regions or total populations with varying ploidy levels, the effective population size is not estimated.

Figure 9. Parameters of effective population size estimation

Pudovkin *et al.*'s (1996) estimator assumes dioecious population, random mating and non-overlapping generations and other assumptions to allow Hardy-Weinberg equilibrium. Under these assumptions, the observed heterozygosity can be slightly higher than expect due to the binomial sampling variance in allele frequencies in the breeding males and females. However, in the presence of population **subdivision, double-reduction and inbreeding/selfing,** the observed heterozygosity is reduced and this estimator can be biased.

Nomura's (2008) estimator use the parental generation as the reference and use parent-based coancestry coefficient to estimate the effective population size. This estimator classified **a predefined proportion of individual pairs as nonrelatives** (e.g., 0.95), and use these individual pairs to estimate squared parental allele frequencies.

Waples & Do's (2010) estimator measuresthe linkage disequilibrium based on Burrows' Δ (Cockerham & Weir 1977). The expectation of squared disequilibrium coefficients r_Δ^2 is influenced by the mating system, the recombination fraction, the effective population size and the sample size. If three in four parameters is known, the remaining parameter can be solved. The **number of chromosomes (#chromosomes)** and the **mating system** can be configured in the parameter box. The **number of chromosomes** is used to determine the recombination fraction. The **mating system** includes: haplotype sampling

(HS), monecious with selfing (MS), monecious excludes selfing (ME), dioecious with random pairing (DR) and dioecious with lifetime pairing (say DH). The DH mating system needs to use the sex ratio as an extra parameter.

An example results of effective population size estimation is shown in Figure 8. Five populations with ploidy levels from 2 to 10 are simulated at 960 tetra-allelic loci and are reproduced for 10 generations.

Pop	Ploidy	#ind	Pudovkin 2009					Nomura 2008				Waples 2010				
			L	D	Ne	SE(Ne)	95% CI	theta_xy	Ne	SE(Ne)	95% CI	r^2_Delta	Ne	SE(Ne)	95% CI
pop1	2	100	960	0.01	86.39	0.00	86.38~86.39	0.00	108.14	0.89	106.4~109.87	0.01	127.42	2.41	122.69~132.15
pop2	4	100	960	0.00	2075.56	0.00	2075.56~2075.56	0.00	100.59	0.84	98.94~102.24	0.01	128.63	1.60	125.48~131.77
pop3	6	100	960	0.00	94.88	0.00	94.88~94.88	0.00	89.60	0.82	88~91.19	0.01	115.82	1.53	112.83~118.81
pop4	8	100	960	0.00	158.79	0.00	158.79~158.79	0.00	105.03	1.01	103.04~107.01	0.01	142.63	2.03	138.66~146.6
pop5	10	100	960	0.00	93.76	0.00	93.76~93.76	0.00	97.62	0.84	95.97~99.27	0.01	133.02	1.67	129.74~136.3
Total	0	500	960	∞	NaN~Nan	NaN	NaN	0.00	∞	NaN	NaN~Nan	0.01	∞	NaN	NaN~Nan

Figure 10. Example results of effective population size estimation

4.10 Genetic differentiation

For the genetic differentiation analysis, several F_{ST} analogous indices are calculated and the differentiation between/among populations or regions is tested. The parameters of genetic differentiation are shown in Figure 11. Among these, Huang unpub estimator can correct the biased from dependent samples (e.g., relatives and correlated subpopulations).

Figure 11. Parameters of genetic differentiation

There are the following five kinds of parameter combinations in the 'range'. For each

4 Usage

kind of combinations, a genetic differentiation test is performed and the corresponding F_{ST} analogous index is calculated.

(i)   'Total population' × 'Region', i.e., among all regions in the total population;

(ii) 'Total population' × 'Population', i.e., among all populations in the total population;

(iii) 'Region' × 'Population', i.e., among all populations in the same region (if a multi-level region is used, the genetic differentiation test and the F_{ST} analogous index among all populations in the same higher-hierarchy region will be calculated);

(iv) only 'Region', i.e., between any two regions;

(v)   only 'Population', i.e., between any two populations.

The F_{ST} estimators are as follows: Weight Genotype (infinity allele model (IAM) or stepwise mutation model (SMM)), Slatkin's (1995) R_{ST}, Nei's (1973) G_{ST}, Hudson *et al.*'s (1992) F_{ST}, Hedrick's (2005) G'_{ST}, Jost's (2008) D and Huang *et al.* (unpublished). Huang *et al.* (unpublished) estimator can eliminate the influence of correlated samples. The index for Weight Genotype is based on the variance decomposition modified from Weir & Cockerham's (1984) estimator θ (see Section 5.6 for details), where the genetic distance between two alleles can use either the IAM distance or the SMM distance. The other indices are calculated based on the allele frequencies. An example results for these analyses is shown in Box 6.

```
Method: Huang et al. unpublished Weight Genotype IAM
Among all regions = 0.02

Among all populations = 0.00

Among all populations in reg1 = 0.02

Between regions:   reg1
reg1   0.00

Between populations:    pop1 pop2
pop1   0.00    0.02
pop2   0.02    0.00
```
Box 6. Example results of genetic differentiation test (part 1)

4 Usage

The tests for genetic differentiation are performed by using either the Fisher's *G*-test or the Raymond & Rousset's (1995) Markov Chain test based on the allelic phenotype/genotype distribution or the allele distribution. An example results of such an analysis is shown in Figure 12.

Phenotypic differentiation test														
		All loci			Loc1						Loc2				
		Fisher's G test			Fisher's G test			Raymond & Rousset's te			Fisher's G test			Raymond & Rousset's test		
A	B	G	d.f.	P	G	d.f.	P	switches	S.E.	P	G	d.f.	P	switches	S.E.	P
Among all populations		13.5	15	0.56	7.27	8	0.51	162016	0	0.91	6.22	7	0.51	169835	0	0.95
Among all populations in reg1		13.5	15	0.56	7.27	8	0.51	162016	0	0.91	6.22	7	0.51	169835	0	0.95
pop2	pop1	13.5	15	0.56	7.27	8	0.51	162036	0	0.91	6.22	7	0.51	169716	0	0.95
Allelic differentiation test															
		All loci			Loc1						Loc2				
		Fisher's G test			Fisher's G test			Raymond & Rousset's te			Fisher's G test			Raymond & Rousset's test		
A	B	G	d.f.	P	G	d.f.	P	switches	S.E.	P	G	d.f.	P	switches	S.E.	P
Among all populations		6.25	6	0.4	4.83	3	0.18	362869	0	0.18	1.42	3	0.7	357064	0	0.71
Among all populations in pop1		6.25	6	0.4	4.83	3	0.18	362869	0	0.18	1.42	3	0.7	357064	0	0.71
pop2	pop1	6.25	6	0.4	4.83	3	0.18	363289	0	0.18	1.42	3	0.7	356766	0	0.71

Figure 12. Example results of genetic differentiation test (part 2)

4.11 Genetic distance

The genetic distance between individuals, populations or regions can be calculated; calculation options are configured in the box 'Genetic distance' (Figure 13). Various estimators of genetic distances can be selected.

Figure 13. Parameters of genetic distances

Reynolds *et al.*'s (1983) distance D_{RA} and Slatkin's (1995) linearized distance D_{Sl} can be obtained by transforming the corresponding F_{ST} estimators shown in the

groupbox 'Genetic differentiation' (Figure 11).

The results are placed on the page 'Distance' and an example of these results is shown in Box 7 (see Section 5.7 for details).

```
Nei's standard genetic distance for population
          pop1        pop2
pop1      0.000       0.039
pop2      0.039       0.000

Cavalli-Sforza chord distance for population
          pop1        pop2
pop1      0.000       0.919
pop2      0.919       0.000

Reynolds et al.'s theta for population
          pop1        pop2
pop1      0.000       0.186
pop2      0.186       0.000
```

Box 7. Example results of genetic distances

4.12 Ordination analysis

Figure 14. Parameters of Ordination

Ordination analysis groups the individuals into a multi-dimensional space by their dissimilarity, reduce the dimension of data with the least loss of information, and eliminate the correlation between coordinates. POLYGENE provides two types of ordination analysis, principal coordinate analysis (PCoA) and principal component analysis (PCA).

PCoA is performed based on the calculated genetic distance matrices, while PCA is performed based on allele frequencies. The parameters are shown in Figure 14, including

4 Usage

the number of outputted coordinates the preview scatter plot.

POLYGENE is able to present the scatter plot as a preview by using the first two principal coordinates. An example result is shown in Figure 15, including the preview scatter plot, the principal coordinates and the corresponding eigenvalues. The scatter plot can be switched to different distance matrices by the combobox. The populations of different regions or the individuals in different populations are each represented by different color and markers, the marker style, marker size, marker character and identifier size can be configured in the toolbar of the scatter plot.
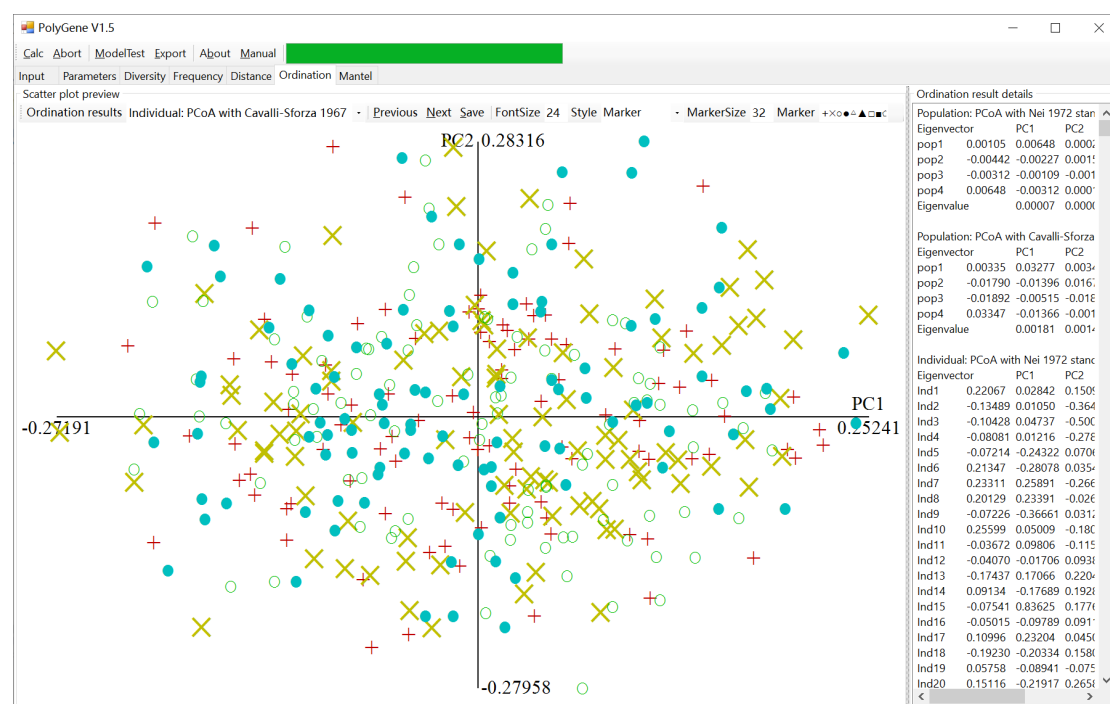
Figure 15. Example results of a PCoA

The text results are given in the right textbox, including the transformed coordinates in the new space, the eigen-vectors in PCA, the variances in each new axis, the percentage of variances explained by each new axis, and the cumulative percentage of variances explain by first several axes.

4.13 Hierarchical clustering

Figure 16. Parameters of hierarchical clustering

POLYGENE can also be used to perform the hierarchical clustering of the resulting matrices of genetic distance calculations. Various parameters can be found in Figure 16. Seven clustering methods can be used. The options '**Font size**' and '**Line skip**' are used to control the style of the resulting dendrogram.

An example results of hierarchical clustering is shown in Figure 17. The dendrogram can be manipulated by the combobox. The tree files are shown to the right, which can be pasted into the file `*.trees` and can be viewed by other software packages such as FIGTREE (http://tree.bio.ed.ac.uk/software/figtree/).

Figure 17. Example results of hierarchical clustering

4.14 Individual inbreeding coefficient

POLYGENE provides three methods-of-moment estimators (Loiselle *et al.* 1995; Ritland 1996a; Weir 1996) to estimate the coancestry coefficient (also known as the kinship coefficient). These estimators can be selected in the groupbox 'Individual inbreeding coef.' (Figure 18). After that, these estimated coancestry coefficients will be converted into the individual inbreeding coefficients. A jackknife method is used to estimate the standard error of inbreeding coefficient estimates.

Figure 18. Parameters of individual inbreeding coefficient

An example of the data of individual inbreeding coefficients is shown in Box 8. Because these methods-of-moment estimators are unbiased, the expected values of estimates of individuals within an outcrossing population is zero, and so negative inbreeding coefficients may occur.

```
Ind     Pop     Reg     Ploidy  Ritland 1996    Loiselle 1995   Weir 1996
Ind1    pop1    reg1    4       0.024           -0.074          -0.265
Ind2    pop1    reg1    4       -0.027          0.072           0.367
Ind3    pop1    reg1    4       -0.060          -0.057          0.141
Ind4    pop1    reg1    4       -0.181          -0.193          -0.265
Ind5    pop1    reg1    4       0.078           -0.047          -0.163
Ind6    pop1    reg1    4       -0.096          -0.091          -0.064
Ind7    pop1    reg1    4       -0.181          -0.169          -0.392
Ind8    pop1    reg1    4       -0.027          -0.009          0.145
Ind9    pop1    reg1    4       -0.037          -0.068          -0.064
Ind10   pop1    reg1    4       -0.181          -0.193          -0.265
Ind11   pop2    reg1    4       -0.020          0.009           0.147
Ind12   pop2    reg1    4       -0.194          -0.218          -0.176
Ind13   pop2    reg1    4       -0.044          -0.021          -0.046
Ind14   pop2    reg1    4       0.167           0.112           -0.046
Ind15   pop2    reg1    4       -0.021          -0.015          -0.089
Ind16   pop2    reg1    4       -0.060          -0.098          -0.298
Ind17   pop2    reg1    4       0.169           0.281           0.378
Ind18   pop2    reg1    4       -0.194          -0.218          -0.176
Ind19   pop2    reg1    4       -0.022          -0.051          -0.035
```

```
 Ind20   pop2    reg1    4       -0.152          -0.156          -0.035
```
Box 8. Example of the individual inbreeding coefficients

4.15 Individual heterozygosity-index

For a given individual, the heterozygosity-index (H-index) is defined as the probability of randomly sampling two distinct alleles within this individual without replacement. This H-index can be considered as the average heterozygosity of this individual at each locus or across all loci. There are no parameters for this function. An example of the individual H-indices is shown in Box 9.

```
Ind     Pop     Reg     Ploidy  Average  Loc1    Loc2
Ind1    pop1    reg1    4       0.833    0.833   0.833
Ind2    pop1    reg1    4       0.417    0.000   0.833
Ind3    pop1    reg1    4       0.566    0.564   0.568
Ind4    pop1    reg1    4       0.833    0.833   0.833
Ind5    pop1    reg1    4       0.766    0.532   1.000
Ind6    pop1    reg1    4       0.701    0.833   0.568
Ind7    pop1    reg1    4       0.917    1.000   0.833
Ind8    pop1    reg1    4       0.563    0.564   0.563
Ind9    pop1    reg1    4       0.701    0.833   0.568
Ind10   pop1    reg1    4       0.833    0.833   0.833
Ind11   pop2    reg1    4       0.571    0.571   0.571
Ind12   pop2    reg1    4       0.833    0.833   0.833
Ind13   pop2    reg1    4       0.695    0.833   0.558
Ind14   pop2    reg1    4       0.687    0.833   0.541
Ind15   pop2    reg1    4       0.786    0.571   1.000
Ind16   pop2    reg1    4       0.917    1.000   0.833
Ind17   pop2    reg1    4       0.285    0.571   0.000
Ind18   pop2    reg1    4       0.833    0.833   0.833
Ind19   pop2    reg1    4       0.702    0.833   0.571
Ind20   pop2    reg1    4       0.702    0.833   0.571
```
Box 9. Example of the individual H-indices
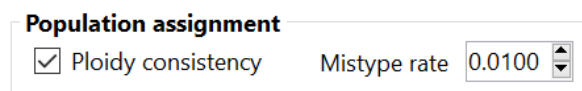
4.16 Population assignment

For the population assignment, the logarithm of the likelihood for each individual in a target population is calculated by using the allele frequencies of the respective population or region (Paetkau *et al.* 2004). Each individual is assigned to the population with the largest likelihood. Such assignment can help in identifying the natal population

4 Usage

of an individual.

The parameters for the population assignment are shown in Figure 19, where the first option 'Ploidy consistency' is used to exclude the populations with different ploidy levels from that of the target individual. The second option, 'Mistype rate', is used to configure the mistyping rate of genotypes/phenotypes. Such mistyping rate can avoid an individual being excluded from its true natal population due to genotyping errors (e.g., false allele, Taberlet *et al*. 1996).

Figure 19. Parameters of population assignment

An example results of population assignment is shown in Figure 20, where the first four columns are the sample information, the fifth column is the assigned populations, and the successive columns are either the value of the LOD score or the common logarithm of likelihood for an individual in each population or in each region. The value of the LOD score for an individual is the difference between the largest and the second largest common logarithms of the likelihoods for this individual in all populations or in all regions. Since there is only one region in this example, the value of the LOD score cannot be calculated, and so this is set as NaN (not a number).

Ind	Pop	Reg	Ploidy	Assiged po	LOD	Log10 L for each population			Assiged	LOD	Log10 L for each region	
						pop1	pop2				reg1	
Ind1	pop1	reg1	4	pop2	0.125	-2.041	-1.915		reg1	NaN	-1.908	
Ind2	pop1	reg1	4	pop1	1.076	-1.684	-2.76		reg1	NaN	-2.147	
Ind3	pop1	reg1	4	pop1	0.23	-1.167	-1.397		reg1	NaN	-1.256	
Ind4	pop1	reg1	4	pop2	0.053	-1.174	-1.12		reg1	NaN	-1.088	
Ind5	pop1	reg1	4	pop1	0.546	-2.719	-3.265		reg1	NaN	-2.927	
Ind6	pop1	reg1	4	pop2	0.267	-1.306	-1.039		reg1	NaN	-1.122	
Ind7	pop1	reg1	4	pop2	0.038	-1.913	-1.875		reg1	NaN	-1.835	
Ind8	pop1	reg1	4	pop1	0.704	-1.359	-2.063		reg1	NaN	-1.678	
Ind9	pop1	reg1	4	pop1	0.141	-1.562	-1.703		reg1	NaN	-1.607	
Ind10	pop1	reg1	4	pop2	0.053	-1.174	-1.12		reg1	NaN	-1.088	
Ind11	pop2	reg1	4	pop1	0.23	-1.167	-1.397		reg1	NaN	-1.256	
Ind12	pop2	reg1	4	pop2	0.053	-1.174	-1.12		reg1	NaN	-1.088	
Ind13	pop2	reg1	4	pop2	0.156	-1.911	-1.755		reg1	NaN	-1.776	
Ind14	pop2	reg1	4	pop2	1.136	-3.911	-2.774		reg1	NaN	-3.28	
Ind15	pop2	reg1	4	pop2	0.075	-2.422	-2.347		reg1	NaN	-2.317	
Ind16	pop2	reg1	4	pop2	0.11	-2.78	-2.67		reg1	NaN	-2.655	
Ind17	pop2	reg1	4	pop2	0.563	-3.07	-2.507		reg1	NaN	-2.739	
Ind18	pop2	reg1	4	pop2	0.053	-1.174	-1.12		reg1	NaN	-1.088	
Ind19	pop2	reg1	4	pop1	0.141	-1.562	-1.703		reg1	NaN	-1.607	
Ind20	pop2	reg1	4	pop2	0.267	-1.306	-1.039		reg1	NaN	-1.122	

Figure 20. Example results of population assignment

4.17 Spatial pattern analysis

Spatial pattern analysis is used for analyze isolation by distance in a continuous population (Hardy & Vekemans 1999). It calculates the autocorrelation coefficient (Moran's I) of individual allele frequency at different distance classes, which is equivalent to the average Moran's I relatedness coefficient (Hardy & Vekemans 1999) between individuals at different distance classes with the reference population being the total population.

Figure 21. Parameters of spatial pattern analysis

The distances between individuals are calculated from coordinates. The distance type can

be configured in the 'Haversine distance' checkbox (Figure 21). For cartesian coordinates, the Euclidean distance can be used. For latitude and longitude coordinates, the haversine distance (in unit kilometer) can be used. The distance intervals can be manually input in the 'Distance intervals' box with 'Manually input intervals' option selected in the '#distance classes' box, or automatically assigned into several classes with equal individual pairs by choosing the remaining digital options in the '#distance classes' box. The standard error of Moran's *I* can be estimated by checking 'Estimate SE by Jackknife method' box.

An example results is shown in Figure 22. The header row descriptions are as follows. Class denotes the order of distance class; Range denotes the range; #pairs denotes number of individual pairs, Mean(d) and Mean(ln d) is the mean of distance and log-distance, respectively; partic% denotes the proportion of all individuals represented, CV(#partic) denotes the coefficient of variation of the number of times each individual is represented, Morans I is the autocorrelation coefficient, SE is the standard error obtained by Jackknife method, P is the significance of the test with null hypothesis that there is not spatial autocorrelation within each distance class, i.e., the Morans *I* is equal to $1/(n-1)$.

Spatial pattern analysis for individuals								
Sample size n: 628								
Distance type: Euclidean								
Class	Range	#pairs	Mean(d)	Mean(ln d)	partic%	CV(#partic	Moran's I	SE	P
1	0~40.81	32187	15.47	1.91	100	0.17	0.05	0.02	0.05
2	40.81~577	31913	354.35	5.47	100	0.55	0.03	0.03	0.27
3	577.00~68	32211	620.87	6.38	74.68	0.35	-0.01	0.01	0.3
4	683.21~11	32128	944.28	6.79	71.82	0.6	-0.01	0.02	0.45
5	1118.13~1	31817	1372.47	7.14	84.87	0.6	-0.02	0.02	0.29
6	1571.22~9	32301	1718.39	7.37	71.34	0.43	-0.03	0.02	0.1

Figure 22. Example results of spatial pattern analysis

4.18 Relationship coefficient

The relationship coefficients include relatedness coefficient, coancestry coefficient, Moran's *I*. POLYGENE provides two polyploid relatedness estimators, three coancestry

4 Usage

coefficient estimators, and a Moran's *I* coefficient.

The first two estimators are the moment estimator devised by Huang *et al.* (2014) and the maximum-likelihood estimator devised by Huang *et al.* (2015a), and the three coancestry coefficient estimators are first published by Loiselle *et al.* (1995), Weir (1996) and Ritland (1996a), respectively. Moran's *I* coefficient is introduced into the analysis of spatial genetic structure by Hardy and Vekemans (1999).

For the first two estimators, the supported maximum level of ploidies is eight (in Huang *et al.* 2014, 2015a). For two individuals with different ploidy levels, their relatedness from the higher ploidy to the lower ploidy can also be calculated (see Huang *et al.* 2015b). The remaining estimators support a maximum level of ploidies of 10. For the three coancestry coefficient estimators, the estimates of Loiselle *et al.* (1995) and Ritland (1996a) can be converted into the relatedness coefficient (see Section 5.12 for details).

Figure 23. Parameters of relationship coefficient

The parameters of relationship coefficient estimation are shown in Figure 23, where the first three options are used to configure the range of individuals, so that the relatedness between two individuals within the selected range can be calculated. Because the estimation of relatedness depends on the allele frequencies in the reference population, the two estimates between the same pair of individuals in two different ranges are generally different. There are eight relationship coefficient estimators, which can be

selected in the latter options. The standard error of estimated relationship coefficient can be estimated with a Jackknife method by checking the 'Estimate SE by Jackknife method' box.

An example results of pairwise relatedness is shown in Box 10.

```
Huang et al. 2014 moment estimator
pop1   Ind1   Ind2   Ind3   Ind4   Ind5   Ind6   Ind7   Ind8   Ind9   Ind10
Ind1   0.824  0.006  -0.491 0.125  0.111  -0.724 0.125  0.390  -0.724 0.125
Ind1   0.006  1.000  0.100  -0.120 0.403  -0.015 -0.259 0.126  -0.005 -0.120
Ind1   -0.491 0.100  0.614  0.101  -0.128 0.528  -0.038 0.022  0.537  0.101
Ind1   0.125  -0.120 0.101  0.000  -0.220 0.105  0.000  0.115  0.105  0.000
Ind1   0.111  0.403  -0.128 -0.220 0.838  -0.362 -0.052 -0.083 0.016  -0.220
Ind1   -0.724 -0.015 0.528  0.105  -0.362 0.618  0.105  -0.279 0.618  0.105
Ind1   0.125  -0.259 -0.038 0.000  -0.052 0.105  0.000  -0.003 0.105  0.000
Ind1   0.390  0.126  0.022  0.115  -0.083 -0.279 -0.003 0.651  -0.271 0.115
Ind1   -0.724 -0.005 0.537  0.105  0.016  0.618  0.105  -0.271 0.618  0.105
Ind1   0.125  -0.120 0.101  0.000  -0.220 0.105  0.000  0.115  0.105  0.000
pop2   Ind11  Ind12  Ind13  Ind14  Ind15  Ind16  Ind17  Ind18  Ind19  Ind20
Ind11  0.602  0.134  -0.059 -0.341 -0.304 -0.358 -0.130 0.134  0.466  0.458
Ind11  0.134  0.000  0.180  -0.022 0.156  0.152  0.030  0.000  0.119  0.123
Ind11  -0.059 0.180  0.637  0.143  0.239  -0.422 -0.650 0.180  -0.128 -0.305
Ind11  -0.341 -0.022 0.143  0.711  -0.253 -0.272 -0.365 -0.022 -0.045 -0.194
Ind11  -0.304 0.156  0.239  -0.253 0.670  0.362  -0.297 0.156  -0.458 0.099
Ind11  -0.358 0.152  -0.422 -0.272 0.362  0.786  -0.073 0.152  -0.249 -0.597
Ind11  -0.130 0.030  -0.650 -0.365 -0.297 -0.073 0.676  0.030  -0.258 -0.010
Ind11  0.134  0.000  0.180  -0.022 0.156  0.152  0.030  0.000  0.119  0.123
Ind11  0.466  0.119  -0.128 -0.045 -0.458 -0.249 -0.258 0.119  0.654  0.462
Ind11  0.458  0.123  -0.305 -0.194 0.099  -0.597 -0.010 0.123  0.462  0.530
```
Box 10. Example results of pairwise relatedness

4.19 Heritability estimation

POLYGENE provides five heritability estimators (Mousseau *et al.* 1998; Ritland 1996b; Thomas *et al.* 2000), these can be configured in the parameter box in Figure 24. The principal of estimating heritability is comparing the similarity in quantitative trait among different class of relatives. If close related individuals are more similar in quantitative trait, then the heritability is higher.

4 Usage

Figure 24. Parameters of heritability estimation

Ritland (1996b) MOM and Huang et al. (unpub) MOM estimators are based on regression analysis for $Z_X Z_Y$ (product of standardized quantitative trait values) on the kinship coefficient $\hat{\theta}_{XY}$. Huang unpub applied correction for inbreeding and sampling correlations based on Nomura (2008). Therefore, a predefined proportion of individuals pairs are used as putative nonrelatives as in estimating effective population size. This can be configured in the 'p(nonrelative)' box.

Mousseau *et al.* (1998) ML, Thomas *et al.* (2000) ML and Huang et al. (unpub) ML are maximum-likelihood estimators, which use genetic marker data to identify the posterior probability of different relationships (e.g., full-sibs, half-sibs and nonrelatives). Mousseau *et al.* (1998) ML only consider two full-sibs and nonrelatives, while Thomas *et al.* (2000) ML consider all three relationships. Huang et al. (unpub) ML consider inbreeding and sample correlations based on Nomura (2008) and also requires a predefined proportion of putative nonrelatives.

An example results of pairwise relatedness is shown in Figure 25. Mean and SD are respectively the mean and standard deviation of the quantitative traits, follow by the estimated heritability and standard error of for selected estimators.

pop1	n	Mean	SD	Ritland1996 h2	SE	Mousseau1998 h2	SE	Thomas2000 h2	SE	HuangML h2	SE	HuangMOM h2	SE
Q1	250	100.037	0.968	0.039	2.484	-0.4	0.1	-0.03	0.114	-0.31	1.162	0.012	0.14
Q2	250	99.974	1.007	0.403	2.435	-0.39	0.096	0.066	0.069	-0.41	0.835	0.027	0.077
Q3	250	99.984	0.951	-2.66	2.271	-0.42	0.138	-0.15	0.153	1.355	2.137	-0.16	0.149
Q4	250	100.008	1.021	0.296	1.823	-0.38	0.155	-0.02	0.106	-0.01	0.673	-0.03	0.187

Figure 25. Example results of heritability estimation

*4.20 Q_{ST} estimation

4.21 Parentage analysis

There are three typical categories in a parentage analysis: (i) identifying the father when the mother is unknown; (ii) identifying the father when the mother is known, and (iii) identifying the father and the mother jointly. There are two methods to perform a parentage analysis: one is the exclusion method and the other is the likelihood method (Marshall *et al.* 1998).

The exclusion method is performed by excluding the alleged parent(s) based on mismatched genotypes or allelic phenotypes, but the true parent(s) may also be excluded due to mistyping, and some false parent(s) may not be excluded.

The likelihood method can resolve the problem of genotyping errors and find the optimal alleged parent even if some alleged parents cannot be excluded (Kalinowski *et al.* 2007). In this method, the two likelihoods for an alleged parent are calculated based on the two hypotheses that the alleged parent is the true parent and the alleged parent is not the true parent. Each alleged parent will be assigned an LOD score, which is the natural logarithm of the ratio of these two likelihoods. If an LOD score is positive, it means that the first hypothesis is more likely to be true than the second hypothesis. Marshall *et al.* (1998) provided a statistic Δ for resolving the paternity, which is the difference between the highest and the second highest LOD scores. Monte-Carlo simulations are subsequently used to assess the confidence level of each value of Δ.

4 Usage
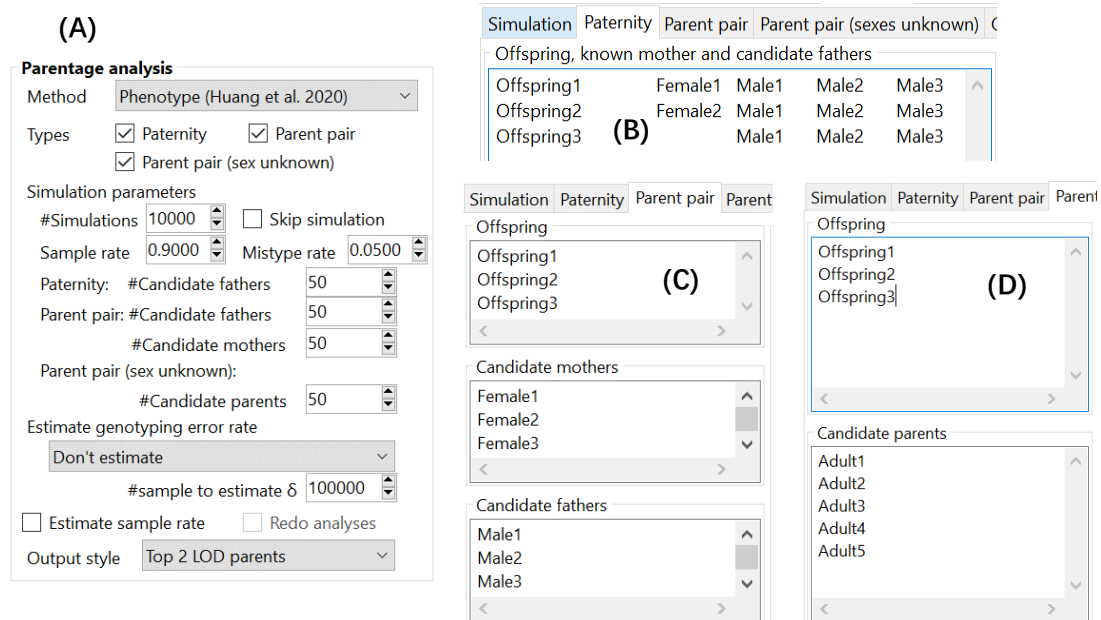
Figure 26. (A) Parameters of parentage analysis; (B) offspring, known mother and candidate fathers of paternity analysis; (C) offspring, candidate mothers and candidate fathers of parent pair analysis; (D) offspring and candidate parents of parent pair analysis with unknown sexes.

The parameters of a parentage analysis are shown in Figure 26 (A), where the methods consist of allelic phenotype, exclusion and dominant, the allelic phenotype method used to calculate the LOD and the likelihood based on allelic phenotypic frequency and transitional probabilities; the exclusion method uses negative mismatch as the LOD, which does not use simulation and the confidence level is unavailable; the dominant method converts the data into pseudo-dominant marker data, where each visible allele defines a dominant locus, and then uses a diploid equation to calculate the likelihoods and the LOD.

The 'Skip simulation' can help to skip the Monte-Carlo simulation and uses a previously calculated or manually inputted values as the critical thresholds of Δ. This can save time by preventing repeated calculations. 'Paternity' can help to identify the true father from several candidate fathers when the mother is either known or unknown; 'Parent pair' can help to identify the father and the mother jointly; 'Parent pair (sex unknown)' is similar

to the 'Parent pair', but the sexes of parents are unknown; and 'Allow selfing when sexes are unknown' denotes whether selfing is considered. The subsequent options are the simulation parameters, including the sampling rate, genotyping rate, mistype rate and the numbers of offspring and candidate parent(s) in a simulation. 'Output style' is used to control the number of candidate parent(s) with larger LOD scores in the output results.

POLYGENE is able to perform a parentage analysis in the presence of self-fertilization. This can be enabled by checking the 'Consider selfing' in the allele frequency estimation box. The selfing rate will be estimated from the input data, and the estimated selfing rate will be used in the parentage analyses. In order to accommodate self-fertilization, the procedure of a parentage analysis needs to be modified as follows: (i) the offspring will be produced by selfing at a probability of s during simulation; (ii) the likelihood equation will incorporate selfing (see Section 5.13 for details); (iii) the genotypic frequency will use the selfing version.

Besides, POLYGENE is able to estimate the genotyping error rate and the sample rate. These two analyses are performed only for the allelic phenotype method. The genotyping error rate is estimated from any mismatches in reference pairs or trios. The reference pairs and trios are extracted from the known parents and the identified parents at a selected confidence level. The confidence level and number of sampling pairs or trios to estimate δ (some exclusion rates) can be configured (Figure 26A).

The sample rate is estimated from the assignment rate. The assignment rates in case of the true parent (or parent-pair) being sampled or not can be obtained from the Monte-Carlo simulation, and the observed assignment rate is a weighted average among these two assignment rates. The 'Redo simulations and analyses' will reperform the simulations and analyses with the newly calculated genotyping error rate (if estimated), sample rate, selfing rate in this application as *a priori*, then use the new assignment rates to estimate the sample rate.

4 Usage

The information of offspring and parent follows the format of CERVUS V3.0 (Kalinowski *et al.* 2007). The individual information for a paternity analysis is shown in Figure 26 (B), where each row denotes a case; the first and second columns denote the offspring and the known mother, respectively (if the mother is unknown or untyped, the second column remains blank). Beginning with the third column, each denotes an alleged father. The individual information of a parent pair analysis is shown in Figure 26 (C), where there are three textboxes used to identify the offspring, the candidate mothers and the candidate fathers. For each textbox, each row identifies an individual. Furthermore, when the sexes are unknown, all candidate parents are set in the same textbox (see Figure 26 (D)).
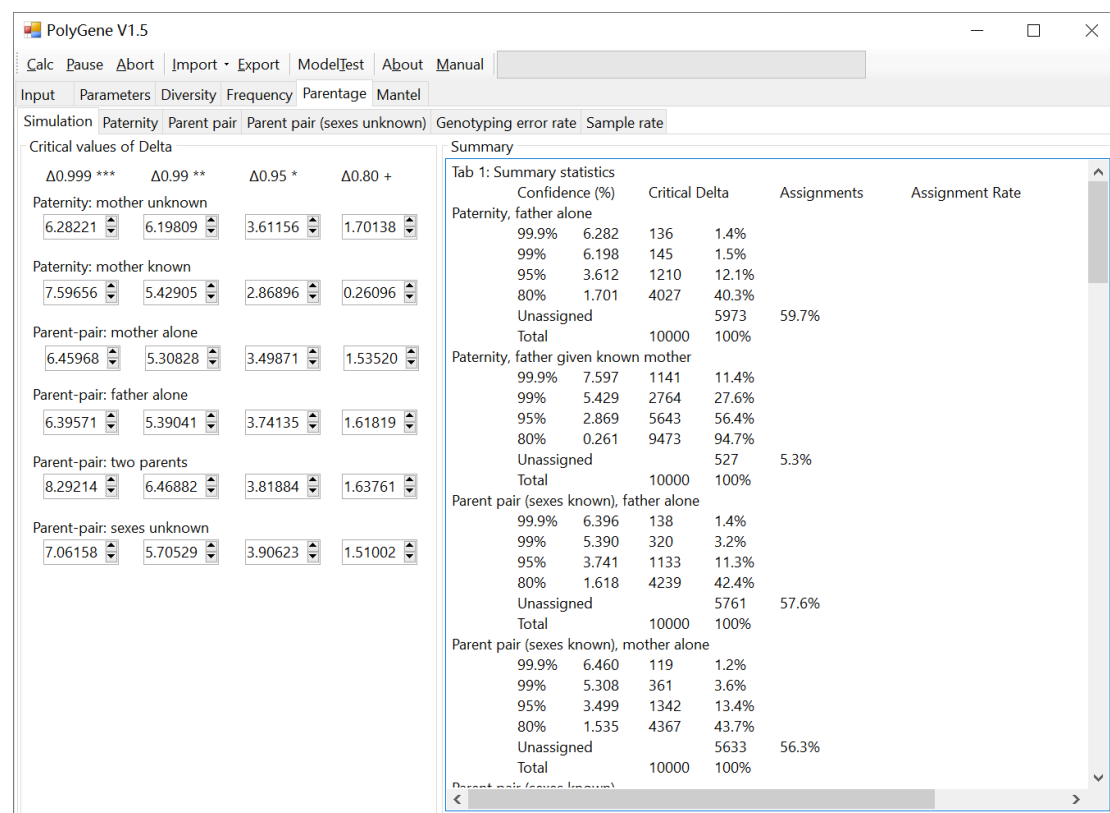
Figure 27. Example of outputting the simulation results of a parentage analysis

An example of the simulation results of a parentage analysis is shown in Figure 27, where the critical thresholds of Δ are in the left panel, and the summary of results is in

4 Usage

the right textbox.

The output style also follows the format of CERVUS V3.0 (Kalinowski *et al.* 2007), see Figure 28 for details).

Offsprin	Loci typ	Mother	Loci typ	Pair loc	Pair loc	Pair LOI	Candid:	Loci typ	Pair loc	Pair loc	Pair LOI	Pair De	Pair cor	Trio loc	Trio loc	Trio LO	Trio De	Trio confidence
Ind8	2	Ind2	2	2	0	0.783	Ind3	2	2	0	0.169	0.026	-		2	0	0.135	0.135	-
Ind8	2	Ind2	2	2	0	0.783	Ind4	2	2	0	0.143	0			2	0	-0.02	0	
Ind9	2	Ind2	2	2	0	-0.57	Ind7	2	2	0	0.352	-0.33			2	0	0.857	0.761	-
Ind9	2	Ind2	2	2	0	-0.57	Ind3	2	2	0	0.677	0			2	1	0.095	0	
Ind10	2	Ind2	2	2	0	-0.11	Ind4	2	2	0	0.673	0.513	-		2	0	0.446	0.119	-
Ind10	2	Ind2	2	2	0	-0.11	Ind7	2	2	0	0.16	0			2	0	0.326	0	
Ind11	2	Ind2	2	2	0	0.456	Ind3	2	2	0	1.359	0.716	-		2	0	1.048	0.549	-
Ind11	2	Ind2	2	2	0	0.456	Ind6	2	2	0	0.642	0			2	0	0.499	0	
Ind12	2	Ind2	2	2	0	-0.11	Ind4	2	2	0	0.673	0.513	-		2	0	0.446	0.119	-
Ind12	2	Ind2	2	2	0	-0.11	Ind7	2	2	0	0.16	0			2	0	0.326	0	
Ind13	2	Ind2	2	2	0	-0.14	Ind4	2	2	0	0.644	0.513	-		2	0	0.505	0.119	-
Ind13	2	Ind2	2	2	0	-0.14	Ind7	2	2	0	0.131	0			2	0	0.386	0	
Ind14	2	Ind2	2	2	1	-3.53	Ind5	2	2	0	-1.04	0			2	1	-0.27	0	
Ind14	2	Ind2	2	2	1	-3.53	Ind7	2	2	0	-0.35	0			2	2	-0.55	0	

Figure 28. Example results of parentage analysis

For a paternity analysis, the meanings of various columns are described as follows:

1 `Offspring ID`
   Offspring identifier
2 `Loci typed`
   Number of loci typed for each offspring
3 `Mother ID`
   Known mother identifier
4 `Loci typed`
   Number of loci typed for the known mother
5 `Pair loci compared`
   Number of loci typed for both the offspring and the known mother
6 `Pair loci mismatching`
   Identifying the mismatches between the offspring and the known mother
7 `Pair LOD score`
   LOD score of the known mother (without the candidate father's genotypes)
8 `Candidate father ID`
   Candidate father identifier
9 `Loci typed`
   Number of loci typed for the candidate father
10 `Pair loci compared`
   Number of loci typed for both the offspring and the candidate father
11 `Pair loci mismatching`
   Identifying the mismatches between the offspring and the candidate father
12 `Pair LOD score`
   LOD score of the candidate father (without the known mother's genotypes)

13 `Pair Delta`

Delta of the most-likely father (without the known mother's genotypes)

14 `Pair confidence`

Confidence of the most-likely father (without the known mother's genotypes)

15 `Trio loci compared`

Total number of loci typed in all three individuals

16 `Trio loci mismatching`

Mismatches among the offspring, the known parent and the candidate parent

17 `Trio LOD score`

LOD score of the candidate father (considering the known mother's genotypes)

18 `Trio Delta`

Delta of the most-likely father (considering the known mother's genotypes)

19 `Trio confidence`

Confidence of the most-likely father (considering the known mother's genotypes)

The results of the exclusion method can be found in the three columns in Figure 24: 'Loci typed', 'Loci compared' and 'Loci mismatching'.

For a parent pair analysis, the meanings of various columns are described as follows:

1 `Offspring ID`

Offspring identifier

2 `Loci typed`

Number of loci typed for each offspring

3 `Candidate mother ID`

Candidate mother identifier

4 `Loci typed`

Number of loci typed in the candidate mother

5 `Pair loci compared`

Number of loci typed in both the offspring and the candidate mother

6 `Pair loci mismatching`

Identifying the mismatches between the offspring and the candidate mother

7 `Pair LOD score`

LOD score of the candidate mother (without the candidate father's genotypes)

8 `Pair Delta`

Delta of the most-likely mother (without the candidate father's genotypes)

9 `Pair confidence`

Confidence of the most-likely mother (without the candidate father's genotypes)

10 `Candidate father ID`

Candidate father identifier

11 `Loci typed`

Number of loci typed in the candidate father

12 `Pair loci compared`

4 Usage

Number of loci typed in both the offspring and the candidate father

13 Pair loci mismatching

Identifying the mismatches between the offspring and the candidate father

14 Pair LOD score

LOD score of the candidate father (without the candidate mother's genotypes)

15 Pair Delta

Delta of the most-likely father (without the candidate mother's genotypes)

16 Pair confidence

Confidence of the most-likely mother (without the candidate father's genotypes)

17 Trio loci compared

Number of loci typed in the offspring, the candidate mother and the candidate father

18 Trio loci mismatching

Identifying the mismatches among the offspring, the candidate mother and the candidate father

19 Trio LOD score

LOD score of the candidate parents

20 Trio Delta

Delta of the most-likely candidate parents

21 Trio confidence

Confidence of the most-likely parents

For a parent pair (sex unknown) analysis, the headings of the columns are as follows:

1 Offspring ID

Offspring identifier

2 Loci typed

Number of loci typed in the offspring

3 First candidate ID

First candidate parent identifier

4 Loci typed

Number of loci typed in the first candidate parent

5 Pair loci compared

Number of loci typed in both the offspring and the first candidate parent

6 Pair loci mismatching

Identifying the mismatches between the offspring and the first candidate parent

7 Pair LOD score

LOD score of the first candidate parent (without the second candidate parent's genotypes)

8 Second candidate ID

Second candidate parent identifier

9 Loci typed

Number of loci typed in the second candidate parent

10 Pair loci compared

4 Usage

Number of loci typed in both the offspring and the second candidate parent

11 Pair loci mismatching

Identifying the mismatches between the offspring and the second candidate parent

12 Pair LOD score

LOD score of the second candidate parent (without the first candidate parent's genotypes)

13 Trio loci compared

Number of loci typed in the offspring and the two candidate parents

14 Trio loci mismatching

Identifying the mismatches among the offspring and the two candidate parents

15 Trio LOD score

LOD score of the candidate parents

16 Trio Delta

Delta of the most-likely candidate parents

17 Trio confidence

Confidence of the most-likely parents

An example of the genotyping error rate estimation is shown in Figure 29. For each application, first comes a table that lists the locus identifier, number of pairs or trios, γ (frequency of mismatch in reference pairs or trios), δ (some exclusion rates or coefficient), genotyping error rate e, E $(= 1 - (1 - e)^2)$ and the weight.

Left window (PolyGene V1.5 — Genotyping error rate tab):

Result

Estimate error rate by pair mismatch in paternity analyse (father only)
Estimating genotyping error rate from pair mismatch

Locus	n_pair	gamma	delta	E	weight
CA190	286	0.049	0.084	0.583	24.001
CA344	286	0.000	0.050	0.000	14.286
CA787	286	0.000	0.045	0.000	12.819
CA855	286	0.007	0.234	0.030	67.053

Number of pairs sampled to estimate delta: 100000
Number of reference pairs: 286
Multi-locus E: 0.053
Estimated genotyping error rate: 0.027
Variance of genotyping error rate: e / 1137.067
Weight of genotyping error rate: 1137.067

Estimate error rate by trio mismatch in paternity analyse (mother known)
Estimating genotyping error rate from trio mismatch

Locus	n_trio	gamma	delta	deltao	deltas	deltap	e	weight
CA190	143	0.098	1.051	0.350	0.530	0.350	0.104	150.270
CA344	143	0.000	1.051	0.350	0.576	0.350	0.000	150.223
CA787	143	0.000	1.241	0.414	0.605	0.414	0.000	177.529
CA855	143	0.021	2.240	0.747	0.911	0.747	0.009	320.364

Number of trios sampled to estimate delta: 100000
Number of reference pairs: 288
Number of reference trios: 143
Selfing rate in the reference trios: 0 / 143
Estimated genotyping error rate: 0.016
Variance of genotyping error rate: e / 3130.928
Weight of genotyping error rate: 3130.928

Estimated genotyping error rate in all applications: 0.017
Weight of genotyping error rate in all applications: 7882.290

Right window (PolyGene V1.5 — Sample rate tab):

Result

Simulations and analyses are reperofrmed!
Prior genotyping error rate: 0.017
Prior sample rate: 0.955
Prior selfing rate: 0.000

Estimate sample rate in paternity analyse (father only)

Confidence Level	ac	au	na	nc	ps	weight
99%	0.939	0.194	257	286	0.946	1703.918
95%	1.000	0.976	286	286	1.004	45.845
80%	1.000	1.000	286	286	0.500	0.000

Weighted average of sample rate: 0.947
Weight: 1749.763

Estimate sample rate in parent-pair analyse (sexes unknown)

Confidence Level	ns	nc
99%	0	62
95%	0	82
80%	0	83

su: 0.000

Confidence Level	ac	au	na	nc	ps	weight
99%	0.991	0.106	62	143	0.608	667.490
95%	1.000	0.561	82	143	0.167	40.159
80%	1.000	0.999	83	143	0.500	0.000

Weighted average of sample rate: 0.583
Weight: 707.649

Estimate sample rate in all applications

Confidence Level	ps	weight
99%	0.906	5995.250
95%	0.613	86.004
80%	0.500	0.000

Weighted average of sample rate: 0.901
Weight: 6081.254

Figure 29. Example results of genotyping error rate and sample rate estimation

The following are some statistics (Figure 29). After the results of all applications is shown, the weighted average across all applications is shown.

An example of the sample rate estimation is shown in Figure 29. For each application, first comes a table that lists a_c (assignment rate when the true parent or parent-pair is sampled), a_u (assignment rate when the true parent or parent-pair is not sampled), n_a (number of cases assign a parent or a parent-pair), n_c (number of cases), p_s (estimated sample rate), and the weight. The following is the estimated sample rate and the weight. Similarly, this weight can be used to obtain the weighted average of p_s across applications.

For parent-pair analysis with unknown sexes, an additional table lists the estimation of the selfing rate is provided to show n_s (number assigned selfed cases) and n_c (number of assigned cases) and an additional estimated selfing rate in this application s_u is shown. Finally, a table shows the weighted average of p_s across applications at each confidence level and all confidence levels. After the results of all applications is shown, the weighted average across all applications is shown.

4.22 Analysis of molecular variance

The analysis of molecular variance (AMOVA) can be used to perform the hierarchical partitioning of genetic variation among populations or regions, and to estimate the *F*-statistics at multiple levels. The procedure of AMOVA follows the methods of Excoffier *et al.* (1992) and Weir & Cockerham (1984), but some modifications are made for polysomic inheritance and the definition of a multi-level region.

The parameters used for AMOVA are shown in Figure 30, where the genetic distance model can be either IAM or SMM.

44

Figure 30. Parameters of AMOVA

The option 'Ignore ind. level' is used to configure whether or not to ignore the hierarchy level of 'Within individual'. If this option is checked, the two levels 'Among individual' and 'Within individual' will be merged into the single level 'Within population'. If there is only one region, the level of this region will also be ignored. The final number box is used to control the number of permutations so as to test the differentiation.

POLYGENE supports four methods of AMOVA mentioned below. For the **homoploid method**, the calculation speed is most rapid and the results are unbiased for perfect genotypic data but will have bias in the presence of missing data or for allelic phenotype data. Both missing data and allelic phenotype data can introduce bias, while the first bias can be eliminated by checking **p(type) corr**. For the **anisoploid method**, the calculation speed is slower than for the homoploid method, but the results are unbiased for missing data while biased for the allelic phenotype data. For the **weight genotype method,** the calculation speed is relatively slow and the results are least affected by allelic phenotype data. These three methods are methods-of-moment, whose estimates of variance components or F-statistics can be negative when the sample size is small and/or the differentiation is low. The **maximum-likelihood method** can avoid this problem and ensure the estimates fall within a biological meaningful range. However, this method has the longest computing time and can only be applied to small datasets. More detailed descriptions can be found in Section 5.14.

An example results of AMOVA is shown in Figure 31. In this figure, for the first table, the

first line shows the random number generator seed, followed by a summary of the AMOVA, which shows the degrees of freedom (d.f.), sum of squares (SS), mean squares (MS), variance (Var), percentage of variance explained (%), permuted mean, permuted variance and right-tailed P value of each variance component at each hierarchy level. The second table is the permutation test results of the F-statistics together with the P values. Finally, the sums of squares (SS) within each population or within each individual are shown in the remaining two tables.

	A	B	C	D	E	F	G	H	I
1	Summary AMOVA Table (IAM), model: homoploid								
2	Source	d.f.	SS	MS	Var	%	PermMean	PermVar	P(rand>obs)
3	Within Ind	200	61.018	0.305	0.305	74.872	0.339	0.000	1.000
4	Among Ind	98	23.681	0.242	-0.021	-5.190	0.000	0.000	0.979
5	Among Po	1	16.721	16.721	0.124	30.319	0.000	0.000	0.000
6	Total	299	101.419	0.339	0.407	100.000	-	-	-
7									
8	F-statistic	Value	PermMea	PermVar	P(rand>obs)				
9	FIS	-0.074	-0.001	0.001	0.992				
10	FIT	0.251	-0.001	0.001	0.000				
11	FST	0.303	0.000	0.000	0.000				
12									
13	SS within each population								
14	ID	#Hap	IAM						
15	pop1	100	17.270						
16	pop2	200	67.428						
17									
18	SS within each individual								
19	ID	#Hap	IAM						
20	Ind1	2	0.000						
21	Ind2	2	0.000						
22	Ind3	2	0.500						
23	Ind4	2	0.000						
24	Ind5	2	0.000						
25	Ind6	2	0.000						
26	Ind7	2	0.000						

Figure 31. Example results of AMOVA

4.23 Bayesian clustering

The Bayesian clustering uses a Markov chain Monte Carlo (MCMC) algorithm to cluster individuals into clusters, which can be used to infer the individual ancestry and the population genetic structure. It follows the software named 'STRUCTURE' (Pritchard *et al*. 2000), whose parameters are shown in Figure 32. The three checkboxes at the top of this figure are used to configure three models, which are the ADMIXTURE model (Pritchard *et*

4 Usage

al. 2000), the LOCPRIORI model (Hubisz *et al*. 2009) and the F model (Falush *et al*. 2003).

In the non-ADMIXTURE model, each individual is considered as complete, which can only originate from one cluster. However, in the ADMIXTURE model, each allele copy within an individual and at a locus can originate from different clusters. In the LOCPRIORI model, the *a priori* information of the sample group (population sampled) is used to help the clustering. Moreover, in the F model, it is assumed that the allele frequencies between each cluster and its corresponding ancestral cluster will be correlated. In addition, null alleles are considered for all three models, and missing data are weighted to generate the dummy haplotypes according to the allele frequencies in the population.

Figure 32. Parameters of Bayesian clustering

For the subsequent options, 'K runs from X to Y' is used to configure the range of numbers (K) of the calculated putative clusters. The MCMC parameters are '#Burnin' (the length of burnin period), '#Rep' (the length of record iteration), '#Thinning' (the number of thinning intervals) and '#Run' (the number of independent runs), respectively.

In the part 'Allele frequency & admixture burnin', λ is a Dirichlet parameter to update the allele frequencies in each cluster, whose value can also be updated if the option 'Infer λ' is checked. The two options 'std(λ)' and 'max(λ)' are used to configure the behavior of the updating procedure of λ. The option 'Diff λ' enables this algorithm to use different values of λ for different clusters. The option 'Admburnin'

4 Usage

is used in both the non-ADMIXTURE and the non-LOCPRIORI models, which is able to generate a proper initial state so as to prevent the Markov chain from becoming blocked at the local maxima.

In the part 'ADMIXTURE model parameters', α is a Dirichlet parameter used to update the admixture proportions of individuals, the initial value of which is '$\alpha 0$'. When the option 'Infer α' is checked, the value of α will be updated according to the values of 'std(α)' and 'max(α)'. The *a priori* distribution of α is either a uniform distribution or a gamma distribution, which is used to evaluate the new value of α. If the parameters 'α prior A' and 'α prior B' are checked, the *a priori* distribution of α is switched to a gamma distribution. The option 'Diff α' enables this algorithm to use different values of α for different clusters. The option 'MetroFreq' is used to control the Metropolis-Hastings update frequency of admixture proportions.

In the part 'LOCPRIORI model', a weak population structure can be inferred with the assistance of sample group information. The parameter r is used to estimate the informativeness of data for sampling location, and the parameters η and γ are used in the non-ADMIXTURE model, where η reflects the relative proportion of individuals assigned to a cluster and γ reflects the relative proportion of individuals sampled from a location and assigned to a cluster. They are updated by drawing a new value from the uniform distribution:

$$U\big(\eta - \text{eps}(\eta), \eta + \text{eps}(\eta)\big) \quad \text{or} \quad U\big(\gamma - \text{eps}(\gamma), \gamma + \text{eps}(\gamma)\big),$$

in which eps($\cdot$) denotes the maximum value of allowable changes when $\cdot$ is updated. For the ADMIXTURE model, LOCPRIORI will help to add one Dirichlet parameter, called the local α, and the original parameter α is called the global α to distinguish this from the local α. The global α and local α are updated by drawing two new values from two normal distributions, respectively.

In the part 'F model parameters', each allele frequency is correlated with its

4 Usage

corresponding ancestral cluster, and the parameter F is evaluated by the correlation, where F is analogous to Wright's F_{ST}. In the update of F, the new value F' is drawn from the normal distribution $N(F, \text{std}^2(F))$, and is evaluated according to its *a priori* gamma distribution by 'priori mean' and 'priori std'. The allele frequency of an ancestral cluster is also updated by one of two approaches: (i) an independent approach and (ii) the Metropolis-Hastings approach.
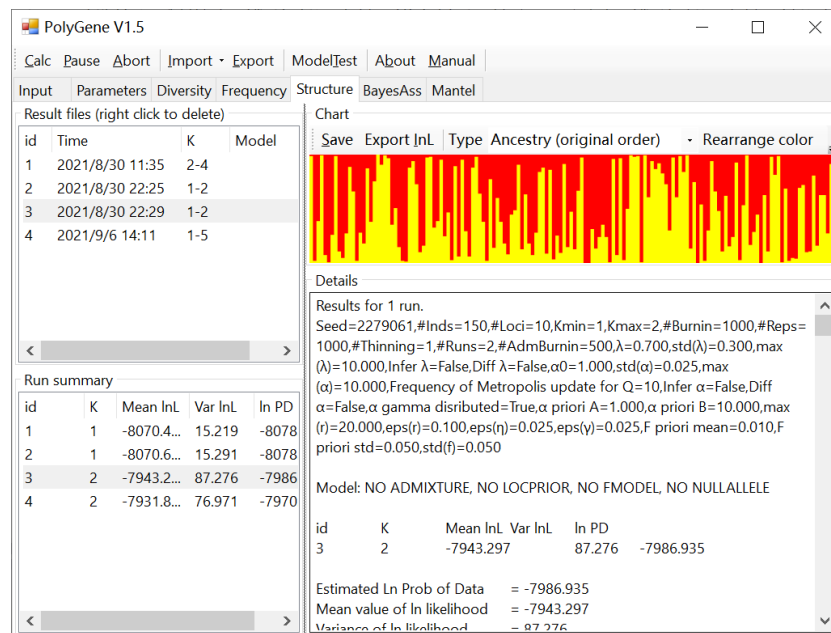
Figure 33. Example results for Bayesian clustering

The result files are shown at the top-right 'Structure' page (Figure 33). If the user wants to delete a result file, right-click the corresponding item, and then click the option 'Delete'. By selecting a result file, the summary statistics for all runs will be shown in the bottom-right box. An example result details are shown in Box 11.

```
...
id      K       Mean lnL     Var lnL      ln PD
1       2       -7607.087    300.666      -7757.420
2       2       -7607.194    299.706      -7757.047
3       3       -7358.984    367.177      -7542.573
4       3       -7359.195    383.276      -7550.833
5       4       -7112.708    366.960      -7296.188
```

Box 11. Example results of run summy

4 Usage

By selecting one or multiple runs on the bottom-left, then the run details will be displayed on the right-side (see Box 12 for details). If multiple runs are selected and they have the same K, the average statistics will be calculated, e.g., the likelihood, allele frequency, ancestry, differentiation coefficient. If there are multiple K in the selected runs, then only the likelihood can be calculated. The cluster will be rearranged to ensure ancestry among different runs are the same. The figure will also be generated and shown in the top-right. The bar plot shows the averaged ancestry or the likelihood line chart (to check convergency) shows the likelihood for each run (Figure 34). The figure can be saved as an image file by click the 'Save' button, where the file format can be selected as `jpg`, `tiff`, `gif`, `bmp`, `png`, `emf` and `wmf`, where `emf` and `wmf` are vector images and can be converted into `eps, ai` or `pdf` with additional software such as Adobe Illustrator. The likelihoods of selected runs can be exported as plain text by clicking the 'Export lnL' button, and the exported text is shown in Figure 34.

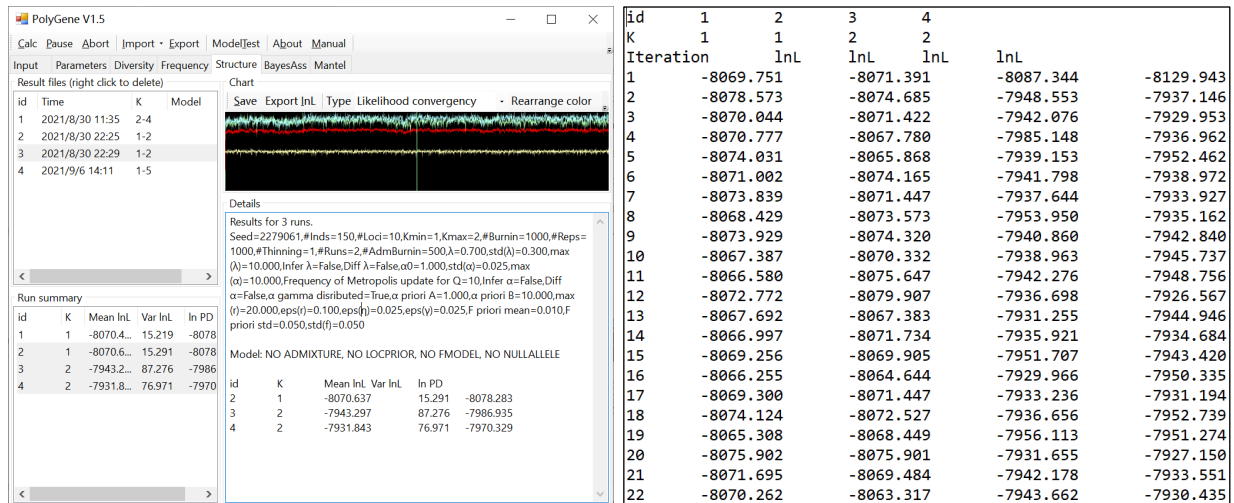Figure 34. The likelihood line chart and exported likelihood text file. For the line chart, the likelihood of each run is plot with a different color and the averaged likelihood is shown in red, the green vertical line separates the burnin and sampling periods.

The output data follow the format of STRUCTURE, including the Ln P(D), the mean and the variance of the Ln likelihoods, as well as the estimated parameters used for the

4 Usage

corresponding model. For the ADMIXTURE model, there is an extra parameter α. For the LOCPRIORI model, there are three extra parameters r, η and γ. For the LOCPRIORI + ADMIXTURE model, there are three extra parameters r, global α and local α. For the F model, the value of F (also known the value of global F) for each cluster is given in the output results. The next stage of the output consists of five tables (see Box 12): (i) proportion of membership of each pre-defined population in a cluster; (ii) allele-frequency divergence among populations (Net nucleotide distance); (iii) expected heterozygosity between individuals within the same cluster; (iv) inferred ancestry of individuals; and (v) estimated allele frequencies in each cluster.

```
Seed=62814158,#Inds=200,#Loci=8,Kmin=2,Kmax=5,#Burnin=5000,#Reps=10000,
#Thinning=1,#Runs=2,#AdmBurnin=500,λ=1.000,std(λ)=0.300,max(λ)=1.000,In
fer λ=False,Diff λ=False,α0=1.000,std(α)=0.025,max(α)=10.000,Frequency
of Metropolis update for Q=10,Infer α=True,Diff α=False,α gamma
disributed=False,α priori A=0.025,α priori
B=0.001,max(r)=20.000,eps(r)=0.100,eps(η)=0.025,eps(γ)=0.025,F priori
mean=0.010,F priori std=0.050, std(f)=0.050

Model: NO ADMIXTURE, NO LOCPRIOR, NO FMODEL, NO NULLALLELE
Estimated Ln Prob of Data    = -7300.835
Mean value of ln likelihood = -7113.007
Variance of ln likelihood   = 375.656

Proportion of membership of each pre-defined population in each of the
4 clusters
       Cluster
Pop    1     2     3     4
pop1   0.002 0.007 0.964  0.027
pop2   0.984 0.011 0.005  0.000
pop3   0.001 0.998 0.000  0.001
pop4   0.002 0.024 0.014  0.960

Allele-frequency divergence among populations (Net nucleotide distance)
       Cluster
Cluster      1     2     3     4
1      0.000 0.131 0.023  0.072
2      0.131 0.000 0.046  0.094
3      0.023 0.046 0.000  0.050
4      0.072 0.094 0.050  0.000

Expected heterozygosity between individuals within the same cluster
Cluster     He    Loc1  Loc2  Loc3  Loc4  Loc5  Loc6
1      0.687 0.687 0.687 0.687 0.687 0.687 0.687
2      0.633 0.633 0.633 0.633 0.633 0.633 0.633
3      0.736 0.736 0.736 0.736 0.736 0.736 0.736
4      0.662 0.662 0.662 0.662 0.662 0.662 0.662

Inferred ancestry of individuals
```

```
          Cluster
Individual  Population   1       2       3       4
Ind1   pop1   pop1   0.000   0.000   1.000   0.000
Ind2   pop1   pop1   0.000   0.000   0.746   0.254
Ind3   pop1   pop1   0.000   0.000   1.000   0.000
Ind4   pop1   pop1   0.000   0.000   1.000   0.000
Ind5   pop1   pop1   0.000   0.018   0.982   0.000
Ind6   pop1   pop1   0.002   0.010   0.987   0.001
Ind7   pop1   pop1   0.015   0.015   0.945   0.025
Ind8   pop1   pop1   0.000   0.000   1.000   0.000
Ind9   pop1   pop1   0.000   0.003   0.996   0.002
Ind10  pop1   pop1   0.000   0.003   0.996   0.001
...


Estimated allele frequencies in each cluster
Locus: Loc1, 4 alleles
      Cluster
Allele 1      2       3       4
1      0.171  0.071   0.158   0.137
2      0.457  0.144   0.314   0.189
3      0.242  0.263   0.240   0.505
4      0.130  0.522   0.287   0.169
...
```

Box 12. Example results of a run for Bayesian clustering

4.24 Migration rate estimation

The migration rate estimation is performed with MCMC follows the software BayesAss (Wilson & Rannala 2003), whose parameters are shown in Figure 35.

Figure 35. Parameters of migration rate estimation

There are five methods to perform migration rate estimation (Figure 32), where 'Dummy Genotype' denotes the dummy diploid genotypes are used to obtain the likelihood, 'Genotype' denotes the polyploid genotypes are used to obtain the likelihood, 'Phenotype' denotes the allelic phenotype are used to obtained the likelihood. 'Variable'

4 Usage

denotes the genotypes are updated during iteration, while 'Fixed' do not update the genotypes. The dummy genotype methods are faster and less accurate, the genotype/phenotype methods are slower.

The MCMC parameters are '#Burnin' (the length of burnin period), '#Rep' (the length of record iteration), '#Thinning' (the number of thinning intervals) and '#Run' (the number of independent runs), respectively. The updating parameters are 'deltaA' (Mixing parameter for allele frequencies), 'deltaF' (Mixing parameter for inbreeding coefficients), 'deltaM' (Mixing parameter for migration rates). 'Fix likelihood to 1' denotes fixing likelihood to 1 and generate priors.

An example results of migration rate estimation is shown in Figure 36. As well as Bayesian clustering, the result files are listed in the top-left and can be deleted by right-click and click 'Delete'. By selecting a results file, the summary for all runs will be shown in the bottom-right.
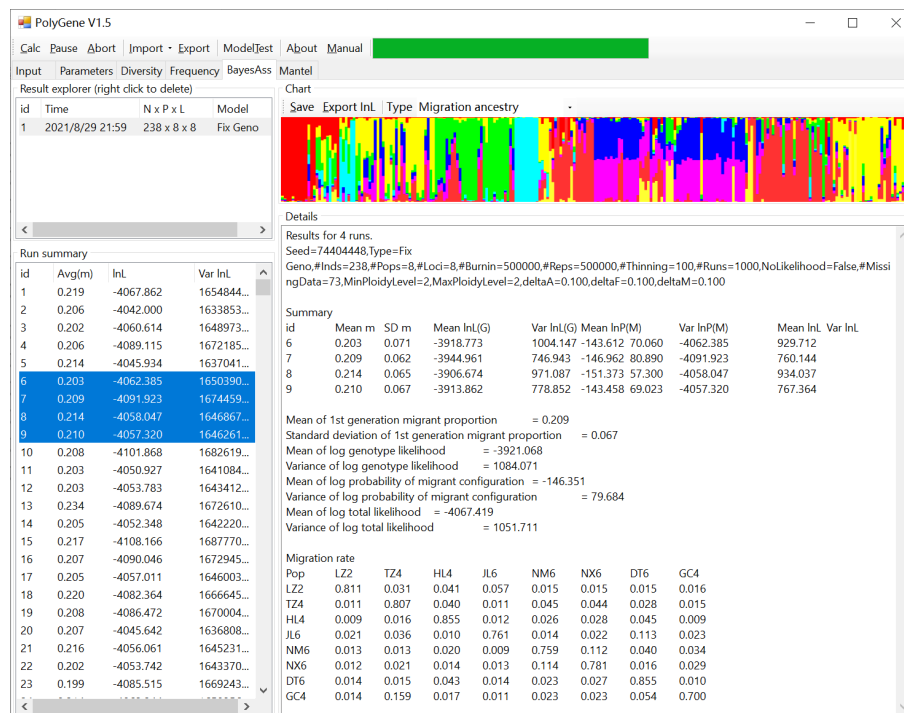

Figure 36. Example results of migration rate estimation

By selecting one or multiple runs in the bottom-left, the results for the selected runs will

be averaged (including migration rate, inbreeding coefficient, allele frequency, individual

ancestry). The simplified results are shown in the bottom-right (Box 13), the ancestry box

plot and the likelihood line chart will also be shown (Figure 36). The detailed results

including the migration rate, individual ancestry, inbreeding coefficient and allele

frequency can be exported by clicking the 'Export Details' button.

```
Results for 1 run.
Seed=74404448,Type=Fix
Geno,#Inds=238,#Pops=8,#Loci=8,#Burnin=500000,#Reps=500000,#Thinning=10
0,#Runs=1000,NoLikelihood=False,#MissingData=73,MinPloidyLevel=2,MaxPlo
idyLevel=2,deltaA=0.100,deltaF=0.100,deltaM=0.100

Summary
id Mean m SD m    Mean lnL(G)    Var lnL(G) Mean lnP(M)    Var lnP(M)
   Mean lnL    Var lnL
2   0.206   0.069   -3895.082   818.030 -146.918    62.119 -4042.000
    770.723

Mean of 1st generation migrant proportion              = 0.206
Standard deviation of 1st generation migrant proportion = 0.069
Mean of log genotype likelihood                        = -3895.082
Variance of log genotype likelihood                    = 818.030
Mean of log probability of migrant configuration       = -146.918
Variance of log probability of migrant configuration   = 62.119
Mean of log total likelihood                           = -4042.000
Variance of log total likelihood                       = 770.723

Migration rate
Pop    LZ2     TZ4     HL4     JL6     NM6     NX6     DT6     GC4
LZ2    0.807   0.040   0.040   0.048   0.009   0.021   0.016   0.019
TZ4    0.012   0.813   0.030   0.010   0.009   0.067   0.044   0.016
HL4    0.009   0.021   0.850   0.011   0.009   0.027   0.064   0.009
JL6    0.030   0.050   0.010   0.756   0.010   0.016   0.108   0.019
NM6    0.015   0.021   0.011   0.009   0.680   0.158   0.071   0.035
NX6    0.012   0.022   0.013   0.014   0.009   0.879   0.015   0.035
DT6    0.017   0.014   0.042   0.013   0.009   0.030   0.866   0.010
GC4    0.013   0.160   0.015   0.011   0.009   0.018   0.071   0.705
```
Box 13. Example results of a run for migration rate estimation

4.25 Mantel tests

The Mantel test is used to measure the degree of association between two dissimilarity

matrices (Mantel 1967). This test is an extension of the partial Mantel test (Smouse & Sokal 1986) to control an additional matrix. In POLYGENE, the Mantel test is further extended to control for up to five additional matrices. The interface of the Mantel test is shown in Figure 34, where the box at the top-left of the screen is used to configure the number of permutations. The input matrices are in the textbox at the right of the screen, with the output results in the textbox at the bottom-left.
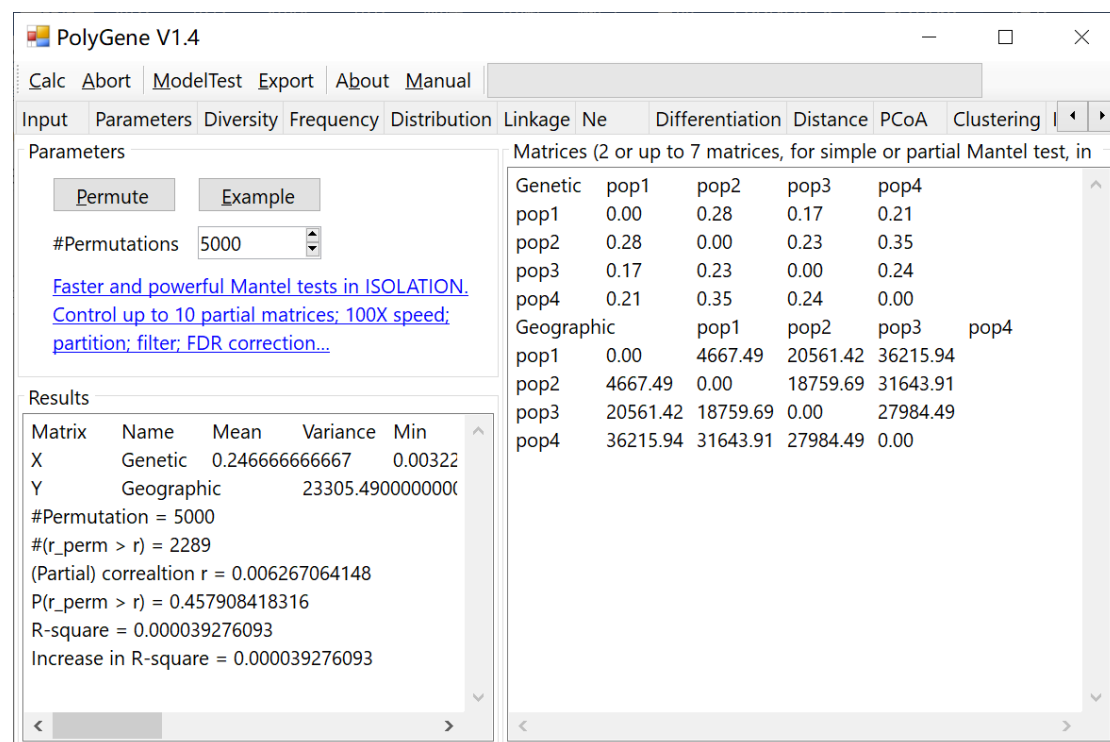
Figure 34. Mantel test

An example of the input matrices is shown in Box 13. This presents the results of a simple Mantel test, which uses only two input matrices. Generally, the input matrices are in order of Y, X, Z_1, Z_2, Z_3, Z_4 and Z_5. Each matrix is a table, whose name is in the first cell, and the first row and the first column are the headers consisting of the population names except for the first cell. Note that each dissimilarity matrix must be symmetrical, whose diagonal elements must all be zero.

```
Genetic      pop1      pop2      pop3      pop4
pop1         0.00      0.28      0.17      0.21
pop2         0.28      0.00      0.23      0.35
```

```
pop3          0.17       0.23       0.00       0.24
pop4          0.21       0.35       0.24       0.00
Geographic    pop1       pop2       pop3       pop4
pop1          0.00       4667.49    20561.42   36215.94
pop2          4667.49    0.00       18759.69   31643.91
pop3          20561.42   18759.69   0.00       27984.49
pop4          36215.94   31643.91   27984.49   0.00
```

Box 13. Example of the input matrices for Mantel test

An example results for a Mantel test is shown in Box 14, where the first three rows form a table, whose columns are in turn the identifier and the name of each input matrix, as well as the mean, variance, minimum, maximum and the sum of squares of all non-diagonal elements in each input matrix. Next comes the permutation results, including the random number generator seed, the number of permutations, the number of permutations that the correlation (between the permuted matrices **X** and **Y** with a controlled matrix **Z**) is greater than the original value, the original correlation, single-tailed P value and the R-square value (i.e., the coefficient of determination).

```
Matrix Name       Mean        Variance      Min        Max       SS
X      Genetic    0.247       0.003         0.170      0.350     0.019
Y      Geographic 23305.490   105611878.351 4667.490   36215.940 633671270.104

Random number generator seed: 23483077
#Permutation = 10000
#(r_perm > r) = 4658
r = 0.006
P(r_perm > r) = 0.466
R-square = 0.000
```

Box 14 Example results of a Mantel test

5 Methodology

5.0 Frequently used symbols

A, B allele
P allele frequency
v ploidy level or number of haplotypes in a group in AMOVA
L number of loci
l locus loop variable
J number of alleles

j allele loop variable
H heterozygosity/H-Index
$\widehat{M}$ unbiased homozygosity estimate
h haplotype or haplotype loop variable
K number of clusters in Bayesian clustering
k cluster loop variable in Bayesian clustering
p subpopulation
t total population
x, y individual or population loop variable
r relatedness coefficient or linkage disequilibrium measure or LOCPRIORI model parameter in Bayesian clustering
θ coancestry (kinship) coefficient
F Fixation index or F model parameter in Bayesian clustering
f inbreeding coefficient
s selfing rate or location loop variable in Bayesian clustering
$\mathcal{L}$ likelihood
$\mathcal{P}$ allelic phenotype
$\mathcal{G}$ zygote genotype
g gamete genotype
G Fisher's G statistics or two-locus genotypic frequency in Burrow's Δ
β negative amplification rate
e mistyping rate
r_s recombination fraction
α double-reduction rate or Dirichlet parameter for allele frequencies in Bayesian clustering

5.1 Allele frequency estimation

The estimation of allele frequencies is fundamental to most population genetics analyses. In POLYGENE, four polysomic inheritance models are adopted to calculate the allelic phenotypic or genotypic frequencies, the disomic inheritance model is equivalent to the RCS model in the single-locus genotypic frequency. These polysomic inheritance models are RCS, PRCS, CES and PES (see Section 4.3 for details). The diagram of these models under tetrasomic inheritance is presented in Figure 35. The double-reduction rates for these models at the ploidy levels 4, 6, 8, 10 and 12 are shown in Table 1, where the probability that a gamete carrying i pairs of IBDR alleles is called a double-reduction rate, denoted by α_i.
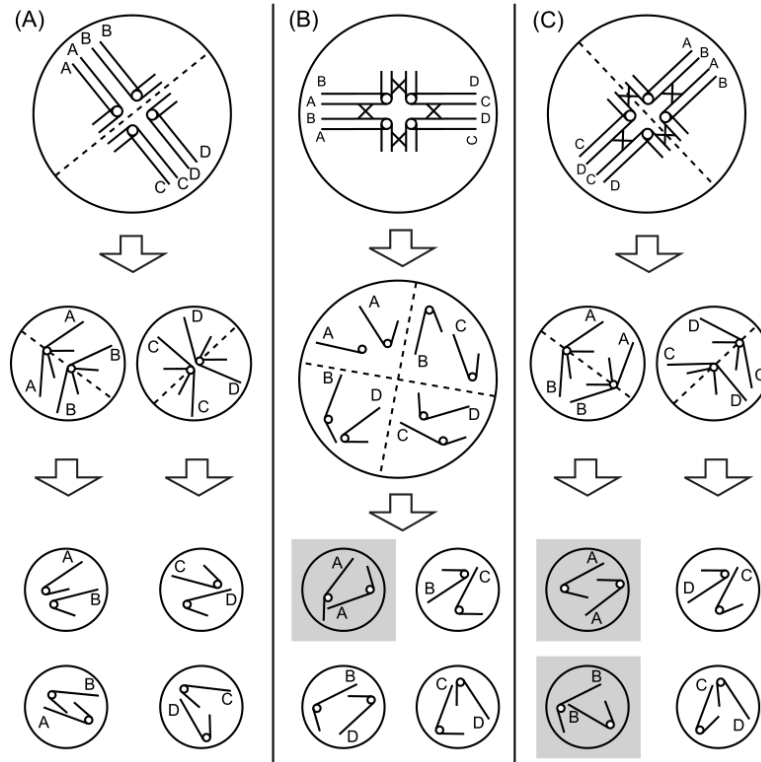
Figure 35. Diagram of polysomic models under tetrasomic inheritance modified from Huang *et al.* (2019). These inheritance model assumes 100% multivalent formation. The leftmost column shows the primary oocytes, the middle column shows the secondary oocytes in (A) and (C) or the tetrad in (B), and the rightmost column shows the gametes. The gametes with a gray background carry the identical-by-double-reduction (IBDR) alleles. Dashed lines denote cellular fission, solid lines denote the arms of the chromosome, and the circles connecting solid lines denote the centromere. The target locus is located in the long arm of the chromosome and the identical-by-descent alleles are denoted by the same letter. (A) Random chromosome segregation (RCS) ignores the crossover between the target locus and the centromere (Muller 1914). In the absence of crossing over, gametes may originate from any combination of homologous chromosomes, and two sister chromatids never sort into the same gamete (Parisod *et al.* 2010). (B) Pure random chromatid segregation (PRCS) accounts for the crossing over between the target locus and the centromere, and assumes that the chromatids behave independently and randomly segregate into gametes (Haldane 1930). When sister chromatids are segregated into the same gamete, double-reduction occurs. The probability that the two chromatids in a gamete are sister chromatids is $4/\binom{8}{2}$, i.e., 1/7, where 4 is the number of sister chromatids pairs) divided by, and $\binom{8}{2}$ is the number of ways to sample two chromatids from eight chromatids. (C) Complete equational segregation (CES), homologous chromosomes pair and chromatids exchange via recombination (Mather 1935). The whole arms of sister chromatids are exchanged into different chromosomes. The probability that two homologous chromosomes within a single secondary oocyte were previously paired at a target locus in Prophase I is 1/3. In this case, these sister chromatid fragments will become segregated into a single gamete at a ratio of 1/2, so the rate of double-reduction is 1/6 for tetrasomic inheritance.

5 Methodology

Table 1. The double-reduction rates in four models

Model	Alpha	Ploidy level				
		4	6	8	10	12
RCS	α_1	0	0	0	0	0
	α_2			0	0	0
	α_3					0
PRCS	α_1	1/7	3/11	24/65	140/323	1440/3059
	α_2			1/65	15/323	270/3059
	α_3					5/3059
CES	α_1	1/6	3/10	27/70	55/126	285/616
	α_2			3/140	5/84	65/616
	α_3					5/1848
PES	α_1	$r_s/6$	$3r_s/10$	$\frac{3}{70}r_s(10-r_s)$	$\frac{5}{126}r_s(14-3r_s)$	$\frac{5}{616}r_s(84-28r_s+r_s^2)$
	α_2			$\frac{3}{140}r_s^2$	$\frac{5}{84}r_s^2$	$\frac{5}{616}r_s(14-r_s)$
	α_3					$\frac{5}{1848}r_s^3$

The symbol r_s denotes the recombination fraction.

Under each of these models, the frequency of each genotype can be calculated. The natural logarithm of genotypic likelihood at a given locus is given by

$$\ln \mathcal{L} = \sum_{\mathcal{P}} \sum_{\mathcal{G} \rhd \mathcal{P}} \Pr(\mathcal{G}|\mathcal{P}) \ln \Pr(\mathcal{G}),$$

where $\mathcal{P}$ is taken from all allelic phenotypes of the whole individuals at this given locus, $\mathcal{G} \rhd \mathcal{P}$ denotes that $\mathcal{G}$ is a genotype determining $\mathcal{P}$ or determining $\mathcal{P} \cup \{A_y\}$ if the null allele A_y is considered, $\Pr(\mathcal{G}|\mathcal{P})$ is the posterior probability of $\mathcal{G}$ conditional on $\mathcal{P}$, and $\Pr(\mathcal{G})$ is the genotypic frequency under one of the four inheritance models mentioned above.

If self-fertilization is not considered, the genotypic frequency can be calculated under a number of inheritance model. Otherwise, the approximated frequency $\Pr(\mathcal{G}, f)$ of a genotype $\mathcal{G}$ is derived by using the inbreeding coefficient f as an intermediate variable, that is

$$\Pr(\mathcal{G}, f) = \binom{v}{n_1, n_2, \ldots, n_J} \prod_{j=1}^{J} \prod_{i=0}^{n_j-1} \left[\left(\frac{1}{f} - 1 \right) P_j + i \right] \bigg/ \prod_{i=0}^{v-1} \left(\frac{1}{f} - 1 + i \right),$$

where v is the ploidy level of individuals in the same population, J is the number of alleles in $\mathcal{G}$, n_j in the multinomial coefficient is the number of copies of the jth allele in $\mathcal{G}$,

and P_j is the frequency of jth allele in the reference population. The inbreeding coefficient f under both a double-reduction and self-fertilization is calculated by

$$f = \frac{8\alpha + sv}{8\alpha + v(s + v - sv)},$$

where α is the expected number of IBDR alleles in a gamete, that is $\alpha = \sum_{i=0}^{\lfloor v/4 \rfloor} i\alpha_i$, v is the ploidy level and s is the selfing rate (Huang *et al.* 2019). Because f will vary as s varies, it can be regarded as a function of s. Moreover, it can be seen from Table 1 that f can also be regarded as a function of the recombination fraction r_s under the PES model.

The posterior probability $\Pr(\mathcal{G}|\mathcal{P})$ can be calculated by following the Bayes equation:

$$\Pr(\mathcal{G}|\mathcal{P}) = \frac{T(\mathcal{P}|\mathcal{G})\Pr(\mathcal{G})}{\Pr(\mathcal{P})},$$

where $\Pr(\mathcal{P})$ is the frequency of $\mathcal{P}$, $T(\mathcal{P}|\mathcal{G})$ is the transitional probability from $\mathcal{G}$ to $\mathcal{P}$, and they are calculated by

$$\Pr(\mathcal{P}) = \begin{cases} (1-\beta)\sum_{G \triangleright \mathcal{P}} \Pr(\mathcal{G}) & \text{if } \mathcal{P} \neq \emptyset, \\ \beta + (1-\beta)\sum_{G \triangleright \mathcal{P}} \Pr(\mathcal{G}) & \text{if } \mathcal{P} = \emptyset; \end{cases}$$

$$T(\mathcal{P}|\mathcal{G}) = \begin{cases} 1 & \text{if } \mathcal{P} = \emptyset \text{ and } G \triangleright \mathcal{P}, \\ \beta & \text{if } \mathcal{P} = \emptyset \text{ and } G \not\triangleright \mathcal{P}, \\ 1 - \beta & \text{if } \mathcal{P} \neq \emptyset \text{ and } G \triangleright \mathcal{P}, \\ 0 & \text{if } \mathcal{P} \neq \emptyset \text{ and } G \not\triangleright \mathcal{P}, \end{cases}$$

where β is the negative amplification rate.

Under self-fertilization, $\Pr(\mathcal{G})$ should be replaced by $\Pr(\mathcal{G}, f)$ in the above expression of $\Pr(\mathcal{P})$. Because the intermediate variable f is a function of the selfing rate s, so is the frequency $\Pr(\mathcal{P})$. Similarly, $\Pr(\mathcal{P})$ is a function of the recombination fraction r_s under the PES model.

Following the methods of Kalinowski *et al.* (2006), an EM algorithm is used to maximize the genotypic likelihood at each locus. For all alleles at the same locus, the initial values of their frequencies are assumed to be equal and are updated during iteration. The

5 Methodology

relevant iterative equations for the EM optimization are

$$\hat{P}'_j = \frac{\sum_{\mathcal{P}} \sum_{G \triangleright \mathcal{P}} v \, \Pr(A_j | \mathcal{G}) \Pr(\mathcal{G} | \mathcal{P})}{\sum_{\mathcal{P}} v}, \qquad j = 1, 2, \cdots, J,$$

where $\hat{P}'_j$ is the updated frequency of the jth allele A_j at a locus, $\mathcal{P}$ is taken from all allelic phenotypes at this locus, J is the number of alleles in the total population (including the null allele if it is considered), $\Pr(A_j | \mathcal{G})$ is the frequency of A_j in G (thus $v \, \Pr(A_j | \mathcal{G})$ is the dosage of A_j in $\mathcal{G}$), and $\Pr(\mathcal{G} | \mathcal{P})$ is calculated from the allele frequencies at the current step.

Besides, at the same locus as above, the negative amplification rate is updated by

$$\hat{\beta}' = \frac{N_\emptyset \hat{\beta} / \Pr(\mathcal{P} = \emptyset)}{N},$$

where $N_\emptyset$ is the number of negative allelic phenotypes at this locus, N is the number of individuals in the total population, and $\hat{\beta} / \Pr(\mathcal{P} = \emptyset)$ is the posterior probability that a negative allelic phenotype is due to the negative amplification.

If the maximum of the absolute values of all differences, each of which is the difference between the two frequencies of an allele under two adjacent iterations, reaches the threshold 10^{-8}, or if the number of iterations reaches 2000, then the procedure of iterations is terminated. For null alleles, their frequency can also be estimated by adding them into the candidate genotypes.

The selfing rate s is assumed to be equal in the same population. For maximum-likelihood estimator estimate s with a down-hill simplex algorithm (Nelder & Mead 1965) by maximizing the allelic phenotypic likelihood. The estimated selfing rate $\hat{s}$ can be expressed as

$$\hat{s} = \arg\max_{s \in [0,1]} \sum_{l=1}^{L} \sum_{\mathcal{P}_l} \ln \Pr(\mathcal{P}_l),$$

in which L is the number of loci, $\mathcal{P}_l$ is taken from all allelic phenotypes at the lth locus, and $\Pr(\mathcal{P}_l)$ is a function of the selfing rate s as mentioned above.

5 Methodology

The remaining two selfing rate estimators do not update $\hat{s}$ during the estimation of allele frequencies, but estimate $\hat{s}$ before estimating allele frequencies. Hardy's (2016) Fz-based estimator estimate s from inbreeding coefficient estimate, the inbreeding coefficient is estimated by Hardy's (2016) Eqn (10):

$$\hat{F}_z = 1 - \frac{\sum_l \sum_x h_{xl}}{\sum_l \left[\frac{(N-M_l)^2}{(N-M_l-1)} \left(1 - \sum_a \hat{P}_{al}^2 - \frac{J-1}{J(N-M_l)^2} \sum_x \hat{h}_{xl} \right) \right]}.$$

where N is the sample size, M_l is the number of individuals with missing data at locus l, h_{xl} is the heterozygosity of individual i at locus l, and $\hat{P}_{al}$ is the allele frequency estimate of a rough method: if an individual has n alleles, then the count of each allele is added by v/n. If the genotype of individual i at locus l is missing, then h_{xl} is set to 0, otherwise h_{xl} is obtained from Hardy's (2016) Table 1. The selfing rate is estimated by Hardy's (2016) Eqn (15):

$$\hat{s} = \frac{v\hat{F}_z - (v-2)(1-\hat{F}_z)\alpha}{1 + (v-1)\hat{F}_z}.$$

where α is the Hardy's (2016) double-reduction rate, which is the probability that a random pair of homologous gene copies sampled without replacement within a gamete derived from the same chromosome during meiosis, and can be convert from double reduction rate α_i by

$$\alpha = \frac{8}{v(v-2)} \cdot \sum_{i=1}^{\lfloor v/4 \rfloor} i\alpha_i.$$

If the loci have different α, then α is averaged across loci.

Hardy's (2016) g2z-based estimator estimate s from identity disequilibrium coefficients, which is a measure of the correlation between the heterozygosities of distinct loci. The identity disequilibrium coefficient between two loci (l and m) is estimated by Hardy's (2016) Eqn (14):

$$\hat{g}_{2z} = \frac{\sum_l \sum_{m \neq l} (\sum_x h_{xl} h_{xm})}{\sum_l \sum_{m \neq l} \frac{N - M_l - M_m + M_{lm}}{(N-1)(N-M_l-M_m) + M_l M_m - M_{lm}} [(\sum_x h_{xl})(\sum_x h_{xm}) - \sum_x h_{xl} h_{xm}]} - 1.$$

where M_{lm} number of individuals with missing data at locus l and m. The selfing rate is

estimated by Hardy's (2016) Eqn (16):

$$\hat{s} = \frac{X - \sqrt{X^2 - YZ}}{Y}.$$

where $X = 1 + (\alpha_l + \alpha_m)(v - 2) + \alpha_l\alpha_m(v - 2)^2 + \hat{g}_{2z}[9 - 16v + 7v^2 - (\alpha_l + \alpha_m)(2 - 3v + v^2)]$, $Y = \hat{g}_{2z}[14 - 20v + 7v^2 - (\alpha_l + \alpha_m)(8 - 10v + 3v^2) + \alpha_l\alpha_m(v - 2)^2]$, and $Z = \hat{g}_{2z}[4 - 12v + 7v^2 + (\alpha_l + \alpha_m - \alpha_l\alpha_m)(v - 2)^2]$.

The recombination fraction r_s in the PES model is assumed to be equal at the same locus among all populations, which is also estimated by a down-hill simplex algorithm (Nelder & Mead 1965) by maximizing the allelic phenotypic likelihood. The estimated recombination fraction $\hat{r}_s$ can be expressed as

$$\hat{r}_s = \underset{r_s \in [0,1]}{\arg\max} \sum_{p=1}^{N_{\text{pop}}} \sum_{\mathcal{P}_p} \ln \Pr(\mathcal{P}_p | p),$$

in which N_{pop} is the number of populations, $\mathcal{P}_p$ is taken from all allelic phenotypes of the whole individuals in the pth population, $\Pr(\mathcal{P}_p | p)$ denotes the frequency of $\mathcal{P}_p$ in the pth population, which is a function of r_s as mentioned above.

It should be noted that because both r_s and s can influence the inbreeding coefficient f, these two parameters cannot be estimated simultaneously, i.e., the options 'Consider selfing' and 'PES estimate rs' are incompatible.

5.2 Genetic diversity

In order to extend the discussion of disomic inheritance into polysomic inheritance, some parameters of genetic diversity need to be modified. In POLYGENE, there are several parameters of genetic diversity listed in the following text.

- k, the number of distinct alleles at a locus, where the **null allele** is also considered.

- n, the total number of individuals genotyped at a locus.

- Ho, the observed heterozygosity, which is extended to polysomic inheritance. There are four hierarchical observed heterozygosities: the Ho for an individual is defined

as the probability of randomly choosing two non-IBS alleles within this individual without replacement; the Ho for a population (a region or the total population) is the weighted average of those Ho for all individuals within this population (this region or the total population).

- He, the expected heterozygosity in the current group, like the disomic inheritance, whose value at a locus is $1 - \sum_{j=1}^{J} \hat{P}_j^2$, where $\hat{P}_j$ is the frequency of the jth allele A_j at this locus.

- PIC, the polymorphic information content, like the disomic inheritance, whose value at a locus is $2 \sum_{1 \leq i < j \leq J} \hat{P}_i \hat{P}_j (1 - \hat{P}_i \hat{P}_j)$.

- Ae, the effective number of alleles, like the disomic inheritance, whose value at a locus is $1 / \sum_{j=1}^{J} \hat{P}_j^2$.

- Ar, the allelic richness, like the disomic inheritance, whose value at a locus in a population/region p is $k_t + \sum_{A_j \notin p} (1 - \hat{P}_{tj})^{v_{pl}}$, where k_t is the number of alleles in the total population, $A_j \notin p$ denotes the alleles A_j is not sampled in p, $\hat{P}_{tj}$ is the allele frequency in the total population, and v_{pl} is the number of allele copies in the current population/region.

- I, Shannon's Information Index, like the disomic inheritance, whose value at a locus is $-\sum_{j=1}^{J} \hat{P}_j \ln \hat{P}_j$.

- NP, number of private alleles, where private alleles are the alleles only preset in this population/region.

- NE-1P, the average probability of not excluding a candidate parent from the parentage of an arbitrary offspring, given only the genotype of this offspring, like the disomic inheritance, whose value at a locus is

$$4a_2 - 2a_2^2 - 4a_3 + 3a_4,$$

where $a_i = \sum_{j=1}^{J} \hat{P}_j^i$, $i = 1, 2, 3, 4$.

- NE-2P, the average probability of not excluding a candidate parent from the parentage of an arbitrary offspring, given the genotype of this offspring and of the known parent of the opposite sex, like the disomic inheritance, whose value at a

locus is

$$2a_2^2 + 2a_2 - a_3 - 3a_2a_3 + 3a_5 - 2a_4.$$

- NE-PP, the average probability of not excluding a candidate parent pair from the parentage of an arbitrary offspring, given only the genotype of this offspring, like the disomic inheritance, whose value at a locus is

$$4a_5 - 4a_4 + 3a_6 + 8a_2^2 - 8a_2a_3 - 2a_3^2.$$

- NE-I, the average probability that the genotypes at a single locus do not differ between two unrelated individuals, like the disomic inheritance, whose value at this locus is

$$4a_2^2 - 3a_4.$$

- NE-SI, the average probability that the genotypes at a single locus do not differ between two full siblings, like the disomic inheritance, whose value at this locus is

$$\frac{1}{4} + \frac{1}{2}a_2 + \frac{1}{2}a_2^2 - \frac{1}{4}a_4.$$

- Fix, the inbreeding coefficient in this group, which is defined as the probability of sampling two identical-by-descent (IBD) alleles from an individual without replacement, whose value is estimated by $1 - H_I/H_X$, where the H_I is the heterozygosity index within individuals (i.e., Ho), H_X is the expected heterozygosity in the current group (i.e., He).

- Fsx (or Fc1x/Fc2x/…), the F-statistics among populations (or regions) within this group (say X), which is defined as the probability of sampling two identical-by-descent (IBD) alleles from a population (or a region) without replacement, whose value is estimated by $1 - H_S/H_X$, where the H_S is the weighted average of expected heterozygosity of populations (or regions) with the number of allele copies as the weight, and H_X is the expected heterozygosity of this group.

5.3 Allelic phenotype or genotype distribution test

In diploids, the Markov chain test (Guo & Thompson 1992) or Fisher's G-test is usually used to test whether the genotypic frequencies accord with the Hardy-Weinberg

5 Methodology

equilibrium. Because the Markov chain test (Guo & Thompson 1992) can only be applied to the disomic/RCS model, only Fisher's *G*-test is employed to perform the test of allelic phenotype or genotype distribution.

Let Λ denote the possible genotypes at a target locus. Then there is $\Lambda = \binom{J+1}{2}$ in diploids. Because the number of allele copies is fixed, the genotypic distribution is subjected to the following J constraints:

$$w_i = w_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{J} w_{ij}, \qquad i = 1, 2, \cdots, J,$$

where w_i is the number of copies of the ith allele A_i at this locus, w_{ii} is the number of homozygotes $A_i A_i$, and w_{ij} is that of the heterozygotes $A_i A_j$ ($i \neq j$) at the same locus. Similarly, there is $\Lambda = \binom{J+v-1}{v}$ in polyploids. Likewise, the genotypic distribution is subjected to J constraints as follows:

$$w_i = \sum_{\mathcal{G}} v \Pr(A_i | \mathcal{G}), \qquad i = 1, 2, \cdots, J,$$

where $\mathcal{G}$ is taken from all possible genotypes at this locus. Therefore, the degrees of freedom for the genotypic distribution are $\Lambda - J$.

The number of allelic phenotypes consisting of i distinct alleles at target locus is equal to $\binom{J}{i}$, $1 \leq i \leq v$. The number of possible allelic phenotypes at the same locus is therefore $\sum_{i=1}^{\min(v,J)} \binom{J}{i}$, denoted by Λ^*. Because there is only one constraint for the allelic phenotype distribution, that is $\sum_{\mathcal{P}} \Pr(\mathcal{P}) = 1$, the degrees of freedom for this distribution are $\Lambda^* - 1$. This number of degrees of freedom is lower than for the genotypic distribution of the same ploidy level v ($v > 2$). For haploids and diploids, each allelic phenotype has only one candidate genotype. That is because the possible genotypes determining an allelic phenotype $\mathcal{P}$ can only be either $\mathcal{P}$ itself or a homozygote in the form AA. Therefore, regardless of whether haploid or diploid, the allelic phenotype distribution has the same number of degrees of freedom as the genotype distribution.

5 Methodology

The meanings of the expected and the observed occurrences together with the contingency table can be seen from Section 5.4. If there is at least one expected occurrence of a allelic phenotype or a genotype whose value is less than one, or if there are at least 20% of the expected occurrences whose values are less than five, then there are two rows or two columns in the contingency table that are combined.

The expected occurrence of each allelic phenotype or each genotype can be calculated under a specific inheritance model. After that, Fisher's G statistic is calculated by

$$G = \sum_j 2O_j \ln\left(\frac{O_j}{E_j}\right).$$

where O_j and E_j are the values of the observed and the expected occurrences of the allelic phenotype $\mathcal{P}_j$ ($j = 1, 2, \cdots, \Lambda^*$) or of the genotype $\mathcal{G}_j$ ($j = 1, 2, \cdots, \Lambda$). The P-value can be obtained by the right-tailed probability of a Chi-squared distribution.

It noteworthy that if the null allele or negative amplification are considered, then the negative allelic phenotypes are also taken into account in allelic phenotype distribution test. Otherwise, the negative allelic phenotypes cannot be explained and is discarded. In allelic phenotype distribution test, if the expected occurrence of any allelic phenotype is less than 5, then the two rare most allelic phenotypes are collapsed and treated as one allelic phenotype. In genotype distribution test, if the expected occurrence of any genotype is less than 5, then two minor alleles are collapsed and treated as one allele.

The number of parameters k_{par} to describe the data are the allele frequencies ($J - 1$ parameters), the selfing rate s, the null allele frequency p_y, the negative amplification rate β and the single-chromatid recombination fraction r_s. The latter four parameters are accounted for if the corresponding factors are considered. The likelihood for the genotype/allelic phenotype data in a population/region are then calculated, which is the product of genotypic/allelic phenotypic frequencies across individuals (including the negative allelic phenotype in calculating the allelic phenotypic likelihood):

$$\mathcal{L} = \sum_{\mathcal{P}} \Pr(\mathcal{P}) \quad \text{or} \quad \mathcal{L} = \sum_{\mathcal{G}} \Pr(\mathcal{G}).$$

The values of AICc and BIC are also calculated, which can be used to choose the optimal

inheritance model. The model with a lower AICc or BIC is better.

$$\text{AICc} = -2\ln\mathcal{L} + 2k_{\text{par}} + 2\frac{k_{\text{par}}(k_{\text{par}} + 1)}{n - k_{\text{par}} - 1},$$

$$\text{BIC} = -2\ln\mathcal{L} + k_{\text{par}}\ln n.$$

5.4 Linkage disequilibrium test

Fisher's *G*-test

For all individuals in a population, if the numbers of their observed different allelic

phenotypes at two loci are n_1 and n_2, respectively, then a contingency table including two

matrices of type $n_1 \times n_2$ can be established, where the elements in these two matrices are

the counts of the observed and the expected allelic phenotypes (also known the values of

the observed and the expected occurrences). For convenience, these two matrices are

called the observed and the expected matrices, respectively. An example of such a

contingency table is shown in Table 2.

The number of degrees of freedom for the contingency table is $(n_1 - 1)(n_2 - 1)$, and

Fisher's *G* statistic is calculated from the contingency table. If there is at least one

expected occurrence, whose value is less than one, or if there are at least 20% of expected

occurrences, whose values are less than five, then two rows or two columns with the

smallest and the second smallest sum of occurrences are chosen to be combined. By

calculating the right-tailed probability of a Chi-squared distribution, the *P*-value can be

obtained.

Table 2. Counts of the observed and expected allelic phenotypes at two loci

Observed	Locus 2		Expected	Locus 2	
Locus 1	A	AB	Locus 1	A	AB
M	3	5	M	3.15	4.85

MN	6	8	MN	5.52	8.48
N	4	7	N	4.33	6.67

For each pair of tested loci, the FDR correction (Benjamini & Hochberg 1995) is also applied to multiple tests. This test is not extended to varying ploidy levels.

Raymond & Rousset's (1995) Markov Chain test

Raymond & Rousset's (1995) Markov Chain test is a numerical method to obtain the significance of a contingency table, which is performed by first randomly choose two different rows (say the i_1^{th} and i_2^{th} rows) and two different columns (say the j_1^{th} and j_2^{th} columns) in the observed matrix, and then denote the elements located in the four crossover points of these rows and columns by $O_{i_1 j_1}$, $O_{i_1 j_2}$, $O_{i_2 j_1}$ and $O_{i_2 j_2}$ in turn. Next, replace $O_{i_1 j_1}$ by $O_{i_1 j_1} - 1$, $O_{i_1 j_2}$ by $O_{i_1 j_2} + 1$, $O_{i_2 j_1}$ by $O_{i_2 j_1} + 1$ and $O_{i_2 j_2}$ by $O_{i_2 j_2} - 1$ in this matrix. Such a process is called a switch, which needs to be made many times. After each switch, the statistic ρ is updated by the following formula:

$$\rho' = \rho + \ln\left(\frac{O_{i_1 j_1} O_{i_2 j_2}}{(O_{i_1 j_2} + 1)(O_{i_2 j_1} + 1)}\right).$$

For a Markov Chain test, it is necessary to make some iterations before sampling. Such a process is called the burnin or the dememorization. After the burnin period, such an algorithm will continuously perform several batches, each of which consists of several iterations. This algorithm can be used to calculate the probability for each batch that ρ' is less than or equal to zero, denoted by $\Pr(\rho' \leq 0)$. The arithmetic means and the standard error of the probabilities (such as $\Pr(\rho' \leq 0)$) for all batches are the unbiased P value and the SE in the results, respectively. The total number of switches will also be included in the output. For each pair of tested loci, the FDR correction (Benjamini & Hochberg 1995) is also applied to multiple tests.

Burrows's Δ and squared correlation coefficient

Linkage disequilibrium (LD) is the non-random association of alleles at different loci in a given population. The most frequently used measure of disequilibrium is the square

5 Methodology

correlation coefficient r^2 (Hill & Weir 1994). The value of r^2 is calculated from D statistics. which is defined as

$$r^2 = \frac{\sum_{AB} D_{AB}^2}{\sum_{AB} Q_{AB}}, \qquad D_{AB} = P_s^{AB} - P_A P_B,$$

$$Q_{AB} = P_A(1 - P_A)P_B(1 - P_B),$$

where D_{AB} is the deviation of the haplotypic frequency from expect that assuming the two loci are independent, also the correlation coefficient between alleles A and B each at a different locus, Q_{AB} is the theoretical maximum of D_{AB}^2, P_s^{AB} is the frequency that the two alleles in the same haplotype are A and B, and P_A and P_B are respectively the allele frequencies of A and B among individuals with both loci genotyped.

Unfortunately, these statistics are calculated from phased genotypes. For unphased genotypes, Burrows's Δ statistics (Cockerham & Weir 1977) is used as an alternative to measure the LD between two alleles, and Huang *et al.* (unpublished) have expanded it to the polysomic inheritance. Δ_{AB} is defined as the deviation of a linear combination of observed frequencies of some two-locus genotypes (that have both A and B) from the expect assuming linkage equilibrium. The above equation can be revised as

$$r_\Delta^2 = \frac{\sum_{AB} \Delta_{AB}^2}{\sum_{AB} R_{AB}}, \qquad \Delta_{AB} = \left(\sum_{i=1}^{v}\sum_{j=1}^{v} \frac{ij}{v}G_{ij}\right) - vP_A P_B,$$

$$R_{AB} = [P_A(1 - P_A) + (v - 1)(P_{AA} - P_A^2)][P_B(1 - P_B) + (v - 1)(P_{BB} - P_B^2)].$$

where G_{ij} is the two-locus unphased genotypic frequencies, in which the genotype at the first locus has i copies of A and that at the second locus has j copies of B, R_{AB} is the theoretical maximum of Δ_{AB}^2, and P_{AA} is the probability that sampling two allele copy A in the same individual without replacement.

There are in total Nv^2 allele pairs within individuals in the sample. The observed expected occurrence of AB pair is $NvP_s^{AB} + Nv(v - 1)P_d^{AB}$ and $Nv^2 P_A P_B$ respectively. Therefore the statistics $N\sum_{AB} \frac{\hat{\Delta}_{AB}^2}{\hat{P}_A \hat{P}_B} \sim \chi_{J_1-1, J_2-1}^2$ (Weir & Cockerham 1979). It noteworthy that J_1 and J_2 is only calculated from the individuals with both loci genotyped. To perform the linkage disequilibrium test, Huang *et al.* (unpublished) follow Weir and

5 Methodology

Cockerham (1979) and establish a contingency table with J_1 rows and J_2 columns, and perform another Fisher's G test and another Raymond & Rousset's (1995) Markov Chain test to test the null hypothesis that all $\Delta_{AB} = 0$. For each pair of tested loci, the FDR correction (Benjamini & Hochberg 1995) is also applied to multiple tests.

Huang *et al.* (unpublished) also expand the above statistics and tests to allow varying ploidy levels, in which Δ_{AB} is redefined as $P_s^{AB} + (\tilde{v} - 1)P_d^{AB} - \tilde{v}P_A P_B$, where $\tilde{v}$ is the adjusted ploidy level and is equal to $\frac{\sum_i v_i^2}{\sum_i v_i}$ with v_i being the ploidy level of individual i. Δ_{AB} can be estimated by

$$\hat{\Delta}_{AB} = \frac{\sum_i v_i \sum_{jk} \frac{jk}{v} \hat{G}_{vjk}}{\sum_i v_i} - \tilde{v}\hat{P}_A\hat{P}_B.$$

where $\hat{G}_{vij}$ is the observed two-locus genotypic frequency at ploidy level v. $\hat{R}_{AB}$ should be revised as

$$\hat{R}_{AB} = [\hat{P}_A(1 - \hat{P}_A) + (\tilde{v} - 1)(\hat{P}_{AA} - \hat{P}_A^2)][\hat{P}_B(1 - \hat{P}_B) + (\tilde{v} - 1)(\hat{P}_{BB} - \hat{P}_B^2)].$$

5.5 Effective population size

Wright-Fisher population

The Wright–Fisher model describes a population with discrete, nonoverlapping generations. This virtual population is monoecious and self-fertilization is allowed. During reproduction, each individual generates infinity but equal haploid gametes. They are gathered in a gamete pool, and are randomly merged to produce the individuals in the next generation.

For haploids and diploids Wright–Fisher population, each offspring is generated by randomly sampling one and two gametes from the gamete pool, respectively. Huang *et al.* (unpublished) extended such mating system (termed haplotype sampling, HS) to polyploids, and assume each offspring is generated by sampling v gametes. The expected heterozygosity in generation t can be expressed by

5 Methodology

$$E(H_t) = \left(1 - \frac{1}{vN_e}\right) H_{t-1}, \text{ or}$$

$$E(H_t) = \left(1 - \frac{1}{vN_e}\right)^t H_0 \approx e^{-t/(vN_e)} H_0.$$

If the changing pattern in allele frequency of a real population is identical to that of a Wright-Fisher population, the size of this Wright-Fisher population is termed the effective population size N_e of this real population.

Pudovkin's (1996) heterozygosity excess estimator

We modified Pudovkin's (1996) estimator to account for polysomic inheritance. This estimator is based on the deviation of observed heterozygosity from Hardy-Weinberg expectation. This model assumes a dioecious population with random mating. In this population, each generation has infinity individuals, but a small number of males and females are involved in reproduction. The numbers of reproductive males and females are N_f and N_m, respectively. Therefore, this estimator estimates the effective number of breeders N_{eb}, but N_{eb} is nearly equivalent to N_e with non-overlapping generations. We will use N_e to denote both notations in the following text.

For a target allele A at a diallelic locus, its frequencies in the parental and offspring generations are denoted as P_0 and P_1. The allele frequencies of A in the reproductive females and males are differ due to the binomial sampling error, whose frequencies are denoted by P_f and P_m, respectively. The expectations of $E(P_f)$, $E(P_m)$, $E(P_f^2)$, $E(P_m^2)$ and $E(P_f P_m)$ can be derived. The expectation of observed and expected heterozygosity in the offspring generation is derived by

$$E(H_O) = \binom{v}{2}^{-1} \left\{ \binom{v/2}{2} [4P_0(1 - P_0)] + \left(\frac{v}{2}\right)^2 [P_f(1 - P_m) + P_m(1 - P_f)] \right\},$$

$$E(H_E) = 2 \frac{P_f + P_m}{2} \left(1 - \frac{P_f + P_m}{2}\right).$$

The Selander's (1970) heterozygosity excess index D is calculated by

$$D = \frac{H_O}{H_E} - 1.$$

By ignoring the correlation between H_O and H_E (which can be eliminated by summing the nominator and denominator across loci and alleles), the approximated expectation of

5 Methodology

D is

$$E(D) \approx \frac{E(H_O)}{E(H_E)} - 1 \approx \frac{1}{N_e v - 1}.$$

The effective population size can be estimated by

$$\widehat{N}_e = \frac{\widehat{D} + 1}{\widehat{D} v}.$$

Note that this equation can only be used for polyploid with even ploidy levels. For diploids, Pudovkin's (1996) original equation is used

$$\widehat{N}_e = \frac{1}{2\widehat{D}} + \frac{1}{2\widehat{D} + 2}.$$

For multi-locus and multi-allelic locus, $\widehat{D}$ is calculated by:

$$\overline{D} = \frac{\sum_l \sum_j H_{Olj}}{\sum_l \sum_j H_{Elj}} - 1,$$

where H_{Olj} and H_{Elj} is the observed and expected heterozygosity of the jth allele at the lth locus. A Jackknife method is used to estimate the SE and 95% CI of $\widehat{N}_e$.

$$SE = \sqrt{(L - 1)\text{Var}(\widehat{N}_{e\bar{l}})},$$

where $\widehat{N}_{e\bar{l}}$ is the estimate without using the data at locus l and $\text{Var}(\widehat{N}_{e\bar{l}})$ is its population variance.

Nomura's (2008) molecular coancestry estimator

This estimator assumes a monoecious Wright-Fisher population. Under the assumption of random mating, the expected coancestry coefficient between individuals in generations t is

$$E(\theta_{xyt}) = \frac{1}{N_e}\theta_{xxt-1} + \left(1 - \frac{1}{N_e}\right)\theta_{xyt-1},$$

where $1/N_e$ is the probability that two randomly sampled alleles each from different individuals in generation t come from the same parent, θ_{xxt-1} and θ_{xyt-1} are the average coancestry coefficient within individual and between individuals in the previous generation, respectively.

Nomura (2008) used the parental generation (assuming $t = 0$) as the reference population which ignore the inbreeding within individual ($F_{IS,0} = 0$, equivalent to

5 Methodology

coancestry coefficient within individual $\theta_{xx0} = 1/v$) and the coancestry coefficient between individuals ($\theta_{xy0} = 0$). Under this assumption, the expected coancestry coefficient between offspring is

$$E(\theta_{xy1}) = \frac{1}{N_e}\theta_{xx0} + \left(1 - \frac{1}{N_e}\right)\theta_{xy0} = \frac{1}{vN_e},$$

Such that N_e can be estimated by

$$\widehat{N}_e = \frac{1}{v\widehat{\theta}_{xy1}}.$$

The average parent-based coancestry coefficient between offspring θ_{xy1} can be estimated by modifying Weir (1996) kinship estimator. Hardy (2002) extended Weir's (1996) kinship estimator to include polysomic inheritance, and whose generalized expression is:

$$\widehat{\theta}_{xy1} = \frac{\sum_l \sum_j \widehat{P}_{xlj}\widehat{P}_{ylj} - \sum_l \widehat{M}_l}{L - \sum_l \widehat{M}_l},$$

where $\widehat{P}_{xlj}$ and $\widehat{P}_{ylj}$ are respectively the allele frequency in individual x and y, and $\widehat{M}_l$ is the unbiased estimate of $\sum_j P_{lj}^2$. In order to use the parental generation as the reference population, Nomura (2008) use the molecular coancestry in putative nonrelatives to estimate $\sum_j P_{lj}^2$ (denoted as M_l). M_l is estimated by

$$\widehat{M}_l = \underset{x\text{-}y \in \mathbf{R}_{\bar{l}}}{\text{Avg}} \left(\sum_j \widehat{P}_{xlj}\widehat{P}_{ylj}\right),$$

where $\mathbf{R}_{\bar{l}}$ is the collection of putative nonrelatives at locus l. The putative nonrelatives at locus l is identified from Weir's (1996) kinship estimator from all loci except l:

$$\widehat{\theta}_{xy\bar{l}} = \frac{\sum_{l' \neq l} \sum_j \widehat{P}_{xl'j}\widehat{P}_{yl'j} - \sum_{l' \neq l}\sum_j \widehat{P}_{l'j}^2}{L - 1 - \sum_{l' \neq l}\sum_j \widehat{P}_{l'j}^2}.$$

A predefined proportion of individuals pairs with a smaller $\widehat{\theta}_{xy\bar{l}}$ as putative nonrelatives. A Jackknife method is used to estimate the SE and 95% CI of $\widehat{N}_e$.

Waples & Do's (2010) linkage disequilibrium estimator

This estimator evaluates the linkage disequilibrium with the squared disequilibrium coefficients r_Δ^2 based on Burrows' Δ (Cockerham & Weir 1977). The expectation of r_Δ^2 is strongly affected by the mating system, the recombination fraction c, the effective population size N_e and the sample size n.

Cockerham and Weir's (1977) derived the approximated expectation of r_Δ^2 of diploids under various mating systems: including two monecious mating systems, one allows selfing (say MS), and the other excludes selfing (say ME); and two dioecious mating systems, one with random pairing (say DR) and the other with lifetime pairing (say DH), and f is the average number of females that each male mates. Huang *et al.* (unpublished) extended Cockerham and Weir's (1977) model, and derived the expression of $E(r_\Delta^2)$.

For HS mating system:

$$E(\hat{r}_\Delta^2) \approx \frac{2cv_1 - v_1 - c^2 v}{c_2 c N_e v_1 v} + \frac{1}{n-1}.$$

For MS mating system:

$$E(\hat{r}_\Delta^2) \approx \frac{v^2[4 - 3v + 8c^2 - 14c - cv(2c^2 + 4c - 13) + c_2 cv^2(c+1)]}{c_2 c(cv_2 + v)(3v - 4)(N_e v^2 - 2v^2 + 6v - 4)} + \frac{1}{n-1}.$$

For ME and DR mating system:

$$E(\hat{r}_\Delta^2) \approx \frac{v^2[4 - 3v + 8c^2 - 14c - cv(2c^2 + 4c - 13) + cc_2 v^2(c+1)]}{c_2 c(cv_2 + v)(3v - 4)(N_e v^2 - 4v^2 + 8v - 4)} + \frac{1}{n-1}.$$

For DH mating system:

$$E(\hat{r}_\Delta^2) \approx v^2\{c^3(3+f)v_2 v - (1+f)(3v-4) - c^2[3v^2 - 8 + f(v^2 + 4v - 8)]$$
$$- c[f(2v^2 - 13v + 14) + 3(2v^2 - 7v + 6)]\}$$
$$/[c_2 c(1+f)(cv_2 + v)(3v-4)(N_e v^2 - 4v^2 + 8v - 4)] + \frac{1}{n-1}.$$

where v_i and c_i denotes $v - i$ and $c - i$, respectively.

From the above expressions, N_e can be solved when n, c, v is known. The loci are still assumed to be unlinked. For loci located extreme far in the same chromosome, the recombination fraction c is 0.5 (assuming bivalent pairing); for loci in the different chromosomes, recombination fraction is $1 - 1/v$. Because the corresponding $E(\hat{r}_\Delta^2)$ under these two situations are extreme similar in all mating system (error rate <1.5%), we assume $c = 0.5$ for all loci pairs. The SE and 95% CI is also estimated by the Jackknife method.

5.6 Genetic differentiation

Based on the variance decomposition procedure proposed by Weir & Cockerham (1984),

5 Methodology

either the IAM or the SMM distance between alleles is used for the calculation involving the weighted genotypic F_{ST} estimator. The calculation method is the same as that used for AMOVA, but the individual level is ignored. Some modifications are applied to this estimator so as to accommodate the allelic phenotypes for polyploids (see Section 5.14 for details). The other F_{ST} estimators are calculated based on the allele frequencies. The formulas of several F_{ST} estimators are listed as follows.

- Slatkin's (1995) estimator R_{ST} is calculated by

$$R_{ST} = \frac{\sum_l (S_{Tl} - S_{Wl})}{\sum_l S_{Tl}},$$

where S_{Tl} is the average squared SMM distances within the total population at locus l, symbolically

$$S_{Tl} = \frac{\sum_{j_1 < j_2} v_{tl}^2 \hat{P}_{lj_1} \hat{P}_{lj_2} {}^{\text{SMM}} d_{lj_1 lj_2}^2}{v_{tl}(v_{tl} - 1)/2};$$

and S_{Wl} is the average sum of squares of the SMM distances within each population, symbolically

$$S_{Wl} = \frac{\sum_p \sum_{j_1 < j_2} v_{pl}^2 \hat{P}_{plj_1} \hat{P}_{plj_2} {}^{\text{SMM}} d_{lj_1 lj_2}^2}{\sum_p v_{pl}(v_{pl} - 1)/2},$$

in v_{tl} (or v_{pl}) is the number of allele copies at the lth locus and in the total population (or in the pth population), P_{lj_1} and P_{lj_2} (or P_{plj_1} and P_{plj_2}) are respectively the frequencies of alleles A_{lj_1} and A_{lj_2} in the total population (or in the pth population), and ${}^{\text{SMM}} d_{lj_1 lj_2}$ is the SMM distance between the alleles A_{lj_1} and A_{lj_2}.

- Nei's (1973) estimator G_{ST} is calculated by

$$G_{ST} = \frac{\sum_{l=1}^{L} (H_{Tl} - H_{Sl})}{\sum_{l=1}^{L} H_{Tl}},$$

where H_{Tl} is the expected heterozygosity at the lth locus and in the total population, and H_{Sl} is the weighted average of expected heterozygosity at the lth locus and in all populations, with the number of allele copies in each population as a weight, whose calculating formulas are as follows:

$$H_{Tl} = 1 - \sum_{j=1}^{J_l} \hat{P}_{lj}^2,$$

5 Methodology

$$H_{Sl} = \frac{\sum_{p=1}^{N_{\text{pop}}} v_p \left(1 - \sum_{j=1}^{J_l} \hat{P}_{plj}^2\right)}{v_t},$$

in which P_{lj} (or P_{plj}) is the frequency of the allele A_{lj} and in the total population (or in the pth population), v_p (or v_t) is the number of allele copies in the pth population (or in the total population). Note that monomorphic loci are not accounted.

- Hudson *et al.*'s (1992) estimator F_{ST} is calculated by

$$F_{ST} = \frac{\sum_l (H_{bl} - H_{wl})}{\sum_l H_{bl}},$$

where H_{wl} is the average squared IAM distances between allele copies taken from the same population at locus l, symbolically

$$H_{wl} = \frac{\sum_p \sum_{j_1 < j_2} v_{pl}^2 \hat{P}_{plj_1} \hat{P}_{plj_2} {}^{\text{SMM}} d_{lj_1 lj_2}^2}{\sum_p v_{pl}(v_{pl} - 1)/2},$$

and H_{bl} is the average squared IAM distances between allele copies taken from two different populations at locus l, symbolically

$$H_{bl} = \frac{\sum_{j_1 < j_2} v_{tl}^2 \hat{P}_{lj_1} \hat{P}_{lj_2} {}^{\text{SMM}} d_{lj_1 lj_2}^2 - \sum_p \sum_{j_1 < j_2} v_{pl}^2 \hat{P}_{plj_1} \hat{P}_{plj_2} {}^{\text{SMM}} d_{lj_1 lj_2}^2}{v_{tl}(v_{tl} - 1)/2 - \sum_p v_{pl}(v_{pl} - 1)/2}.$$

- Hedrick's (2005) estimator G'_{ST} is the standardization of Nei's (1973) estimator G_{ST}, which can be calculated by

$$G'_{ST} = \frac{\sum_{l=1}^{L}(H_{Tl} - H_{Sl})(N_{\text{pop}} - 1 + H_{Sl})}{(N_{\text{pop}} - 1)\sum_{l=1}^{L} H_{Tl}(1 - H_{Sl})}.$$

- Jost's (2008) estimator D is calculated by

$$D = \frac{N_{\text{pop}} \sum_{l=1}^{L}(H_{Tl} - H_{Sl})}{(N_{\text{pop}} - 1)\sum_{l=1}^{L}(1 - H_{Sl})}.$$

The test of genetic differentiation is performed by both Fisher's *G*-test and Raymond & Rousset's (1995) Markov Chain test (see Section 5.4 for details). Fisher's *G*-test can be applied to either multiple loci or a single locus. For multiple loci, the statistics (or the

degrees of freedom) will be summed. Raymond & Rousset's (1995) Markov Chain test can only be applied to a single locus.

Huang *et al.*'s (unpublished) F_{ST} estimator

Previous population genetics methods usually assume the samples are independent, and establish the model and estimate the statistics based on such assumption. However, in natural populations, the samples are dependent. In addition to non-independence among individuals (i.e., relatives), subpopulations are also not uniformly differentiated from the ancestral population. Due to the geographical, ecological and historical factors, the gene-flow are unevenly distributed among population pairs. This estimator applied correction for the dependent samples.

This estimator utilize the Nomura's (2008) effective population size estimator to unbiasedly estimate $\sum_j P_{plj}^2$ and $\sum_j P_{tlj}^2$, and the two estimates are denoted as $\hat{M}_{pl}$ and $\hat{M}_{tl}$. They are calculated by

$$\hat{M}_{pl} = \underset{x\text{-}y\in \boldsymbol{R}_{p\bar{l}}}{\text{Avg}} \left(\sum_k \hat{P}_{xlj}\hat{P}_{ylj} \right),$$

$$\hat{M}_{tl} = \underset{x\text{-}y\in \boldsymbol{R}_{t\bar{l}}}{\text{Avg}} \left(\sum_k \hat{P}_{xlj}\hat{P}_{ylj} \right).$$

Nomura's (2008) estimator selected a predefined proportion of individuals pairs as 'nonrelatives'. These putative 'nonrelatives' at locus l is identified from Weir's (1996) kinship estimator from all loci except l, and the collection of such 'nonrelatives' pairs at locus l in subpopulation p or in the total population t are denoted by $\boldsymbol{R}_{p\bar{l}}$ and $\boldsymbol{R}_{t\bar{l}}$. F_{ST} is estimated by

$$\hat{F}_{ST,c} = \frac{\sum_l(\hat{\bar{M}}_{sl} - \hat{M}_{tl})}{\sum_l(1 - \hat{M}_{tl})},$$

where $\hat{\bar{M}}_{sl}$ is the weighted average of $\hat{M}_{pl}$, i.e., $\hat{\bar{M}}_{sl} = \text{Wavg}_{p\in t}\left(\hat{M}_{pl}, N_p^2\right)$.

5.7 Genetic distance

POLYGENE can be used to calculate a variety of genetic distances between two populations, regions and individuals, where all these measures of genetic distances are based on the allele frequencies. The following text details the references and methods.

- Nei's (1972) standard genetic distance D_S: this measure assumes that genetic differences are caused by both mutation and genetic drift. If the mutation rate is constant, the distance D_S between the xth and the yth populations is in proportion to the genetic divergence time, and can be calculated by

$$D_S = -\ln\frac{\hat{J}_{xy}}{\sqrt{\hat{J}_x \hat{J}_y}},$$

where $\hat{J}_x = \sum_{l=1}^{L}\sum_{j=1}^{J_l} \hat{P}_{xlj}^2 /L$, $\hat{J}_y = \sum_{l=1}^{L}\sum_{j=1}^{J_l} \hat{P}_{ylj}^2 /L$ and $\hat{J}_{xy} = \sum_{l=1}^{L}\sum_{j=1}^{J_l} \hat{P}_{xlj}\hat{P}_{ylj} /L$ (J_l is the number of alleles at the lth locus), $\hat{P}_{xlj}$ (or $\hat{P}_{ylj}$) is the frequency of the jth allele A_j at the lth locus and in the xth (or y^{th}) population.

- Cavalli-Sforza's (1967) chord distance D_{CH}: this measure assumes that genetic differences arise only from genetic drift. One major advantage of this measure is that the populations are represented in a hypersphere, the scale of which is one unit per gene substitution. The chord distance D_{CH} in a hyperdimensional sphere is given by

$$D_{CH} = \frac{2}{\pi}\sqrt{2\left(L - \sum_{l=1}^{L}\sum_{j=1}^{J_l}\sqrt{\hat{P}_{xlj}\hat{P}_{ylj}}\right)}.$$

- Reynolds *et al.*'s (1983) distance θ_w: this measure assumes that genetic differentiation occurs only due to genetic drift without any mutation. It is used to estimate the coancestry coefficient θ which provides a measure of genetic divergence, where the expression of θ_w is

5 Methodology

$$\theta_w = \sqrt{\frac{\sum_{l=1}^{L}\sum_{j}^{J_l}\left(\hat{P}_{xlj} - \hat{P}_{ylj}\right)^2}{2\sum_{l=1}^{L}\left(1 - \sum_{j=1}^{J_l}\hat{P}_{xlj}\hat{P}_{ylj}\right)}}\,.$$

- Nei's (1983) distance D_A: this measure assumes that genetic differences arise from both mutation and genetic drift. It is known that such a distance measure gives more reliable population trees than other measures, particularly for microsatellite DNA data. The distance D_A is calculated by

$$D_A = 1 - \frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{J_l}\sqrt{\hat{P}_{xlj}\hat{P}_{ylj}}\,.$$

- Euclidean distance D_{EU}: this is a measure in a Euclidean space with dimension J_l. In this space, the symbol $\left[\hat{P}_{xl1}, \hat{P}_{xl2}, \cdots, \hat{P}_{xlJ_l}\right]$ consisting of the frequencies of alleles in the xth population and at the lth locus denotes a vector. The distance D_{EU} between the xth and the yth populations is the average of the Euclidean distances for the L pairs of vectors from these two populations such that each pair of vectors are at a same locus, that is

$$D_{EU} = \sqrt{\sum_{l=1}^{L}\sum_{j=1}^{J_l}\left(\hat{P}_{xlj} - \hat{P}_{ylj}\right)^2}\,.$$

- Goldstein's (1995) distance $(\delta\mu)^2$: this method was developed based on a stepwise mutation model (SMM), and is specifically used for **microsatellite** markers, whose formula is:

$$(\delta\mu)^2 = \frac{1}{L}\sum_{l=1}^{L}\left(\mu_{xl} - \mu_{yl}\right)^2,$$

where μ_{xl} (or μ_{yl}) is the average size of allele copies at the lth locus and in the xth (or y^{th}) population.

- Nei's (1974) minimum genetic distance D_M: this measure assumes that genetic differences arise from both mutation and genetic drift, whose formula is

$$D_M = \frac{\hat{J}_x + \hat{J}_y}{2} - \hat{J}_{xy},$$

where the definitions of J_x, J_y and $\hat{J}_{xy}$ appear in Nei's (1972) standard genetic distance.

- Roger's (1972) distance D_R: this measure is closely related to the Euclidean distance

D_{EU}, both of which have the relation that $D_R = D_{EU}/\sqrt{2}$, namely

$$D_R = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\sum_{j=1}^{J_l}\frac{\left(\hat{P}_{xlj} - \hat{P}_{ylj}\right)^2}{2}}.$$

- Reynolds *et al.*'s (1983) distance D_{RA}: this measure is defined as the divergence time. If two diploid Wright-Fisher populations of a constant size N have diverged at t generations ago, then the divergence time is $t/2N$. In this situation, Wright's F_{ST} between these two populations can be expressed as $F_{ST} = 1 - \left(1 - \frac{1}{2N}\right)^t$. This can be used to derive that $t/2N \approx -\ln(1 - F_{ST})$. For two Wright-Fisher populations with the same ploidy level v and HS mating system, then the above equation can be expanded as $t/vN \approx -\ln(1 - F_{ST})$. Therefore, the distance D_{RA} can be calculated by

$$D_{RA} = -\ln(1 - F_{ST}).$$

- Slatkin's (1995) linearized distance D_{Sl}: this measure is also defined as the divergence time. Slatkin (1995) considered a simple demographic model. In this model, two diploid populations of size N have diverged from a population of identical size at τ generations ago. Ever since, these two populations have remained isolated, without exchanging any genes. Then the divergence time is $\tau/2N$. On the other hand, because Wright's F_{ST} can be expressed as $F_{ST} = \tau/(\tau + 2N)$, in this model, it is easy to derive that and so $F_{ST}/(1 - F_{ST}) = \tau/(2N)$. In polyploids, this can also be derived that, $F_{ST}/(1 - F_{ST}) = \tau/(vN)$. Therefore, the distance D_{Sl} can be calculated by

$$D_{Sl} = F_{ST}/(1 - F_{ST}).$$

5.8 Ordination analysis

Principal coordinate analysis

PCoA is able ordinate individuals in a new space and make the distance between individuals in the new space is identical to the corresponding distance in the inputting distance matrix **D**.

5 Methodology

Assuming there are n individuals, and each represents a point in a high-dimensional Euclidean space with their distance matrix being $\mathbf{D}$. These points can be placed in a $n-1$ dimension space and ensure their distances matrix are still $\mathbf{D}$. For example, 2 points are in a line, 3 points are in a plane. This process will keep a constant total variance.

Because these n points in the $n-1$ dimension space can further rotate and shift without losing variance. Therefore, to facilitate further dimension reduction, we assume different columns of the new coordinate $\mathbf{X}$ are linearly uncorrelated and each column of $\mathbf{X}$ have been centered. Such that the covariance matrix of $\mathbf{X}$ is a diagonal matrix $\mathbf{V}$, and $\mathbf{1}_{1 \times n}\mathbf{U} = \mathbf{1}_{1 \times n}\mathbf{X} = \mathbf{0}$. $\mathbf{X}$ can be expressed by $\mathbf{U}\sqrt{(n-1)\mathbf{V}}$ where $\mathbf{U}$ is a centered unit vector with a sample variance $1/(n-1)$.

Let $\mathbf{D}' = [d'_{ij}]$ be a $n \times n$ Euclidean distance matrix in the new space. Therefore, the squared Euclidean distance between individuals i and j, d'_{ij}, can be calculated by

$$d'^2_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T = \mathbf{x}_i\mathbf{x}_i^T + \mathbf{x}_j\mathbf{x}_j^T - 2\mathbf{x}_i\mathbf{x}_j^T.$$

The Gram matrix is defined as $\mathbf{G} = \mathbf{X}\mathbf{X}^T = [g_{ij}]$, the above squared distance can be expressed by

$$d'^2_{ij} = g_{ii} + g_{jj} - 2g_{ij}.$$

Substituting $\mathbf{X} = \mathbf{U}\sqrt{(n-1)\mathbf{V}}$ into $\mathbf{G} = \mathbf{X}\mathbf{X}^T$ yields

$$\mathbf{G} = \mathbf{X}\mathbf{X}^T = \mathbf{U}\sqrt{(n-1)\mathbf{V}}\sqrt{(n-1)\mathbf{V}}\mathbf{U}^T = (n-1)\mathbf{U}\mathbf{V}\mathbf{U}^T.$$

Therefore, the new coordinates $\mathbf{X}$ can be solved by performing eigen-value decomposition for $\mathbf{G}$. Now the problem becomes to find a $\mathbf{G}$ that met the constants $d'^2_{ij} = g_{ii} + g_{jj} - 2g_{ij}$. Assuming the distance in the new space d'_{ij} is identical to that in the inputting distance matrix d_{ij}, then $d^2_{ij} = g_{ii} + g_{jj} - 2g_{ij}$. For the same individual, $d_{ii} = 0$ and the equation always holds and does not provide additional constraint. Therefore, the distance matrix $\mathbf{D}$ provides $n(n-1)/2$ constraints but there are $n(n+1)/2$ unknowns in $\mathbf{G}$. Implies that there are insufficient conditional to obtain a unique solution of $\mathbf{G}$. A possible solution is that $\mathbf{G_0} = \mathbf{E} = [e_{ij}] = \left[-\frac{1}{2}d^2_{ij}\right]$, but this solution yields a negative

eigen-value.

The centered $\mathbf{X}$ and $\mathbf{U}$ can provide additional constraints that $\mathbf{1}_{1\times n}\mathbf{U} = \mathbf{0}$, such that $\mathbf{1}_{1\times n}\mathbf{G} = (n-1)\mathbf{1}_{1\times n}\mathbf{U}\mathbf{V}\mathbf{U}^T = \mathbf{0}$. This denotes that $\mathbf{G}$ is also centered, and because $\mathbf{G}$ is symmetric, therefore all columns and rows are centered. This can be done by $\mathbf{G} = \left[g_{ij}\right]$ and $g_{ij} = e_{ij} - e_{i\cdot} - \bar{e}_{\cdot j} + \bar{e}_{\cdot\cdot}$, where $\bar{e}_{i\cdot}$ and $\bar{e}_{\cdot j}$ are respectively the means of ith row and j^{th} column.

Since $\mathbf{G}$ is symmetric and centered, it has at most $n-1$ real eigenvalues, denoted by $\lambda_1, \lambda_2, \cdots, \lambda_p$, and let $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p$ be their corresponding eigenvectors of $\mathbf{G}$. $\mathbf{G}$ can be decomposed as $\mathbf{G} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\mathbf{U} = \left[\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p\right]$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$. The eigen vectors are also centered. It is noteworthy that each eigenvalue λ_i represents a variance that the data are projected into the eigenvector $\boldsymbol{\xi}_i$, and each variance is a non-negative number. Therefore, any negative eigenvalues should be excluded. The transformed coordinate matrix $\mathbf{X}$ can be calculated by $\mathbf{X} = \mathbf{U}\sqrt{\boldsymbol{\Lambda}}$ and the total variance σ_T^2 can be calculated from $\mathbf{X}$.

It does no harm to hypothesize that the first m eigenvalues are the whole non-negative eigenvalues of $\mathbf{G}$. The output matrix $\mathbf{X}'$ can then be expressed as $\mathbf{X}' = \mathbf{U}'\sqrt{\boldsymbol{\Lambda}'}$, i.e.,
$$\mathbf{X}' = \left[\sqrt{\lambda_1}\boldsymbol{\xi}_1, \sqrt{\lambda_2}\boldsymbol{\xi}_2, \cdots, \sqrt{\lambda_m}\boldsymbol{\xi}_m\right],$$
where $\mathbf{U}' = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_m]$ and $\sqrt{\boldsymbol{\Lambda}'} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_m})$. Finally, if all eigenvalues of $\mathbf{G}$ are negative, an output matrix is not produced.

Principal component analysis

If the individuals are originally in a p-dimension space, PCA first placed individuals in a same p-dimension space with different dimension linearly uncorrelated, this can be done by rotation transform, then reduce into a m-dimension space by choosing the first m dimensions with the greatest variances.

5 Methodology

The rotation can be expressed by $\mathbf{X}' = \mathbf{XU}$, where $\mathbf{X}'$ and $\mathbf{X}$ are respectively the new and original coordinate matrix and $\mathbf{U}$ is the rotation matrix. More specifically, $\mathbf{X}_{ij}$ is the allele frequency of individual i of allele j. Therefore $\mathbf{X} = \mathbf{X}'\mathbf{U}^{-1} = \mathbf{X}'\mathbf{U}^T$. Since $\mathbf{X}'$ has been decorrelated, the covariance matrix of $\mathbf{X}'$ is a diagonal matrix $\mathbf{V}$. The covariance matrix of $\mathbf{X}$ is can then be expressed by $\mathbf{C} = \mathbf{UVU}^T$. Therefore, $\mathbf{U}$ and $\mathbf{V}$ can be obtained by performing eigen-value decomposition for the covariance matrix of $\mathbf{X}$.

A coordinate matrix $\mathbf{X} = [x_{ij}]$ can be obtained from the genotype/allelic phenotype data, where x_{ij} is the frequency of jth allele in ith individual/population/region. Therefore, there are $C = \sum_i^L J_l$ columns of matrix $\mathbf{X}$. The variance-covariance matrix $\mathbf{C} = [c_{ij}]$ is calculated for $\mathbf{X}$, where c_{ij} is the covariance between in ith and jth columns of $\mathbf{X}$. Similarly, because $\mathbf{C}$ is symmetric, there are C real eigenvalues. $\mathbf{C}$ can be decomposed as $\mathbf{C} = \mathbf{UVU}^T$, where $\mathbf{U} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_C]$ and $\mathbf{V} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_C)$. It is noteworthy that each eigenvalue λ_i represents the variance that the data are projected into the eigenvector $\boldsymbol{\xi}_i$, and each variance is a non-negative number. The first m greatest eigenvectors and their corresponding eigenvectors are extracted, e.g., $\mathbf{U}' = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_m]$ and $\mathbf{V}' = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m)$. The transformed coordinates are $\mathbf{X}' = \mathbf{XU}'$. The total variance, i.e., the sum of all eigen-vectors is calculated and is denoted as σ_T^2. The variance proportion explained by the first k components is $\sum_i^m \lambda_i / \sigma_T^2$.

5.9 Hierarchical clustering

A hierarchical clustering analysis is based on the genetic distance matrix $\mathbf{D}$ of dimension $n \times n$, where n is the number of clusters. Initially, each individual, each population or each region is defined as a cluster, and the element d_{ij} in $\mathbf{D}$ is the distance from the ith cluster to the jth cluster. The distance matrix $\mathbf{D}$ will be repeatedly updated. The update procedure is described as follows. Assume that the minimum non-diagonal element in $\mathbf{D}$ is d_{ab} ($a \neq b$). First, for the output dendrogram, the two nodes representing the ath and the bth clusters are merged to the node with the coordinate $d_{ab}/2$. Second, update the

5 Methodology

elements in the ath row and the ath column of **D** except for the diagonal element d_{aa}, and the updated elements are written as d'_{ac} and d'_{ca} ($c = 1, 2, \cdots, n$ and $c \neq a$). Third, delete the bth row and the bth column of **D**, such that the type of **D** is reduced to $(n-1) \times (n-1)$. Such a procedure needs to be repeated $n-1$ times, until the type of **D** is reduced to 1×1. Noticing that each distance matrix is symmetric, the updated distances d'_{ac} and d'_{ca} should be equal.

There are several methods to calculate the updated distances d'_{ac} and d'_{ca}, which are listed as follows (here, the symbols N_a, N_b and N_c denote the numbers of the members of the ath, b^{th} and c^{th} clusters, respectively).

- Nearest

$$d'_{ac} = d'_{ca} = \min(d_{ac}, d_{bc}).$$

- Furthest

$$d'_{ac} = d'_{ca} = \max(d_{ac}, d_{bc}).$$

- UPGMA

$$d'_{ac} = d'_{ca} = \frac{N_a d_{ac} + N_b d_{bc}}{N_a + N_b}.$$

- UPGMC

$$d'_{ac} = d'_{ca} = \sqrt{\frac{N_a d_{ac}^2 + N_b d_{bc}^2}{N_a + N_b} - \frac{N_a N_b d_{ab}^2}{(N_a + N_b)^2}}.$$

- WPGMA

$$d'_{ac} = d'_{ca} = \frac{d_{ac} + d_{bc}}{2}.$$

- WPGMC

$$d'_{ac} = d'_{ca} = \sqrt{\frac{d_{ac}^2 + d_{bc}^2}{2} - \frac{d_{ab}^2}{4}}.$$

- WARD

$$d'_{ac} = d'_{ca} = \sqrt{\frac{(N_a + N_c)d_{ac}^2 + (N_b + N_c)d_{bc}^2 - N_c d_{ab}^2}{N_a + N_b + N_c}}.$$

5.10 Individual inbreeding coefficient

The inbreeding coefficient f of an individual is the probability of sampling two IBD alleles from this individual without replacement. The coancestry coefficient θ within an individual is the probability of sampling two IBD alleles from an individual with replacement. According to these definitions, the following expression holds:

$$f = \frac{v\theta - 1}{v - 1}.$$

This expression can be used to covert θ to f if θ is known.

POLYGENE provides the following four methods-of-moment estimators to estimate the coancestry coefficient θ.

- Ritland's (1996a) estimator

$$\hat{\theta}_{\mathrm{RI}} = \frac{\sum_{l=1}^{L}\left[\left(\sum_{j=1}^{J_l} P_{xlj}P_{ylj}/\hat{P}_{lj}\right) - 1\right]}{\sum_{l=1}^{L}(J_l - 1)},$$

  where P_{xlj} (or P_{ylj}) is the frequency of the allele A_{lj} in individual x and y, respectively, and P_{lj} is the frequency of A_{lj} in the reference population.

- Loiselle's (1995) estimator

$$\hat{\theta}_{\mathrm{LO}} = \frac{\sum_{l=1}^{L}\sum_{j}^{J_l}\left(P_{xlj} - \hat{P}_{lj}\right)\left(P_{ylj} - \hat{P}_{lj}\right)}{\sum_{l=1}^{L}(1 - \sum_{j=1}^{J_l}\hat{P}_{lj}^2)},$$

- Weir's (1996) estimator

$$\hat{\theta}_{\mathrm{WE}} = \frac{\sum_{l=1}^{L}\sum_{j=1}^{J_l}\left(P_{xlj}P_{ylj} - \hat{P}_{lj}^2\right)}{\sum_{l=1}^{L}(1 - \sum_{j=1}^{J_l}\hat{P}_{lj}^2)}.$$

- Huang's (unpublished) estimator consider the influence of correlated samples

$$\hat{\theta}_{\mathrm{HU}} = \frac{\sum_{l=1}^{L}\left(\sum_{j=1}^{J_l} P_{xlj}P_{ylj} - \widehat{M}_l\right)}{\sum_{l=1}^{L}(1 - \widehat{M}_l)},$$

  where $\widehat{M}_l$ is the unbiased estimate of homozygosity at locus l, see Section 5.5.

5.11 Individual heterozygosity-index

The heterozygosity-index (H-index) for a genotype $\mathcal{G}$ is defined as the probability of randomly sampling two different alleles within $\mathcal{G}$ without replacement. For diploids, the H-index for a genotype $\mathcal{G}$ is equal to zero if $\mathcal{G}$ is a homozygote, or equal to one if $\mathcal{G}$ is a heterozygote. However, for polyploids, the H-index for a genotype has more than two values ranging from zero to one. POLYGENE enumerates all possible genotypes determining each allelic phenotype, and calculates the H-index for each genotype. The H-index for an individual at a target locus is defined as the weighted average of the H-indices for all genotypes in this individual at this target locus, with the frequency of each genotype in this individual at this target locus as a weight. Moreover, the H-index for an individual is defined as the arithmetic average of the H-indices for this individual across all loci.

5.12 Population assignment

In the population assignment, the likelihoods of individuals for each population are calculated. The likelihood $\mathcal{L}$ of an individual for a target population is the probability of observing the data (i.e., the allelic phenotypes of this individual) under the hypothesis that this individual originated from this target population, whose expression is

$$\mathcal{L} = \prod_{l=1}^{L} \Pr(\mathcal{P}_l),$$

where $\mathcal{P}_l$ is the allelic phenotype of this individual at the lth locus, and $\Pr(\mathcal{P}_l)$ is the frequency of $\mathcal{P}_l$ under the above hypothesis, which is calculated according to the allele frequencies in the target population and a specified inheritance model (e.g., disomic/RCS, PRCS, CES or PES). A higher likelihood implies that this hypothesis is more reliable, i.e., this individual is more likely to have originated from the target population.

Each individual is assigned to the population with the largest likelihood. The difference

between the largest and the second largest common logarithms of likelihoods of an individual in all populations is the LOD score of this individual. Such an LOD score can be used to evaluate the accuracy of the largest likelihood, and the greater the value of the LOD score, the higher the accuracy of the largest likelihood. For example, if the value of the LOD score is up to 3, it means that for the two probabilities of observing allelic phenotypes of this individual in the most possible population and in the second most possible population, the former is $10^3 = 1000$ times that of the latter.

However, the probability of observing allelic phenotypes of an individual will be zero if this individual carries any alleles from outside the target population. Moreover, the individual will be excluded from its true population if at least one of its genotypes is mistyped (e.g., false allele, Taberlet *et al.* 1996).

In POLYGENE, the mistyping rate e is employed into each of the four inheritance models mentioned above, where e is the probability of an allelic phenotype being mistyped. When an allelic phenotype is mistyped, the observed allelic phenotype is randomly drawn according to the allelic phenotypic frequency in the total population. Therefore, the posterior probability $\Pr(\mathcal{P})$ that a given allelic phenotype $\mathcal{P}$ is observed from the pth population is

$$\Pr(\mathcal{P}) = \Pr(\mathcal{P}|p)(1-e) + [1 - \Pr(\mathcal{P}|p)]\, e\, \Pr(\mathcal{P}|t),$$

where $\Pr(\mathcal{P}|p)$ (or $\Pr(\mathcal{P}|t)$) is the frequency of allelic phenotype $\mathcal{P}$ in the pth population (or in the total population).

5.13 Spatial pattern analysis

This method classifies the individual pairs into several distance classes, then calculate the Moran's I coefficient for each distance class, whose formula is

$$I(d) = \frac{\sum_i^n \sum_j^n w_{dij} r_{ij}}{\sum_i^n \sum_j^n w_{dij}}.$$

where d is a distance class, r_{ij} is Hardy and Vekemans (1999) relatedness coefficient using the total population as the reference population, w_{dij} is one if the distance class of

the two individuals is d, otherwise zero. Therefore, Moran's I coefficient for class d, $I(d)$, is the averaged Hardy and Vekemans (1999) relatedness coefficient for individual pairs in distance class d.

A Jackknife method is used to estimate the SE of $I(d)$. The SE of $I(d)$ is calculated by

$$\text{SE} = \sqrt{(L-1)\text{Var}[I_{\bar{l}}(d)]},$$

where $I_{\bar{l}}(d)$ is the estimate of $I(d)$ without the data at locus l and $\text{Var}[I_{\bar{l}}(d)]$ is the population variance of $I_{\bar{l}}(d)$.

5.14 Relationship coefficient

The relationship coefficients include relatedness coefficient r (defined as the Pearson's correlation of allele frequencies between two individuals) and coancestry coefficient θ (defined as the probability of randomly sample IBD alleles between individuals).

POLYGENE provides six relatedness estimators: Huang 2014 MOM (Huang *et al.* 2014), Huang 2015 Likelihood (Huang *et al.* 2015a), Ritland 1996 (r) (Ritland 1996a), Loiselle 1995 (r) (Loiselle *et al.* 1995), Moran's I (r) (Hardy & Vekemans 1999) and Huang unpub (r), and four coancestry coefficient estimators: Ritland 1996 (θ) (Ritland 1996a), Loiselle 1995 (θ) (Loiselle *et al.* 1995), Weir 1996 (θ) (Weir 1996) and Huang unpub (θ).

The formulas of the coancestry coefficient estimators have been listed in Section 5.10. After each coancestry coefficient is calculated, it can be converted to a relatedness coefficient by Equation (8) of Huang *et al.* (2015a). For different ploidy levels, POLYGENE can be used to calculate the relatedness $\hat{r}_{HL}$ from a higher ploidy individual to a lower ploidy individual (Huang *et al.* 2015b). If the ploidy levels of these two individuals are the same, then $\hat{r}_{HL}$ is denoted by $\hat{r}$.

There are two methods to convert a coancestry coefficient into a relatedness coefficient. One is the original conversion, whose formula is $\hat{r}_{HL} = v_{\min}\hat{\theta}$, where $v_{\min}$ is the level of

the lower ploidy individual, and $\hat{\theta}$ is the coancestry coefficient between these two individuals. However, this conversion can only be used for outbred populations and is not used in POLYGENE, and $\hat{r}_{HL}$ may be greater than one. The other is provided by Huang *et al*. (2015a), whose formula is

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

where $\hat{\theta}_{xy}$ is the coancestry coefficient between the xth and y^{th} individuals, $\hat{\theta}_{xx}$ (or $\hat{\theta}_{yy}$) is the coancestry coefficient within the xth (or y^{th}) individual, and $v_{\max}$ is the level of the higher ploidy individual. The latter method can be used for either outbred or inbred populations, with the maximum of $\hat{r}_{HL}$ being equal to 1. This method can be applied to Ritland's (1996) and Loiselle *et al*.'s (1995) estimators, but not to Weir's (1996) estimator because its $\hat{\theta}_{xx}$ maybe zero or negative.

The relatedness coefficient based on Moran's *I* (Hardy & Vekemans 1999) is calculated by

$$r_{xy} = \frac{\sum_{l=1}^{L} \sum_{j=1}^{J_l} \left(\hat{P}_{xlj} - \hat{P}_{lj} \right) \left(\hat{P}_{ylj} - \hat{P}_{lj} \right) + \sum_{l=1}^{L} \sum_{j=1}^{J_l} \text{Var}(\hat{P}_{xlj})/(n_l - 1)}{\sum_{l=1}^{L} \sum_{j=1}^{J_l} \text{Var}(\hat{P}_{xlj})},$$

where $\text{Var}(\hat{P}_{xlj})$ is the variance of individual allele frequency in the reference population, n_l is the number of individuals genotyped at this locus, and $\text{Var}(\hat{P}_{xlj})/(n_l - 1)$ is the sample bias correction.

A Jackknife method is used to estimate the SE of $\hat{r}$. The SE of $\hat{r}$ is calculated by

$$\text{SE} = \sqrt{(L - 1)\text{Var}(\hat{r}_{\bar{l}})}.$$

Where $\hat{r}_{\bar{l}}$ is the relationship estimate without using the data at locus l.

5.15 Heritability estimation

Ritland 1996 MOM estimator

The heritability is estimated by

$$\hat{h}^2 = \frac{\text{Cov}(\hat{\theta}_{xy}, \hat{Z}_x \hat{Z}_y)}{v\widehat{\text{Var}}(\theta_{xy})},$$

where $\hat{\theta}_{xy}$ is the kinship estimate of Ritland (1996a) estimator between two individuals,

5 Methodology

$\hat{Z}_x$ and $\hat{Z}_y$ are respectively the normalized quantitative trait values of the two individuals, v is the ploidy level, and $\widehat{\mathrm{Var}}(\theta_{xy})$ is the variance estimate of actual kinship. $\widehat{\mathrm{Var}}(\theta_{xy})$ is calculated by

$$\widehat{\mathrm{Var}}(\theta_{xy}) = \mathop{\mathrm{Avg}}_{x'<y'} \left(\frac{\hat{\theta}^2_{x'y'} - \sum_l w^2_{x'y'l}\hat{\theta}^2_{x'y'l}}{1 - \sum_l w^2_{x'y'l}} \right) - \mathop{\mathrm{Avg}^2}_{x'<y'}(\hat{\theta}_{x'y'}),$$

where $w_{x'y'l}$ is the unified locus specific weight at the lth locus:

$$w_{xyl} = \frac{J_l - 1}{\sum_{l'}(J_{l'} - 1)}.$$

Such that $\hat{\theta}_{xy} = \sum_l w_{xyl}\hat{\theta}_{xyl}$ and $\sum_l w_{x'y'l} = 1$.

Huang et al. (unpub) MOM estimator

This estimator applied bias correction for inbreeding and sampling correlations based on Nomura (2008).

$$\hat{h}^2 = \hat{\bar{\theta}}_{x,\mathrm{HU}} \frac{\mathrm{Cov}(\hat{\theta}_{xy,\mathrm{HU}}, \hat{Z}_x\hat{Z}_y)}{\widehat{\mathrm{Var}}(\theta_{xy,\mathrm{HU}})}.$$

Where $\hat{\bar{\theta}}_x$ is the average within individual kinship in the population,

$$\hat{\bar{\theta}}_{x,\mathrm{HU}} = \frac{\sum_l \mathrm{Avg}(\sum_k \hat{p}^2_{xlk}) - \sum_l \widehat{M}_l}{\sum_l(1 - \widehat{M}_l)},$$

$\hat{\theta}_{xy,\mathrm{HU}}$ is the kinship estimate between two individuals with bias correction for sample correlations:

$$\hat{\theta}_{xy,\mathrm{HU}} = \frac{\sum_l \sum_k \hat{p}_{xlk}\hat{p}_{ylk} - \sum_l \widehat{M}_l}{\sum_l(1 - \widehat{M}_l)}.$$

Where $\hat{p}_{xlk}$ and $\hat{p}_{ylk}$ are the individual allele frequency, and $\widehat{M}_l$ is the unbiased estimate of homozygosity and can be found in Section 5.5. $\widehat{\mathrm{Var}}(\theta_{xy,\mathrm{HU}})$ is still the variance estimate of actual kinship:

$$\widehat{\mathrm{Var}}(\theta_{xy,\mathrm{HU}}) = \mathop{\mathrm{Avg}}_{x'<y'} \left(\frac{\hat{\theta}^2_{x'y',\mathrm{HU}} - \sum_l w^2_{x'y'l,\mathrm{HU}}\hat{\theta}^2_{x'y'l,\mathrm{HU}}}{1 - \sum_l w^2_{x'y'l,\mathrm{HU}}} \right) - \mathop{\mathrm{Avg}^2}_{x'<y'}(\hat{\theta}_{x'y',\mathrm{HU}}),$$

where $w_{xyl,\mathrm{HU}}$ is the unified weight and is equal to $\frac{1 - \widehat{M}_l}{\sum_{l'}(1 - \widehat{M}_{l'})}$.

Maximum-likelihood estimators

The maximum-likelihood estimators are based on maximize the likelihood

5 Methodology

$$\widehat{h^2} = \arg\max_{h^2} \mathcal{L}(h^2).$$

The likelihood is considered as a function of the heritability

$$\mathcal{L}(h^2) \stackrel{\text{def}}{=} \prod_{x<y} \sum_r \Pr(r) f(V_{xy} \mid r, h^2) \Pr(\mathbf{G}_x\text{-}\mathbf{G}_y \mid r),$$

where r denote a relationship between two individuals, $\Pr(r)$ is the prior probability of r, $\Pr(\mathbf{G}_x\text{-}\mathbf{G}_y \mid r)$ is the likelihood of genotypic/phenotype vectors $\mathbf{G}_x\text{-}\mathbf{G}_y$, and $f(V_{xy} \mid r, h^2)$ is the probability density function of phenotypic variable V_{xy} which is calculated from quantitative trait values.

Mousseau *et al*. (1998) Maximum-likelihood estimator assumes outcrossing, uses $Z_x + Z_y$ as V_{xy} and consider two relationships: nonrelatives and full-sibs. In the outbred populations, $V_{xy} \sim N(0, 2)$ and $N(0, 2 + h^2)$ for nonrelatives and full-sibs. The probability density of V_{xy} can be calculated from the PDF of the normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Thomas et al. (2000) Maximum-likelihood estimator assumes outbreeding, defines V_{xy} as the difference between two quantitative trait values $P_x - P_y$ and further consider half-sibs. Such that, the mean of V_{xy} are zeros for all three relationships, and the variance of V_{xy} are $2\sigma_P^2$, $2\sigma_P^2(1 - h^2/4)$ and $2\sigma_P^2(1 - h^2/2)$ for nonrelatives, half-sibs and full-sibs.

Huang et al. (unpub) maximum-likelihood estimator considers inbreeding and correlated samples, defines V_{xy} as $P_x - P_y$ and consider three relationships: nonrelatives, half-sibs and full-sibs. Such that, the mean of V_{xy} are zeros for all three relationships, and the variance of V_{xy} are $2\sigma_P^2$, $\sigma_P^2[2 - (\frac{1}{2} + f/\tilde{\theta}_x + \frac{1}{2}f^2/\tilde{\theta}_x^2)h^2]$ and $\sigma_P^2[2 - (1 + f/\tilde{\theta}_x)h^2]$ for nonrelatives, half-sibs and full-sibs, where f is the inbreeding coefficient and $\tilde{\theta}_x$ is the averaged kinship coefficient within individuals. They can be estimated from Huang (unpublished) kinship estimator, see Section .

*5.16 Q_{ST} estimation

5.17 Heritability

The heritability is estimated for each population. For Ritland 1996 estimator, the heritability is estimated by

$$h_{\text{RI}}^2 = \frac{\text{Cov}(Z_x Z_y, \hat{\theta}_{xy})}{v \, \widehat{\text{Var}}(\theta_{xy})},$$

Where Z_x and Z_y are the normalized phenotypic value, $\hat{\theta}_{xy}$ is the estimated kinship coefficient using Ritland 1996 estimator, $\widehat{\text{Var}}(\theta_{xy})$ is the estimate of the actual variance of kinship coefficient between individuals.

$$\widehat{\text{Var}}(\theta_{xy}) = \underset{x<y}{\text{Avg}}\left(\frac{\hat{\theta}_{xy}^2 - \sum_l w_{xyl}^2 \hat{\theta}_{xyl}^2}{1 - \sum_l w_{xyl}^2}\right) - \underset{x<y}{\text{Avg}}^2(\hat{\theta}_{xy}),$$

where $\hat{\theta}_{xy}$ is the kinship estimate between individuals x and y, which is an weighted average of locus specific estimate $\hat{\theta}_{xyl}$, and w_{xyl} is the unified locus specific weight.

For Huang MOM estimator, the heritability is estimated by

$$h_{\text{MOM}}^2 = \bar{\hat{\theta}}_x \frac{\text{Cov}(\hat{\theta}_{xy}, \hat{Z}_x \hat{Z}_y)}{\widehat{\text{Var}}(\theta_{xy})}$$

$\bar{\hat{\theta}}_x$ is the estimated average kinship within individuals, it is can be estimated by Huang *et al*. unpublished's estimator,

5.18 Parentage analysis

For the exclusion approach (Zwart *et al.* 2016) in a parentage analysis, the alleged parents are excluded based on the mismatch of genotypes or allelic phenotypes, so it needs to identify the mismatched loci. The identifying procedures are as follows. First, for all known and candidate parents, the possible gamete-types at each locus are extracted. Each genotype (or each gamete-type) is treated as a multi-set. Second, for a parent-

offspring pair, if all gamete-types are not any subsets of the offspring genotypes, then a mismatch error is assigned to this locus. Moreover, for a trio, the set difference $\mathcal{G}_O \setminus g_M$ of multi-sets at a locus is a gamete from the father, where $\mathcal{G}_O$ is an offspring genotype, and g_M is a gamete from the mother. If all possible $\mathcal{G}_O \setminus g_M$ cannot be produced by the alleged father, then another mismatch error is assigned to this locus.

The dominant method (Rodzen *et al.* 2004) converts the allelic phenotype/genotype into pseudo-dominant data: where each allele at a codominant locus is treated as an independent dominant 'locus' with two alleles (dominant and recessive), then use equations for diploids to calculate the likelihoods and LODs (Gerber *et al.* 2000).

For the allelic phenotype method, POLYGENE revise the likelihoods and LODs equations of the traditional likelihood method (Kalinowski *et al.* 2007; Marshall *et al.* 1998), which replaces the genotypic frequency $\Pr(\mathcal{G})$ and the transitional frequency from parent(s) to offspring $T(\mathcal{G}_O|\mathcal{G}_A)$ and $T(\mathcal{G}_O|\mathcal{G}_A, \mathcal{G}_M)$ with the allelic phenotype version: the allelic phenotypic frequency $\Pr(\mathcal{P})$ and the transitional frequency from parent(s) to offspring $T(\mathcal{P}_O|\mathcal{P}_A)$ and $T(\mathcal{P}_O|\mathcal{P}_A, \mathcal{P}_M)$.

In the following, $\mathcal{P}_O$, $\mathcal{P}_A$, $\mathcal{P}_{AM}$, $\mathcal{P}_F$ and $\mathcal{P}_M$ denotes the allelic phenotype of the offspring, the alleged father, the true father and the true mother and the alleged mother; $\mathcal{G}_O$, $\mathcal{G}_A$, $\mathcal{G}_M$, $\mathcal{G}_F$ and $\mathcal{G}_M$ denotes the genotype of these individuals; β denotes the negative amplification rate; e denotes the genotyping error rate; and p_s denotes the sample rate.

When self-fertilization is not considered, for the paternity analysis without known mother, the likelihoods $L(H_1)$ and $L(H_2)$ are

$$\mathcal{L}(H_1) = \Pr(\mathcal{P}_A)\left[(1-e)^2\, T(\mathcal{P}_O|\mathcal{P}_A) + 2e(1-e)\Pr(\mathcal{P}_O) + e^2 \Pr(\mathcal{P}_O)\right],$$

$$\mathcal{L}(H_2) = \Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O),$$

where the hypothesis H_1 is that the alleged father is the true father, and the hypothesis H_2

5 Methodology

is that the alleged father is nonrelatives of the offspring.

For the paternity analysis with known mother, the likelihoods are

$$\mathcal{L}(H_1) = \Pr(\mathcal{P}_M)\Pr(\mathcal{P}_A)\{(1-e)^3\,T(\mathcal{P}_O|\mathcal{P}_A,\mathcal{P}_M)$$
$$+\, e(1-e)^2[T(\mathcal{P}_O|\mathcal{P}_M) + T(\mathcal{P}_O|\mathcal{P}_A) + \Pr(\mathcal{P}_O)] + 3e^2(1-e)\Pr(\mathcal{P}_O)$$
$$+\, e^3\Pr(\mathcal{P}_O)\},$$

$$\mathcal{L}(H_2) = \Pr(\mathcal{P}_M)\Pr(\mathcal{P}_A)\{(1-e)^3\,T(\mathcal{P}_O|\mathcal{P}_M) + e(1-e)^2[T(\mathcal{P}_O|\mathcal{P}_M) + 2\Pr(\mathcal{P}_O)]$$
$$+\, 3e^2(1-e)\Pr(\mathcal{P}_O) + e^3\Pr(\mathcal{P}_O)\},$$

where H_1 and H_2 is the same as above.

For the parent-pair analysis, the likelihoods are

$$\mathcal{L}(H_1) = \Pr(\mathcal{P}_{AM})\Pr(\mathcal{P}_A)\{(1-e)^3\,T(\mathcal{P}_O|\mathcal{P}_A,\mathcal{P}_{AM})$$
$$+\, e(1-e)^2[T(\mathcal{P}_O|\mathcal{P}_{AM}) + T(\mathcal{P}_O|\mathcal{P}_A) + \Pr(\mathcal{P}_O)] + 3e^2(1-e)\Pr(\mathcal{P}_O)$$
$$+\, e^3\Pr(\mathcal{P}_O)\},$$

$$\mathcal{L}(H_2) = \Pr(\mathcal{P}_{AM})\Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O),$$

where H_1 is that the alleged parent-pair is the true father, and the hypothesis H_2 is that the alleged parent-pair is nonrelatives of the offspring.

When the self-fertilization is considered, denoting $T_{S1} = (1-s)\,T(\mathcal{P}_O|\mathcal{P}_A) + s\,T(\mathcal{P}_O|\mathcal{P}_A,\mathcal{P}_A)$, $T_{S2} = (1-s)\,T(\mathcal{P}_O|\mathcal{P}_M) + s\,T(\mathcal{P}_O|\mathcal{P}_M,\mathcal{P}_M)$. The likelihoods for the paternity analysis without known mother are:

$$\mathcal{L}(H_1) = \Pr(\mathcal{P}_A)[(1-e)^2 T_{S1} + 2e(1-e)\Pr(\mathcal{P}_O) + e^2\Pr(\mathcal{P}_O)],$$

$$\mathcal{L}(H_2) = \Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O),$$

For the paternity analysis with known mother, if $A \not\equiv M$, the likelihoods are

5 Methodology

$$\mathcal{L}(H_1) = \Pr(\mathcal{P}_M)\Pr(\mathcal{P}_A)\{(1-e)^3\, T(\mathcal{P}_O|\mathcal{P}_A,\mathcal{P}_M)$$

$$+ e(1-e)^2[T(\mathcal{P}_O|\mathcal{P}_M) + T(\mathcal{P}_O|\mathcal{P}_A) + \Pr(\mathcal{P}_O)] + 3e^2(1-e)\Pr(\mathcal{P}_O)$$

$$+ e^3 \Pr(\mathcal{P}_O)\},$$

$$\mathcal{L}(H_2) = \Pr(\mathcal{P}_M)\Pr(\mathcal{P}_A)\{(1-e)^3 T_{s2} + e(1-e)^2[T_{s2} + 2\Pr(\mathcal{P}_O)] + 3e^2(1-e)\Pr(\mathcal{P}_O)$$

$$+ e^3 \Pr(\mathcal{P}_O)\},$$

if $A \equiv M$, the likelihoods are

$$\mathcal{L}(H_1) = \Pr(\mathcal{P}_A)\{(1-e)^2\, T(\mathcal{P}_O|\mathcal{P}_A,\mathcal{P}_A) + 2e(1-e)\Pr(\mathcal{P}_O) + e^2 \Pr(\mathcal{P}_O)\},$$

$$\mathcal{L}(H_2) = \Pr(\mathcal{P}_A)\{(1-e)^2\, T(\mathcal{P}_O|\mathcal{P}_A) + 2e(1-e)\Pr(\mathcal{P}_O) + e^2 \Pr(\mathcal{P}_O)\},$$

For the parent-pair analysis, if $A \not\equiv AM$, the likelihoods are

$$\mathcal{L}(H_1) = \Pr(\mathcal{P}_{AM})\Pr(\mathcal{P}_A)\{(1-e)^3\, T(\mathcal{P}_O|\mathcal{P}_A,\mathcal{P}_{AM})$$

$$+ e(1-e)^2[T(\mathcal{P}_O|\mathcal{P}_{AM}) + T(\mathcal{P}_O|\mathcal{P}_A) + \Pr(\mathcal{P}_O)] + 3e^2(1-e)\Pr(\mathcal{P}_O)$$

$$+ e^3 \Pr(\mathcal{P}_O)\},$$

$$\mathcal{L}(H_2) = \Pr(\mathcal{P}A_M)\Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O),$$

if $A \equiv AM$, the likelihoods are

$$\mathcal{L}(H_1) = \Pr(\mathcal{P}_A)\{(1-e)^2\, T(\mathcal{P}_O|\mathcal{P}_A,\mathcal{P}_A) + 2e(1-e)\Pr(\mathcal{P}_O) + e^2 \Pr(\mathcal{P}_O)\},$$

$$\mathcal{L}(H_2) = \Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O),$$

The transitional probability $T(\mathcal{P}_O|\mathcal{P}_F)$ and $T(\mathcal{P}_O|\mathcal{P}_F,\mathcal{P}_M)$ are given by

$$T(\mathcal{P}_O|\mathcal{P}_F) = \sum_{\mathcal{G}_F \triangleright \mathcal{P}_F}\sum_{\mathcal{G}_O \triangleright \mathcal{P}_O} \Pr(\mathcal{G}_F|\mathcal{P}_F)\, T(\mathcal{G}_O|\mathcal{G}_F) T(\mathcal{P}_O|\mathcal{G}_O),$$

$$T(\mathcal{P}_O|\mathcal{P}_F,\mathcal{P}_M) = \sum_{\mathcal{G}_F \triangleright \mathcal{P}_F}\sum_{\mathcal{G}_M \triangleright \mathcal{P}_M}\sum_{\mathcal{G}_O \triangleright \mathcal{P}_O} \Pr(\mathcal{G}_F|\mathcal{P}_F)\Pr(\mathcal{G}_M|\mathcal{P}_M)\, T(\mathcal{G}_O|\mathcal{G}_F,\mathcal{G}_M) T(\mathcal{P}_O|\mathcal{G}_O),$$

where $\Pr(\mathcal{G}_F|\mathcal{P}_F)$ (and $\Pr(\mathcal{G}_M|\mathcal{P}_M)$) is the posterior probability of genotype given a allelic phenotype, which is describe above; $T(\mathcal{P}_O|\mathcal{G}_O)$ is the transitional probability from $\mathcal{G}_O$ to $\mathcal{P}_O$, which is calculated by the formula

5 Methodology

$$T(\mathcal{P}|\mathcal{G}) = B_{\mathcal{P}=\emptyset}\beta + B_{G\triangleright\mathcal{P}}(1-\beta).$$

where B_X equals one if the expression X is true, otherwise zero. The transitional probability $T(\mathcal{G}_O|\mathcal{G}_F)$ and $T(\mathcal{G}_O|\mathcal{G}_F,\mathcal{G}_M)$ are calculated by

$$T(\mathcal{G}_O|\mathcal{G}_F) = \sum_{g_F} \Pr(\mathcal{G}_O \setminus g_F)\, T(g_F|\mathcal{G}_F),$$

$$T(\mathcal{G}_O|\mathcal{G}_M,\mathcal{G}_F) = \sum_{g_F} T(\mathcal{G}_O \setminus g_F|\mathcal{G}_M)\, T(g_F|\mathcal{G}_F).$$

where g_F and $\mathcal{G}_O \setminus g_F$ are the genotypes of the sperm and egg that form $\mathcal{G}_O$, $T(g_F|\mathcal{G}_F)$ (and $T(\mathcal{G}_O \setminus g_F|\mathcal{G}_M)$) is the transitional probability from zygote to gamete, and $\Pr(\mathcal{G}_O \setminus g_F)$ is the gamete frequency. The transitional probability from zygote to gamete, the genotypic and allelic phenotypic frequency of zygotes and gametes have been derived by Huang *et al.* (2019).

For the estimation of the genotyping error rate, the known parents and the identified parents at a confidence level (e.g., 99%) is used as a reference. The pair or trio mismatches are used to estimate e.

For a pair mismatch, denote γ as the probability of observing a pair mismatch in a true parent-offspring pair, and denote δ the probability of observing a pair mismatch in a true parent-offspring pair when any individual is erroneously genotyped, and use E to denote the probability that any individual is erroneously genotyped, so $E = 1 - (1-e)^2$. The values of γ and δ can respectively be estimated from reference pairs and from the samples (by Monte-Carlo algorithm). Therefore, the single-locus estimate of E can be calculated by

$$\hat{E} = \frac{\hat{\gamma}}{\hat{\delta}}.$$

The single-locus $\hat{e}$ can be calculated by $\hat{e} = 1 - \sqrt{1-\hat{E}} \approx \hat{E}/2$, and it is variance is $\mathrm{Var}(\hat{e}_l) \approx \frac{e}{2n_{rl}\hat{\delta}_l}$, where n_{rl} is the number of reference pairs or trios at locus l. The multi-locus $\hat{e}$ is the weighted average across locus with $2n_{rl}\hat{\delta}_l$ as the weight

For a trio mismatch, the mismatch in a true trio can be caused by genotyping errors in the offspring or in the parents. If an offspring is erroneously genotyped or both parents are erroneously genotyped, the probability of observing a trio mismatch is equal to the exclusion rate for the third category (say δ_o). If a single parent is erroneously genotyped, the probability of observing a trio mismatch is equal to the exclusion rate for the second category (say δ_p). If any individual in a selfed trio is erroneously genotyped, the probability of observing a trio mismatch is denoted as δ_s. Therefore, the probability of observing mismatch in a true trio γ can be expressed as

$$\gamma = c_1 e + c_2 e^2 + c_3 e^3,$$

$$c_1 = \delta = (1 - s_t)(\delta_o + 2\delta_p) + 2s_t\delta_s,$$

$$c_2 = (1 - s_t)(\delta_o - 4\delta_p) + s_t\delta_s,$$

$$c_3 = (1 - s_t)(2\delta_p - \delta_o).$$

where s_t is the selfing rate in the reference trios. Similarly, the values of γ, δ_o, δ_p and δ_s can respectively be estimated from reference trios and from the samples, and the single-locus estimate of e can be calculated by solving the above equation by neglecting e^3 term:

$$\hat{e} = \frac{\sqrt{\hat{c}_1^2 + 4\hat{c}_2\hat{\gamma}} - c_1}{2\hat{c}_2},$$

The variance of single-locus $\hat{e}$ can be approximately derived by neglecting e^2 and e^3 terms, which is $\text{Var}(\hat{e}_l) \approx \frac{e}{n_{rl}\hat{\delta}_l}$. The multi-locus $\hat{e}$ is the weighted average across locus with $n_{rl}\hat{\delta}_l$ as the weight. The estimate $\hat{e}$ can also be weighed across applications.

The sample rate is estimated from the assignment rate. The assignment rate is the probability that any alleged parent (or alleged parent-pair) in a case is assigned at a confidence level. Denote a_c as the assignment rate when the true parent(s) is sampled, and a_u as the assignment rate when the true parent(s) is not sampled. Such that the actual assignment rate a is a weighted average of a_c and a_u.

5 Methodology

For the paternity analysis with unknown or known mother, $a = p_s a_c + (1 - p_s)a_u$, the sample rate can be estimated by

$$\hat{p}_s = \frac{\hat{a} - \hat{a}_u}{\hat{a}_c - \hat{a}_u}.$$

For the parent-pair analysis with known sexes, $a = p_s^2 a_c + (1 - p_s^2)a_u$, the sample rate can be estimated by

$$\hat{p}_s = \sqrt{\frac{\hat{a} - \hat{a}_u}{\hat{a}_c - \hat{a}_u}}.$$

For the parent-pair analysis with unknown sexes, the probability that the true parents are collected is $p_c = s_u p_s + (1 - s_u)p_s^2$, where s_u is the selfing rate in this application. Hence $a = p_c a_c + (1 - p_c)a_u$, the sample rate can be estimated by

$$\hat{p}_s = \frac{\hat{s}_u - \sqrt{\hat{s}_u^2 + 4\hat{p}_c - 4\hat{s}_u \hat{p}_c}}{2\hat{s}_u - 2}.$$

where $\hat{p}_c = \frac{\hat{a} - \hat{a}_u}{\hat{a}_c - \hat{a}_u}$. The value of $\hat{s}_u$ is estimated by

$$\hat{s}_u = \frac{n_{s,0.99} + n_{s,0.95} + n_{s,0.80}}{n_{a,0.99} + n_{a,0.95} + n_{a,0.80}},$$

where $n_{s,0.99}$, $n_{s,0.95}$ and $n_{s,0.80}$ are the number of assigned selfed cases at 99%, 95% and 80% confidence levels, respectively, and $n_{s,0.99}$, $n_{s,0.95}$ and $n_{s,0.80}$ are the number of assigned cases at these confidence levels, respectively.

The estimated $\hat{p}_s$ is truncated to the range of [0.1]. The value of a is assumed to be drawn from $U(a_u, a_c)$, and a numerical method is used to obtain the variance of $\hat{p}_s$. The inverse of the variance of $\hat{p}_s$ will be used to obtain the weighted average of $\hat{p}_s$ among different confidence levels and different applications. POLYGENE uses three confidence levels, 80%, 95% and 99%, to obtain $\hat{p}_s$ for each application.

Alternative method to obtain thresholds of Δ

5 Methodology

Previous likelihood-based methods use a Monte-Carlo simulation to obtain the threshold of Δ. However, this is inapplicable for some extreme conditions. For example, when the sample rate (the probability of the true parent been sampled) is too high (e.g., 100%) or too low (e.g., 50%), the simulation is unable to find the thresholds of Δ to ensure the correct assignment rate being equal to 80%, 95%, 99% or 100%. This method will be automatically applied if extreme conditions are detected.

Because the LOD of an alleged parent (or parent pair) is the sum of locus-specific LOD across loci. Therefore, according to the central limit theorem, LOD is approximately accord with a normal distribution if there are enough loci ($L > 30$). The Monte-Carlo simulations can be used to estimate the distribution of the LOD of the true parent and the false parent. It is now possible to calculate the posterior probability of correct assignment. Using paternity analysis as an example, there are four types of results:

1. True father is not sampled;

2. True father is sampled and whose LOD is the highest;

3. True father is sampled and whose LOD is the second highest;

4. True father is sampled and whose LOD is not the highest and the second highest;

The joint probability $\Pr(\Delta = x, \text{Type} = T)$ for these types are

$$\Pr(\Delta = x, \text{Type} = 1) = C(1 - p_s)n(n - 1) \int_{-\infty}^{\infty} f_F(u + x)f_F(u)F_F(u)^{n-2}du,$$

$$\Pr(\Delta = x, \text{Type} = 2) = Cp_s(n - 1) \int_{-\infty}^{\infty} f_T(u + x)f_F(u)F_F(u)^{n-2}du,$$

$$\Pr(\Delta = x, \text{Type} = 3) = Cp_s(n - 1) \int_{-\infty}^{\infty} f_F(u + x)f_T(u)F_F(u)^{n-2}du,$$

$$\Pr(\Delta = x, \text{Type} = 4) = Cp_s(n - 1)(n - 2) \int_{-\infty}^{\infty} f_F(u + x)f_F(u)F_T(u)F_F(u)^{n-3}du.$$

These integrals can be calculated by numerical methods. The posterior probability of each type can be calculated using the Bayes equation:

$$\Pr(\text{Type} = T | \Delta = x) = \frac{\Pr(\Delta = x | \text{Type} = T) \Pr(\text{Type} = T)}{\Pr(\Delta = x)} = \frac{\Pr(\Delta = x, \text{Type} = T)}{\sum_{T'} \Pr(\Delta = x, \text{Type} = T')}$$

The posterior probability of correct assignment is the sum of the posterior probabilities of the correct types:

$$\Pr(\text{Correct} | \Delta = x) = \sum_{T \text{ is correct}} \Pr(\text{Type} = T | \Delta = x)$$

The threshold of Δ is the value of Δ with a typical correct assignment rate (e.g., 80%, 95%, 99% and 99.9%). A down-hill simplex algorithm is used to find the thresholds of Δ by minimizing $[\Pr(\text{Correct} | \Delta = x) - \text{Target}]^2$.

5.19 Analysis of molecular variance

Our framework for AMOVA supports any level of ploidy and any number of hierarchies. Huang *et al.* (2021) developed four methods for the multi-locus genotypic/allelic phenotypic data, in which the former three are the method-of-moment estimators, named the homoploid and the anisoploid methods as well as the weighting genotypic method. The fourth is the maximum-likelihood method. The differences among these methods are the calculation speed, the weighting scheme, the strategy for handling any missing data and the method to generate a new dataset in a permutation test. In the following text, the generalized framework for AMOVA is first described, by using the homoploid method as an example.

In the generalized framework AMOVA, the procedures for the homoploid method are roughly as follows: (i) calculating the genetic distance between alleles or haplotypes; (ii) calculating the sums of squares (SS); (iii) solving the variance components; (iv) calculating the *F*-statistics; and (v) performing a permutation test.

Genetic distance

5 Methodology

For the homoploid method, the dummy haplotypes are extracted from a allelic phenotype. Let $\mathcal{P}_l$ be a allelic phenotype at the lth locus, and let A_{hl} be the allele in the hth dummy haplotype extracted from $\mathcal{P}_l$. Denote $\Pr(A_{hl} = A_{lj})$ (or P_{hlj} for short) for the probability that A_{hl} is the jth allele at the lth locus (denoted by A_{lj}). Then

$$P_{hlj} = \Pr(A_{hl} = A_{lj}) = \sum_{G \triangleright \mathcal{P}_l} \Pr(G | \mathcal{P}_l) \Pr(A_{hl} = A_{lj} | G).$$

For missing data, $\hat{P}_{hlj}$ refers the allele frequency of A_{lj} in the population that the hth dummy haplotype is sampled.

The distance between two dummy haplotypes can be calculated under the IAM or the SMM model. For multi-locus data, the square of the distance $d_{hh'}$ between the hth and the h'^{th} dummy haplotypes extracted from $\mathcal{P}_l$ is the following sum:

$$d_{hh'}^2 = \sum_{l=1}^{L} \sum_{j_1=1}^{J_l} \sum_{j_2=1}^{J_l} \hat{P}_{hlj_1} \hat{P}_{h'lj_2} d_{lj_1 j_2}^2,$$

where J_l is the number of alleles at the lth locus, $d_{lj_1 j_2}$ is the distance between the alleles A_{lj_1} and A_{lj_2}. For the IAM model, $d_{lj_1 j_2}^2 = 1$ if $A_{lj_1} \neq A_{lj_2}$, otherwise $d_{lj_1 j_2}^2 = 0$. For the SMM model, $d_{lj_1 j_2}^2 = \left(S_{lj_1} - S_{lj_2}\right)^2$ in which S_{lj_1} and S_{lj_2} are respectively the sizes of A_{lj_1} and A_{lj_2}. Moreover, if A_{lj_1} or A_{lj_2} is the null allele A_y, then $d_{lj_1 j_2}^2$ is defined as the average of squares of all SMM distances between the visible alleles at the lth locus, that is

$$d_{lj_1 j_2}^2 = \frac{2}{(J_l - 1)(J_l - 2)} \sum_{\substack{1 \leq i < j \leq J_l \\ A_{li}, A_{lj} \neq A_y}} d_{lij}^2.$$

Sums of squares

The sums of squares (SS) of the distances within all individuals (WI), within all populations (WP), within all regions (WR) or in the total population (TOT) can be calculated by

$$SS_{WI} = \sum_{i=1}^{N_I} \sum_{1 \leq h_i < h_i' \leq v_{Ii}} \frac{d_{h_i h_i'}^2}{v_{Ii}},$$

$$SS_{WP} = \sum_{i=1}^{N_P} \sum_{1 \leq h_i < h_i' \leq v_{Pi}} \frac{d_{h_i h_i'}^2}{v_{Pi}},$$

5 Methodology

$$SS_{WR} = \sum_{i=1}^{N_R} \sum_{1 \le h_i < h_i' \le v_{Ri}} \frac{d_{h_i h_i'}^2}{v_{Ri}},$$

$$SS_{TOT} = \sum_{1 \le h < h' \le v_T} \frac{d_{hh'}^2}{v_T},$$

where N_I (N_P or N_R) is the number of all individuals (all populations or all regions), v_{Ii} (v_{Pi}, v_{Ri} or v_T) is the number of dummy haplotypes within the ith individual (the ith population, the ith region or the total population), and $d_{h_i h_i'}$ is the distance between the h_i^{th} and the $h_i'^{th}$ dummy haplotypes.

Relations between expected SS and variance components

The relationships between the expected sums of squares of distances and the variance components for various hierarchies can be expressed as

$$E(SS_{WI}) = \sigma_{WI}^2(N_H - N_I),$$

$$E(SS_{WP}) = \sigma_{WI}^2(N_H - N_P) + \sigma_{AI}^2 \left(N_H - \sum_{i=1}^{N_I} \sum_{j=1}^{N_P} \frac{v_{Ii}^2}{v_{Pj}} \right),$$

$$E(SS_{WR}) = \sigma_{WI}^2(N_H - N_R) + \sigma_{AI}^2 \left(N_H - \sum_{i=1}^{N_I} \sum_{j=1}^{N_R} \frac{v_{Ii}^2}{v_{Rj}} \right) + \sigma_{AP}^2 \left(N_H - \sum_{i=1}^{N_P} \sum_{j=1}^{N_R} \frac{v_{Pi}^2}{v_{Rj}} \right),$$

$$E(SS_{TOT}) = \sigma_{WI}^2(N_H - 1) + \sigma_{AI}^2 \left(N_H - \sum_{i=1}^{N_I} \frac{v_{Ii}^2}{v_T} \right) + \sigma_{AP}^2 \left(N_H - \sum_{i=1}^{N_P} \frac{v_{Pi}^2}{v_T} \right) + \sigma_{AR}^2 \left(N_H - \sum_{i=1}^{N_R} \frac{v_{Ri}^2}{v_T} \right),$$

where σ_{WI}^2 is the variance of distances between the dummy haplotypes within all individuals, and σ_{AI}^2 (σ_{AP}^2 or σ_{AR}^2) is the variance of distances between the dummy haplotypes among all individuals (among all populations or among all regions).

According to the concept that each member of a hierarchy is a 'vessel' of genes, the above relational expressions can be rewritten as

$$E(SS_i) = \sum_{j=1}^{i} \sigma_j^2 \left(|V_4| - \sum_{V_i} \sum_{V_{j-1} \in V_i} \frac{|V_{j-1}|^2}{|V_i|} \right), \quad i = 1, 2, 3, 4,$$

where SS_1, SS_2, SS_3, SS_4 denote in turn $SS_{WI}, SS_{WP}, SS_{WR}, SS_{TOT}$; $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ denote in turn $\sigma_{WI}, \sigma_{AI}, \sigma_{AP}, \sigma_{AR}$; V_0, V_1, V_2, V_3, V_4 denote in turn an allele, an individual, a

population, a region, the total population; $|V_i|$ denotes the number of all V_{i-1} in V_i; and the mobile subscript V_i is taken from all vessels of the ith hierarchy.

Extending to M hierarchies

Imitating the previous method, for any positive integer M, it can extend the AMOVA to M hierarchies, i.e., the relations between the expected SS_i and the variance component σ_i $(i = 1, 2, \cdots, M)$ can be expressed as follows:

$$\mathrm{E}(\mathrm{SS}_i) = \sum_{j=1}^{i} \sigma_j^2 \left(|V_M| - \sum_{V_i} \sum_{V_{j-1} \in V_i} \frac{|V_{j-1}|^2}{|V_i|} \right), \quad i = 1, 2, \cdots, M,$$

where V_M denotes the vessel of highest hierarchy, i.e., the total population; when i ranges from 0 to M, the corresponding vessels represent in turn an allele, an individual, a population, a region I (of the third hierarchy), a region II (of the fourth hierarchy), and so on; similarly, the mobile subscript V_i is taken from all vessels of the ith hierarchy. Moreover, if the within individual hierarchy is ignored, then V_1 denotes a population, V_2 denotes a region I, and etc.

The above relational expressions can be expressed as the form of matrices as follows:

$$\mathbf{S} = \mathbf{C}\boldsymbol{\Sigma},$$

where $\mathbf{S} = [\mathrm{E}(SS_1), \mathrm{E}(SS_2), \cdots, \mathrm{E}(SS_M)]^T$, $\boldsymbol{\Sigma} = [\sigma_1^2, \sigma_2^2, \cdots, \sigma_M^2]^T$, and the coefficient matrix $\mathbf{C}$ is a lower triangular matrix of type $M \times M$, whose ij^{th} element C_{ij} is

$$C_{ij} = \begin{cases} |V_M| - \sum_{V_i} \sum_{V_{j-1} \in V_i} \frac{|V_{j-1}|^2}{|V_i|} & \text{if } i \geq j, \\ 0 & \text{if } i < j. \end{cases}$$

A method-of-moment estimation of variance components is given by $\widehat{\boldsymbol{\Sigma}} = \mathbf{C}^{-1}\widehat{\mathbf{S}}$. After that, the F-statistics can be solved by

$$\widehat{F}_{ij} = 1 - \frac{\sum_{k=1}^{i} \widehat{\sigma}_k^2}{\sum_{k=1}^{j} \widehat{\sigma}_k^2}, \quad 1 \leq i < j \leq M.$$

5 Methodology

Genetic differentiation test

Following Excoffier *et al.* (1992), the differentiation test is performed for each F_{ij}. The null hypothesis is $F_{ij} = 0$, which is equivalent to that the vessels j are real but vessels i are artificial. To obtain the null distribution of $\hat{F}_{ij}$, vessels $i - 1$ within vessels j are randomly permuted to generate the new datasets. For each generated dataset, the $\hat{F}_{ij}$ is estimated by the same procedures. Similarly, the probability that $\hat{F}_{ij}$ is greater than the original value is used as a single-tailed *P*-value.

Difference among methods

For the **homoploid method**, all loci are treated as a dummy locus, with the dummy haplotypes extracted from the allelic phenotypes. The allele frequency in missing data refers the that in a population so as to calculate the genetic distance between haplotypes, which assumes $F_{IS} = 0$ and can underestimate F_{IS}. The **p(type) correction** can eliminate such bias and is performed by

$$\hat{F}'_{IS} = \frac{\hat{F}_{IS}}{p_{\text{type}}}.$$

Where p_{type} is the genotyping rate, and the values of SS and MS are also corrected accordingly to allow consistency. For a permutation test, the dummy haplotypes are randomly permuted to generate a new dataset, and the probability that the variance components or the *F*-statistics for the permuting dataset are greater than the original values is calculated, which will be defined as the *P* value.

For the **anisoploid method**, the dummy alleles are extracted instead of the dummy haplotypes, and the genetic distance and the SS are calculated for each locus. The global $\hat{\mathbf{S}}_g$ (or the global $\mathbf{C}_g$) is the sum of matrices $\hat{\mathbf{S}}$ (or matrices $\mathbf{C}$) over all loci, and the global $\hat{\mathbf{\Sigma}}_g$ is solved from the formula $\hat{\mathbf{\Sigma}}_g = \mathbf{C}_g^{-1} \hat{\mathbf{S}}_g$. For this method, any missing data are excluded in the calculating process. Therefore, the matrix $\mathbf{C}$ may be different at different loci due to the missing data (or to the anisoploids, although POLYGENE does not support anisoploids).

5 Methodology

For a permutation test, the alleles are permuted locus-by-locus.

For the **weight genotype method**, neither the dummy alleles nor the dummy haplotypes are extracted. Instead, the SS between or within allelic phenotypes is weighted according to the posterior probabilities of genotypes hidden behind an allelic phenotype. The global $\hat{\mathbf{S}}_g$ (or the global $\mathbf{C}_g$) is also the sum of matrices $\hat{\mathbf{S}}$ (or matrices $\mathbf{C}$) over all loci. For each locus, the SS_i is calculated by

$$SS_i = \sum_{V_i} \frac{1}{|V_i|} \left[\sum_{j=1}^{N} d^2(\mathcal{P}_j) + \sum_{1 \le j < k \le N} d^2(\mathcal{P}_j, \mathcal{P}_k) \right],$$

where V_i is any vessel of the ith hierarchy, N is the number of individuals within V_i and genotyped at this locus, $d^2(\mathcal{P}_j)$ and $d^2(\mathcal{P}_j, \mathcal{P}_k)$ are respectively the weighted sums of squares of distances within the allelic phenotype $\mathcal{P}_j$ and between the allelic phenotypes $\mathcal{P}_j$ and $\mathcal{P}_k$, whose calculation formulas are

$$d^2(\mathcal{P}_j) = \sum_{G \triangleright \mathcal{P}_j} \Pr(G | \mathcal{P}_j) \, d^2(G),$$

$$d^2(\mathcal{P}_j, \mathcal{P}_k) = \sum_{G_1 \triangleright \mathcal{P}_j} \sum_{G_2 \triangleright \mathcal{P}_k} \Pr(G_1 | \mathcal{P}_j) \Pr(G_2 | \mathcal{P}_k) \, d^2(G_1, G_2).$$

In the above formulas, $d^2(G)$ and $d^2(G_1, G_2)$ are the sums of squares of allele distances within G and between G_1 and G_2, respectively, that is

$$d^2(G) = \sum_{i<j} d_{A_i A_j}^2, \qquad d^2(G_1, G_2) = \sum_{i,j} d_{A_{1i} A_{2j}}^2,$$

where A_i and A_j are two alleles in G, and A_{1i} (or A_{2j}) is an allele in G_1 (or G_2). For a permutation test using this method, the dataset is generated from the allele frequencies in the total population based on the hypothesis that there is no differentiation. The remaining steps are the same as for the homoploid method.

For the **maximum-likelihood method**, a reverse procedure is used, where the F-statistics are firstly estimated, and then the variance components and other statistics are solved from the estimated F-statistics. Due to the constraint $(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$, all F-

5 Methodology

statistics can be obtained from $F_{12}, F_{13}, \ldots, F_{1M}$.

The global likelihood for individuals at the ith hierarchy is the product of frequencies of all allelic phenotypes conditional on $\mathbf{p}_{V_i l}$ and F_{1i}, symbolically

$$\mathcal{L}_i = \prod_{V_i} \prod_{l=1}^{L} \prod_{j=1}^{N_{V_i}} \Pr\left(\mathcal{P}_{V_i l j} | \mathbf{p}_{V_i l}, F_{1i}\right), \qquad i = 2, 3, \cdots, M,$$

where V_i is taken from all vessels of the ith hierarchy, N_{V_i} is the number of individuals in V_i, $\mathcal{P}_{V_i l j}$ is the allelic phenotype of jth individual in V_i and at the lth locus, $\mathbf{p}_{V_i l}$ is the vector consisting of the frequencies of all alleles in V_i and at the lth locus, $\Pr\left(\mathcal{P}_{V_i l j} | \mathbf{p}_{V_i l}, F_{1i}\right)$ is the frequency of $\mathcal{P}_{V_i l j}$ conditional on $\mathbf{p}_{V_i l}$ and F_{1i}, that is

$$\Pr\left(\mathcal{P}_{V_i l j} | \mathbf{p}_{V_i l}, F_{1i}\right) = \sum_{\mathcal{G} \triangleright \mathcal{P}_{V_i l j}} \Pr\left(\mathcal{G} | \mathbf{p}_{V_i l}, F_{1i}\right).$$

Under the disomic/RCS model, if the inbreeding is considered, $\Pr\left(\mathcal{G} | \mathbf{p}_{V_i l}, F_{1i}\right)$ is calculated by

$$\Pr\left(\mathcal{G} | \mathbf{p}_{V_i l}, F_{1i}\right) = \binom{|V_1|}{c_1, c_2, \cdots, c_{J_l}} \prod_{j=1}^{J_l} \prod_{k=0}^{c_j - 1} (\alpha_{1ij} + k) \bigg/ \prod_{k'=0}^{|V_1|-1} (\alpha_{1i} + k'),$$

where c_j in the multinomial coefficient is the number of the jth allele copies in $\mathcal{G}$ ($j = 1, 2, \cdots, J_l$), $\alpha_{1i} = 1/F_{1i} - 1$ and $\alpha_{1ij} = \alpha_{1i} \hat{P}_{V_i l j}$ ($\hat{P}_{V_i l j}$ is the jth element in $\mathbf{p}_{V_i l}$). Because the true value of $\mathbf{p}_{V_i l}$ is unavailable, the estimated $\hat{\mathbf{p}}_{V_i l}$ is used as $\mathbf{p}_{V_i l}$ in the calculating process. A down-hill simplex algorithm (Nelder & Mead 1965) can be used to find the optimal F_{1i} ($i = 2, 3, \cdots, M$).

The variance components can be solved from the F-statistics with the following additional constraint:

$$\text{E}(\text{SS}_M) = \sum_{j=1}^{M} \sigma_j^2 \left(|V_M| - \sum_{V_{j-1} \in V_M} \frac{|V_{j-1}|^2}{|V_i|}\right).$$

The SS_M under IAM model can be obtained from the allele frequencies in the total population V_M, that is

$$SS_M = |V_M| \sum_{1 \le i < j \le J} \hat{P}_{Mi} \hat{P}_{Mj} d^2_{A_i A_j},$$

where $\hat{P}_{Mi}$ (or $\hat{P}_{Mj}$) is the frequency of A_i (or A_j) in V_M. Finally, the procedure of permutation test is the same as the weight genotype method.

5.20 Bayesian clustering

MCMC algorithm: Pritchard *et al*. (2000) used a Markov Chain Monte-Carlo (MCMC) algorithm with Gibbs's sampling to infer the population structure. In the MCMC algorithm, various parameters are repeatedly updated by using both the allele frequencies in each cluster and its individual originating clusters so as to obtain the clustering results. The state of this system can be described by **P**, **Q**, Z_i, Z_{ihl}, r, α, λ, η, γ and F, where **P** and **Q** are the vectors consisting of some parameters, the meanings of the vectors **P** and **Q** together with the parameters Z_i and Z_{ihl} will be explained in this section, and the meanings of the other parameters can be seen from Section 4.19. Each new state is generated based on the current state. The sequence consisting of these states is called a Markov chain. In an iteration, all parameters listed above are updated in turn, with each being updated conditional on all other parameters. Such a process is called a Gibbs sampling. The Markov chain starts from a randomly generated initial state, and the state after several iterations becomes stable and independent to the initial state. This iteration period is called a 'burnin' or a 'dememorization', after which this algorithm begins to record the number of times that an allele or an individual is assigned to each cluster. Such a number of iterations is called a 'sampling period'. In order to prevent the results in two adjacent iterations being too similar, the state will be recorded at an interval called the 'thinning interval' so as to eliminate any autocorrelation. Some independent runs (with different random number generator seeds) can be performed so as to repeat the results or to avoid the Markov chain being blocked at local maxima. The number of independent runs is called the 'number of runs'.

5 Methodology

Vector P of allele frequencies: let $A_{l1}, A_{l2}, \cdots, A_{lJ_l}$ be all distinct alleles at the lth locus, and let P_{klj} be the frequency of A_{lj} in the kth cluster, $j = 1, 2, \cdots, J_l$. Denote **P** for the vector $[P_{kl1}, P_{kl2}, \ldots, P_{klJ_l}]$. If **P** is updated as **P'**, it means that the elements $P'_{kl1}, P'_{kl2}, \ldots, P'_{klJ_l}$ in **P'** are randomly drawn from the Dirichlet distribution

$$\mathcal{D}\left(n_{kl1} + \lambda, n_{kl2} + \lambda, \ldots, n_{klJ_l} + \lambda\right)$$

for the non-F model; or from

$$\mathcal{D}\left(n_{kl1} + \varepsilon_{l1} f_k, n_{kl2} + \varepsilon_{l2} f_k, \ldots, n_{klJ_l} + \varepsilon_{lJ_l} f_k\right)$$

for the F model, where λ is the Dirichlet parameter of allele frequencies, n_{klj} is the number of copies of the allele A_{lj} that is assigned to the kth cluster, ε_{lj} is the frequency of A_{lj} in the ancestral clusters, and $f_k = (1 - F_k)/F_k$. The F model together with the parameter F_k will be described in the penultimate paragraph of this section. The parameter λ can be used to prevent the allele frequencies becoming fixed at zero or one. When λ is large, it will result in the allele frequencies becoming more evenly distributed, and thus will reduce any differentiation among clusters and slow the convergence of allele frequencies.

Dirichlet parameter λ of allele frequencies: if the option 'Infer lambda' is checked, then λ will be updated by the Metropolis-Hastings approach. The updated value λ' is randomly drawn from the normal distribution $N(\lambda, \text{std}^2(\lambda))$. If λ' is below zero or above $\max(\lambda)$, it is rejected directly. Otherwise, it is accepted at the probability of $\min(1, E)$, where E is

$$E = \prod_{l=1}^{L} \prod_{k=1}^{K} \left[\frac{\Gamma(J_l \lambda')[\Gamma(\lambda)]^{J_l}}{\Gamma(J_l \lambda)[\Gamma(\lambda')]^{J_l}} \left(\prod_{j=1}^{J_l} \hat{P}_{klj} \right)^{\lambda' - \lambda} \right]$$

for the non-F model, or

$$E = \prod_{l=1}^{L} \left[\frac{\Gamma(J_l \lambda')[\Gamma(\lambda)]^{J_l}}{\Gamma(J_l \lambda)[\Gamma(\lambda')]^{J_l}} \left(\prod_{j=1}^{J_l} \varepsilon_{lj} \right)^{\lambda' - \lambda} \right]$$

for the F model, in which K is the number of clusters, $\Gamma(\cdot)$ is the gamma function, and $\hat{P}_{klj}$ is an element in **P**.

5 Methodology

Moreover, if the option 'Diff λ' is checked, this enables the algorithm to use different values of λ for different clusters. If the option 'Diff λ' is checked, the updated value λ'_k of λ for the kth cluster will be randomly drawn from the normal distribution $N(\lambda_k, \text{std}^2(\lambda))$, whose accepting probability E_k is as follows:

$$E_k = \prod_{l=1}^{L} \left[\frac{\Gamma(J_l \lambda'_k)[\Gamma(\lambda_k)]^{J_l}}{\Gamma(J_l \lambda_k)[\Gamma(\lambda'_k)]^{J_l}} \left(\prod_{j=1}^{J_l} \hat{P}_{klj} \right)^{\lambda'_k - \lambda_k} \right], \qquad k = 1, 2, \cdots, K.$$

Originating clusters of individuals: for the non-ADMIXTURE model, it is assumed that an individual can only originate from one cluster, and all alleles in this individual are assigned to this cluster. Denote Z_i for the current originating cluster of the ith individual. The updated cluster Z'_i of Z_i is randomly drawn from all K clusters, and the accepting probability, denoted by $\Pr(Z'_i = k)$, that Z'_i is equal to the kth cluster is

$$\Pr(Z'_i = k) = \frac{\prod_{l=1}^{L} \prod_{h=1}^{v_i} \sum_{j=1}^{J_l} \hat{P}_{klj} \hat{P}_{ihlj}}{\sum_{k'=1}^{K} \prod_{l=1}^{L} \prod_{h=1}^{v_i} \sum_{j=1}^{J_l} \hat{P}_{k'lj} \hat{P}_{ihlj}},$$

in which $i = 1, 2, \cdots, N$, $k = 1, 2, \cdots, K$, v_i is the ploidy level of ith individual, each $\hat{P}_{klj}$ is an element in $\mathbf{P}$, and $\hat{P}_{ihlj}$ is the posterior probability of A_{lj} in h^{th} allele copy in i^{th} individual. The probability $\hat{P}_{ihlj}$ is calculated from dummy haplotypes. For the genotype data, the value of $\hat{P}_{ilhj}$ is either zero or one, whilst for the allelic phenotype data, the value of $\hat{P}_{ilhj}$ ranges from zero to one.

For the ADMIXTURE model, it is assumed that different alleles in the same individual can originate from different clusters. Let Z_{ihl} be the current originating cluster of the hth allele copy in the ith individual and at the lth locus. Like the situation of non-ADMIXTURE model, the updated cluster Z'_{ihl} of Z_{ihl} is randomly drawn from all K clusters. Using the elements in $\mathbf{P}$ and $\mathbf{Q}$, the accepting probability that Z'_{ihl} is equal to the kth cluster is

$$\Pr(Z'_{ihl} = k) = \frac{Q_{ik} \sum_{j=1}^{J_l} \hat{P}_{klj} \hat{P}_{ihlj}}{\sum_{k'=1}^{K} \left(Q_{ik'} \sum_{j=1}^{J_l} \hat{P}_{k'lj} \hat{P}_{ihlj} \right)}$$

where $h = 1, 2, \cdots, J$, $i = 1, 2, \cdots, N$, $l = 1, 2, \cdots, L$, $k = 1, 2, \cdots, K$, and the meanings of $\mathbf{Q}$

5 Methodology

together with its elements is explained in the next paragraph.

Vector Q of admixture proportions： let Q_{ik} denote the admixture proportion of the ith individual genome that originates from the kth cluster, $k = 1, 2, \cdots, K$, and let **Q** denote the vector $[Q_{i1}, Q_{i2}, \cdots, Q_{iK}]$. As stated previously, **Q** is used in the ADMIXTURE model (Pritchard *et al.* 2000) to update the originating cluster Z_{ihl}. The admixture proportions $Q_{i1}, Q_{i2}, \cdots, Q_{iK}$ are randomly drawn from the Dirichlet distribution

$$\mathcal{D}(m_{i1} + \alpha_1, m_{i2} + \alpha_2, \dots, m_{iK} + \alpha_K),$$

where m_{ik} is the number of allele copies at all loci and in the ith individual assigned to the kth cluster, and α_k is the Dirichlet parameter for the kth cluster, $k = 1, 2, \cdots, K$. These alphas can be used to prevent the proportions becoming fixed. The higher the values of these alphas, the higher the mixture level. Additionally, $Q_{i1}, Q_{i2}, \cdots, Q_{iK}$ are also updated via a Metropolis-Hastings approach at a probability of 0.5. Such an update will improve the mixing when these alphas are small, and will shuffle the individuals to prevent the Markov chain becoming blocked at the local maxima. These alphas are also used to update **Q**. The updated values $Q'_{i1}, Q'_{i2}, \cdots, Q'_{iK}$ are randomly drawn from the Dirichlet distribution

$$\mathcal{D}(\alpha_1, \alpha_2, \dots, \alpha_K)$$

for the non- LOCPRIORI model, or from

$$\mathcal{D}\left(\alpha_{\text{local},s1}, \alpha_{\text{local},s2}, \dots, \alpha_{\text{local},sK}\right)$$

for the LOCPRIORI model, where $\alpha_{\text{local},sk}$ is a local alpha which is used for sampling the individuals from both the sth location and the kth cluster, $k = 1, 2, \cdots, K$. These updated values $Q'_{i1}, Q'_{i2}, \cdots, Q'_{iK}$ are accepted at the probability of $\min(1, E)$, where

$$E = \prod_{h=1}^{v_i} \prod_{l=1}^{L} \frac{\sum_{k=1}^{K} \sum_{j=1}^{J_l} Q'_{ik} \hat{P}_{klj} \hat{P}_{ihlj}}{\sum_{k=1}^{K} \sum_{j=1}^{J_l} Q_{ik} \hat{P}_{klj} \hat{P}_{ihlj}}.$$

Dirichlet parameter α for allele frequencies: this parameter will be updated by a Metropolis-Hastings approach if the option 'Infer alpha' is checked in the LOCPRIORI model or during the admburnin period. For non-LOCPRIORI model, if the values of α for

5 Methodology

the whole clusters are assumed to be all equal, then the updated value α' is randomly drawn from the normal distribution $\mathrm{N}(\alpha, \mathrm{std}^2(\alpha))$, and α' is accepted at the probability of $\min(1, E)$, where

$$E = \frac{\Pr(\alpha')}{\Pr(\alpha)} \prod_{i=1}^{N} \left[\frac{\Gamma(K\alpha')}{\Gamma(K\alpha)} \prod_{k=1}^{K} \left(\frac{\Gamma(\alpha)}{\Gamma(\alpha')} Q_{ik}^{\alpha'-\alpha} \right) \right],$$

in which N is the number of individuals, $\Pr(\alpha)$ is the *a priori* probability of α, and the value of $\frac{\Pr(\alpha')}{\Pr(\alpha)}$ is equal to 1 if α is assumed to be drawn from a uniform distribution, or equal to $\left(\frac{\alpha'}{\alpha}\right)^{A-1} \exp\left(\frac{\alpha-\alpha'}{B}\right)$ if α is assumed to be drawn from a gamma distribution with A and B as the parameters. Moreover, if the values of α for the whole clusters are assumed to be not all equal, then the updated value α_k' for the kth cluster obeys the normal distribution $\mathrm{N}(\alpha_k, \mathrm{std}^2(\alpha))$, and so α_k' is randomly drawn from this distribution, and α_k' is accepted at the probability of $\min(1, E_k)$, where

$$E_k = \frac{\Pr(\alpha_k')}{\Pr(\alpha_k)} \prod_{i=1}^{N} \frac{\Gamma\left(d + \sum_{k'=1}^{K} \alpha_{k'}\right)\Gamma(\alpha_k)}{\Gamma\left(\sum_{k'=1}^{K} \alpha_{k'}\right)\Gamma(\alpha_k')} Q_{ik}^{d}, \qquad k = 1, 2, \cdots, K,$$

in which $d = \alpha_k' - \alpha_k$, and the meanings of $\Pr(\alpha_k)$ and $\frac{\Pr(\alpha_k')}{\Pr(\alpha_k)}$ are similar to $\Pr(\alpha)$ and $\frac{\Pr(\alpha')}{\Pr(\alpha)}$, respectively. For the LOCPRIORI model, the parameter α will be updated by using an alternative approach (see the next paragraph for details).

LOCPRIORI model: in this model, the population information of individuals is used as the *a prior* information to assist our clustering, which is powerful when the population differentiation is relatively weak (Hubisz *et al.* 2009). For the non-ADMIXTURE model, the vectors $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}_s$ are used, where the kth element η_k in $\boldsymbol{\eta}$ is the *a priori* probability of individuals assigned to the kth cluster, and the kth element γ_{sk} in $\boldsymbol{\gamma}_s$ is the *a priori* probability of individuals sampled from the sth location and assigned to the kth cluster, $k = 1, 2, \cdots, K$. For the admixture model, the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_{\mathrm{local}}$ are used, which consist of the global alphas and the local alphas, respectively, where the kth element α_k in $\boldsymbol{\alpha}$ (or $\alpha_{\mathrm{local},sk}$ in $\boldsymbol{\alpha}_{\mathrm{local}}$) reflects the relative level of admixture that the kth cluster is relative to all individuals (or to the individuals sampled from the sth location), $k = 1, 2, \cdots, K$.

5 Methodology

Moreover, the parameter r related to the informativeness of data parameterizes the extent to which the ancestral proportions at the locations of individuals being sampled can deviate from the overall proportion. If r is high (>>1), the prior ancestry proportions at all locations are essentially the same (i.e., it is approximately proportional to η_k or α_k). In contrast, if r is near one or lower, the values of those γ_{sk} or of those $\alpha_{\text{local},sk}$ may vary substantially at various locations, implying that the location data are informative about ancestry.

Update of LOCPRIORI model + non-ADMIXTURE model: all data of parameters for LOCPRIORI are updated by using a Metropolis-Hastings approach. The new value r' of the parameter r is randomly drawn from the uniform distribution $\mathrm{U}(r - \mathrm{eps}(r), r + \mathrm{eps}(r))$, which is checked within the range [0,1], whose acceptance rate E for the non-ADMIXTURE model is

$$E = \prod_{s=1}^{S}\left[\frac{\Gamma(r')}{\Gamma(r)}\prod_{k=1}^{K}\left(\frac{\Gamma(r\eta_k)}{\Gamma(r'\eta_k)}\gamma_{sk}^{\eta_k(r'-r)}\right)\right],$$

where S is the number of sampling locations. Next, the meaning of updating the vector $\boldsymbol{\eta}$ (or $\boldsymbol{\gamma}_s$) is that only needs to update two elements in $\boldsymbol{\eta}$ (or in $\boldsymbol{\gamma}_s$). The updating procedures are as follows. Firstly, randomly sample two elements from $\boldsymbol{\eta}$ (or $\boldsymbol{\gamma}_s$), denoted by η_a and η_b (or γ_{sk} and γ_{sl}). Secondly, randomly draw a difference variable d from the uniform distribution $\mathrm{U}(0, \mathrm{eps}(\eta))$ for $\boldsymbol{\eta}$, or from $\mathrm{U}(0, \mathrm{eps}(\gamma))$ for $\boldsymbol{\gamma}_s$. Finally, the updated values η_a' and η_b' for $\boldsymbol{\eta}$ are given by

$$\eta_a' = \eta_a + d,$$
$$\eta_b' = \eta_b - d.$$

The two updated values are checked within the range [0,1], whose acceptance rate E is

$$E = \prod_{s=1}^{S}\left[\frac{\Gamma(r\eta_a)\Gamma(r\eta_b)}{\Gamma(r\eta_a')\Gamma(r\eta_b')}\left(\frac{\gamma_{sa}}{\gamma_{sb}}\right)^{rd}\right].$$

Moreover, the updated values γ_{sk}' and γ_{sl}' for $\boldsymbol{\gamma}_s$ are given by

$$\gamma_{sk}' = \gamma_{sk} + d,$$
$$\gamma_{sl}' = \gamma_{sl} - d.$$

5 Methodology

The two updated values are checked within the range [0,1], whose acceptance rate E_s is

$$E_s = \left(\frac{\gamma'_{sk}}{\gamma_{sk}}\right)^{r\eta_k - 1 - N_{sk}} \left(\frac{\gamma'_{sl}}{\gamma_{sl}}\right)^{r\eta_l - 1 - N_{sl}},$$

where N_{sk} (or N_{sl}) is the number of the individuals sampled from the sth location and assigned to the kth (or l^{th}) cluster, $s = 1, 2, \cdots, S$.

Update of LOCPRIORI model + ADMIXTURE model: for these models, the new value r' of the parameter r is sampled from the same uniform distribution as described in the preceding paragraph, whose acceptance rate E is

$$E = \prod_{k=1}^{K} \prod_{s=1}^{S} \left[\frac{r'^{r'\alpha_k}}{r^{r\alpha_k}} \frac{\Gamma(r\alpha_k)}{\Gamma(r'\alpha_k)} \alpha_{\text{local},sk}^{\alpha_k d} \exp\left(-d\alpha_{\text{local},sk}\right)\right],$$

where $d = r' - r$. Next, the kth updated value α'_k of the global alphas is randomly drawn from the normal distribution $\text{N}(\alpha_k, \text{std}^2(\alpha))$, whose acceptance rate E_k is

$$E_k = \prod_{s=1}^{S} \left[\alpha_{\text{local},sk}^{r(\alpha'_k - \alpha_k)} \frac{\Gamma(\alpha_k r)}{\Gamma(\alpha'_k r)} r^{r(\alpha'_k - \alpha_k)}\right], \qquad k = 1, 2, \cdots, K.$$

This is followed by the kth updated value $\alpha'_{\text{local},sk}$ of the local alphas, which is randomly drawn from the normal distribution $\text{N}\left(\alpha_{\text{local},sk}, \text{std}^2(\alpha)\right)$, whose acceptance rate E_{sk} is

$$E_{sk} = \left(\frac{\alpha'_{\text{local},sk}}{\alpha_{\text{local},sk}}\right)^{r\alpha_k - 1} \left[\frac{\Gamma\left(d + \sum_{k'=1}^{K} \alpha_{\text{local},sk'}\right)\Gamma(\alpha_{\text{local},sk})}{\Gamma\left(\sum_{k'=1}^{K} \alpha_{\text{local},sk'}\right)\Gamma(\alpha'_{\text{local},sk})}\right]^{N_s} \left(\prod_{i=1}^{N_s} Q_{ik}\right)^{d} \exp(-rd),$$

where $d = \alpha'_{\text{local},sk} - \alpha_{\text{local},sk}$, and N_s is the number of individuals sampled from the sth location, $s = 1, 2, \cdots, S, k = 1, 2, \cdots, K$.

F model: in this model, it is assumed that all K clusters have undergone independent drift away from an ancestral cluster, and their allele frequencies are correlated. The measure F in this model is analogous to Wright's F_{ST}. For the kth cluster, the differentiation from the ancestral cluster is measured by F_k, where F_k is the value of F for the kth cluster, $k = 1, 2, \cdots, K$. The allele frequencies of the kth cluster at the lth locus are drawn from the Dirichlet distribution

$$\mathcal{D}\left(\varepsilon_{l1} f_k, \ \varepsilon_{l2} f_k, \dots, \ \varepsilon_{lJ_l} f_k\right), \qquad k = 1, 2, \cdots, K,$$

where $\varepsilon_{l1}, \ \varepsilon_{l2}, \dots, \ \varepsilon_{lJ_l}$ are the frequencies of alleles in the ancestor clusters and at the lth

5 Methodology

locus, and $f_k = (1 - F_k)/F_k$. The values of F are also updated. If the values of F for the whole clusters are assumed to be not all equal, the new value F'_k is randomly drawn from the normal distribution $N(F_k, \text{std}^2(F))$, which is accepted at the probability of $\min(1, E_k)$, where

$$E_k = \left(\frac{F'_k}{F_k}\right)^{\mu^2/\sigma^2 - 1} \exp\left(\frac{\mu(F_k - F'_k)}{\sigma^2}\right) \prod_{l=1}^{L} \frac{\Gamma(f'_k) \prod_{j=1}^{J_l} \Gamma(f_k \varepsilon_{lj}) \hat{P}_{klj}^{f'_k \varepsilon_{lj}}}{\Gamma(f_k) \prod_{j=1}^{J_l} \Gamma(f'_k \varepsilon_{lj}) \hat{P}_{klj}^{f_k \varepsilon_{lj}}}, \qquad k = 1, 2, \cdots, K,$$

in which μ and σ are the *a priori* mean and the *a priori* standard deviation of F, and $f'_k = (1 - F'_k)/F'_k$. Moreover, if the values of F for the whole clusters are assumed to be equal, the new value F' of F is randomly drawn from the normal distribution $N(F, \text{std}^2(F))$, whose acceptance rate E is

$$E = \left(\frac{F'}{F}\right)^{\mu^2/\sigma^2 - 1} \exp\left(\frac{\mu(F - F')}{\sigma^2}\right) \prod_{k=1}^{K} \prod_{l=1}^{L} \frac{\Gamma(f') \prod_{j=1}^{J_l} \Gamma(f \varepsilon_{lj}) \hat{P}_{klj}^{f' \varepsilon_{lj}}}{\Gamma(f) \prod_{j=1}^{J_l} \Gamma(f' \varepsilon_{lj}) \hat{P}_{klj}^{f \varepsilon_{lj}}}.$$

Updating the allele frequencies of ancestral clusters: two approaches are available for selection to update the allele frequencies of the ancestral clusters, each of which is selected at the probability 0.5. For the first approach, the updated values $\varepsilon'_{l1}, \varepsilon'_{l2}, \cdots, \varepsilon'_{lJ_l}$ of $\varepsilon_{l1}, \varepsilon_{l2}, \cdots, \varepsilon_{lJ_l}$ are randomly sampled from the Dirichlet distribution

$$\mathcal{D}\left(\lambda + \sum_{k=1}^{K} \hat{P}_{kl1} f_k, \lambda + \sum_{k=1}^{K} \hat{P}_{kl2} f_k, \cdots, \lambda + \sum_{k=1}^{K} \hat{P}_{klJ_l} f_k\right), \qquad l = 1, 2, \cdots, L,$$

and the acceptance rate $E_{lj'}$ is

$$E_{lj'} = \prod_{j=1}^{J_l} \left[\left(\frac{\varepsilon_{lj}}{\varepsilon'_{lj}}\right)^{\sum_{k=1}^{K} \hat{P}_{klj} f_k} \prod_{k=1}^{K} \frac{\Gamma(f_k \varepsilon_{lj})}{\Gamma(f_k \varepsilon'_{lj})} \hat{P}_{klj}^{f(\varepsilon'_{lj} - \varepsilon_{lj})}\right], \qquad l = 1, 2, \cdots, L, j' = 1, 2, \cdots, J_l.$$

For the second approach, this only needs to be updated to the frequencies of the two randomly chosen alleles. Let ε_{la} and ε_{lb} be these two frequencies. Next, a difference variable d is drawn from the uniform distribution $U(0, N^{-1/2})$, where N is the number of individuals. Then the updated values ε'_{la} and ε'_{lb} are respectively

$$\varepsilon'_{la} = \varepsilon_{la} + d,$$

$$\varepsilon'_{lb} = \varepsilon_{lb} - d.$$

Now, if ε'_{la} or ε'_{lb} is not in the acceptable range $[0, 1]$, then both are rejected, and the original ε_{la} and ε_{lb} are regarded as the updated values. If ε'_{la} and ε'_{lb} are in the acceptable

range $[0, 1]$, then both are accepted at the acceptance rate E, where

$$E = \left(\frac{\varepsilon'_{la}\varepsilon'_{lb}}{\varepsilon_{la}\varepsilon_{lb}}\right)^{\lambda-1} \prod_{k=1}^{K} \left[\frac{\Gamma(f_k\varepsilon_{la})\Gamma(f_k\varepsilon_{lb})}{\Gamma(f_k\varepsilon'_{la})\Gamma(f_k\varepsilon'_{lb})} \left(\frac{\hat{P}_{kla}}{\hat{P}_{klb}}\right)^{f_k d}\right].$$

5.21 Migration rate estimation

The migration rate estimation is based on extending Wilson and Rannala (2003) to polysomic inheritance. There are two likelihoods, the likelihood of genotype/phenotype vector $\mathbf{X}$, and the likelihood of individual migration ancestor vector $\mathbf{M}$ and migration generation $\mathbf{t}$. The migration ancestor M can be any population including the sampling population, and the migration generation has three possible values: 0 for native, 1 for first generation immigrant, and 2 for second generation immigrant. Besides, the necessary variables include the sampling population vector $\mathbf{S}$, the inbreeding coefficient vector $\mathbf{f}$, and the allele frequency vector $\mathbf{p}$. The number of populations is denoted as N_P.

The first likelihood is given by

$$\Pr(\mathbf{X}|\mathbf{S}; \mathbf{M}, \mathbf{t}, \mathbf{f}, \mathbf{p}) = \prod_{i=1}^{n} \prod_{l=1}^{L} \Pr(X_{il}|S_i; M_i, t_i, \mathbf{F}, \mathbf{p}),$$

where $\Pr(X_{il}|S_i; M_i, t_i, \mathbf{F}, \mathbf{p})$ is the frequency of the genotype/phenotype of individual i at locus l, X_{il}, and it is given by a piecewise function:

$$\Pr(X_{il}|S_i; M_i, t_i, \mathbf{f}, \mathbf{p}) = \begin{cases} \Phi(X_{il}, S_i) & \text{if } M_i = S_i \text{ and } t = 0, \\ \Phi(X_{il}, M_i) & \text{if } M_i \neq S_i \text{ and } t = 1, \\ p_{AA}\Phi(X_{il}, S_i) + p_{BB}\Phi(X_{il}, M_i) \\ +p_{AB}\varphi(X_{il}, S_i, M_i) & \text{if } M_i \neq S_i \text{ and } t = 2. \end{cases}$$

$\Phi(X, s)$ is the frequency of X in subpopulation s, and $\varphi(X, s, r)$ is the frequency of X of hybrid between subpopulation s and r. p_{NN} and p_{FF} is the probability that two alleles are both native and foreign, p_{FN} is the probability that one allele is native and the other is foreign. In diploids, $p_{NN} = p_{FF} = 0$, and $p_{FN} = 1$.

For diploids, $\Phi(X, s)$ and $\varphi(X, s, r)$ are given by

5 Methodology

$$\Phi(X, s) = \begin{cases} f_s P_{sA} + (1 - f_s) P_{sA}^2 & \text{if } X = AA, \\ 2(1 - f_s) P_{sA} P_{sB} & \text{if } X = AB. \end{cases}$$

$$\varphi(X, s, r) = \begin{cases} P_{rA} P_{sA} & \text{if } X = AA, \\ P_{rA} P_{sB} + P_{rB} P_{sA} & \text{if } X = AB. \end{cases}$$

The second likelihood is given by the probability mass function of a multinomial distribution. $\mathbf{M}$ and $\mathbf{t}$ can be combined and define an additional variable $\mathbf{q}$ (termed individual source), which has $2N_P - 1$ possible values, coded from 1 to $2N_P - 1$, and the probability of each value can be calculated from the migration rate matrix $\mathbf{m}$, denotes as P_q:

$$P_q = \begin{cases} 3m_{ss} - 2 & \text{if } S = s, M = s \text{ and } t = 0, \\ m_{sr} & \text{if } S = s, M = r \text{ and } t = 1, \\ 2m_{sr} & \text{if } S = s, M = r \text{ and } t = 2. \end{cases}$$

The probability of observing $\mathbf{M}$ and $\mathbf{t}$ given $\mathbf{m}$ is

$$\Pr(\mathbf{M}, \mathbf{t} | \mathbf{m}) = \prod_s^{N_p} \binom{n_s}{n_{s1}, n_{s2}, \dots, n_{s(2N_p-1)}} \prod_{q=1}^{2N_p-1} P_q^{n_{sq}},$$

where n_s is the number of individuals in population s, n_{sq} is the number of individuals been assigned to source q.

MCMC

If the 'Fix likelihood to 1' is used, then only the second likelihood is used to calculate the acceptance ratio, otherwise their product is used. But both likelihoods are recorded. The algorithm is described as follows:

Step 0. Initialize. The inbreeding coefficient are set as zero, $m_{sr} = 1/(3N_p)$, $m_{ss} = 1 - (N_p - 1)/(3N_p)$, the initial allele frequency is copies from allele frequency estimates, all individuals are assigned as natives.

Step 1, update $\mathbf{M}, \mathbf{t}$. Randomly select a population s, randomly select an individual source q that existing in the individuals in s, randomly selected an individual i with the source q, and randomly selected another individual source q'. The new source q' is accepted at a

probability of $\min(1, L'/L)$, where L' and L are respectively the new and the current likelihood.

Step 2, update **m.** Randomly select a population s and a foreign population r. The new migration rate m'_{sr} is drawn from a uniform distribution $U(m_{sr} - \Delta_M/2, m_{sr} + \Delta_M/2)$. To allow the sum of $m'_{s.}$ being equal to one, m_{ss} should also be updated as $m'_{ss} = 1 - \sum_{r \neq s} m_{sr}$. The two migration rates m'_{sr} and m'_{ss} are accepted at a probability of $\min(1, L'/L)$.

Step 3. update **p.** Randomly select a population s, a locus l, and an allele a at locus l. The new allele frequency p'_{sla} is drawn from a uniform distribution $U(p_{sla} - \Delta_P/2, p_{sla} + \Delta_P/2)$. Similarly, to allow the sum of allele frequency being equal to one, the frequencies of the remaining alleles are also updated: $p'_{slb} = p_{slb}(1 - p'_{sla})/(1 - p_{sla})$. The new allele frequency vector $\mathbf{p}'_{sl}$ is accepted at a probability of $\min(1, L'/L)$.

Step 4. update **F.** Randomly select a population s. The new inbreeding coefficient f'_s is randomly drawn from a uniform distribution $U(f_s - \Delta_F/2, f_s + \Delta_F/2)$. The new inbreeding coefficient f'_s is accepted at a probability of $\min(1, L'/L)$.

Step 5. update missing genotype. Randomly sample a missing genotype X. The new genotype X' is randomly generated according to the Hardy-Weinberg equilibrium assuming the allele frequencies are equal. The new genotype X' is accepted at a probability of $\min(1, L'/L)$.

Step 6, record. If current iteration is in the burnin period, then only record the two likelihoods for plotting the likelihood line chart. Otherwise, record the parameters including migration rate matrix $\mathbf{m}$, individual source $\mathbf{q}$ (equivalent to $\mathbf{M}, \mathbf{t}$), allele frequency $\mathbf{p}$, inbreeding coefficient $\mathbf{f}$.

Step 7, if the number of iterations reach the limit, then exit the loop, otherwise goto step 1.

5 Methodology

Dummy diploid genotype methods

During initialization, the dummy diploid genotypes are generated by two steps: firstly, randomly sample a genotype from an allelic phenotype according to the posterior probability $\Pr(\mathcal{G}|\mathcal{P})$; secondly, randomly sample two alleles from the genotype and create a diploid dummy genotype.

The two genotypic frequency functions $\Phi(X, s)$ and $\varphi(X, s, r)$ can still be used in calculating the likelihood. However, the values of p_{AA}, p_{AB} and p_{BB} in $\varphi(X, s, r)$ should be revised.

The dummy genotype is generated by randomly sampling two allele copies from the genotype without replacement. When sampling two allele copies from a hybrid (e.g., tetraploid *AABB*, hexaploid *AAABBB*), the probability that both allele copies are native (p_{AA}) or foreign (p_{BB}) is

$$p_{AA} = p_{BB} = \frac{\binom{v/2}{2}}{\binom{v}{2}} = \frac{v-2}{4v-4}.$$

The probability that one is native and the other is foreign is

$$p_{AB} = \frac{(v/2)^2}{\binom{v}{2}} = \frac{v}{2v-2}.$$

Genotype methods

During initialization, the genotype data are generated by randomly sampling a genotype for each allelic phenotype according to the posterior probability $\Pr(\mathcal{G}|\mathcal{P})$. The genotype for missing data is randomly generated according to RCS model assuming no inbreeding. For phenotype method, this genotype is further converted into allelic phenotype.

In the update of missing data, the new genotype/allelic phenotype is randomly generated with the same method describe above, and is accepted at a probability of $\min(1, L'/L)$.

5 Methodology

The genotypic/allelic phenotype frequencies are used in calculating the likelihood. The genotypic frequency for non-hybrids $\Phi(X, s)$ is replaced with the $\Pr(\mathcal{G}, f)$ for genotype methods or $\Pr(\mathcal{P}, f)$ for phenotype method (see Section 5.1).

For genotype methods, X is a genotype and the genotypic frequency of hybrid $\varphi(X, s, r)$ should be revised as

$$\varphi(X, s, r) = \sum_{g_1 \uplus g_2 = X} \Pr(g_1|s) \Pr(g_2|r).$$

Where the genotype X is a multiset with v elements, g_1 and g_2 are gamete-types, and $g_1 \uplus g_2 = X$ denotes their union set is X. $\Pr(g_1|s)$ and $\Pr(g_2|r)$ can be calculated by $\Pr(\mathcal{G}, f)$.

For phenotype methods, X is an allelic phenotype and $\varphi(X, s, r)$ should be revised as

$$\varphi(X, s, r) = \sum_{g_1 \uplus g_2 \rhd X} \Pr(g_1|s) \Pr(g_2|r).$$

Where $g_1 \uplus g_2 \rhd X$ denotes the union set of g_1 and g_2 is a genotype that determining X. $\Pr(g_1|s)$ and $\Pr(g_2|r)$ can be calculated by $\Pr(\mathcal{P}, f)$.

Variable Genotype

For variable genotype, the genotypes are not only randomly sampled at step 0, but also updated in each iteration. This step is inserted between Step 4 and Step 5.

This is performed by: randomly sample a non-missing genotype X, and randomly generate the new genotype (or dummy genotype) X' with the same method in step 0, and is accepted at a probability of $\min(1, L'/L)$.

5.22 Mantel tests

The Mantel test is used to measure the degree of association between two dissimilarity

matrices (Mantel 1967). For example, the significant Mantel correlation between the genetic matrix and a barrier matrix would indicate that the genetic structure of a population is correlated with a specific isolation-by-barrier hypothesis.

Partial Mantel tests (Smouse & Sokal 1986) are used to measure the effect of a matrix on the genetic dissimilarity matrix whilst the effect(s) of an additional partial matrix (analogous to the partial correlation coefficients) are statistically controlled. Partial Mantel tests are expanded to allow for multiple additional partial matrices.

By expanding the method described in (Smouse & Sokal 1986), the effects of k additional partial matrices $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k$ of type $n \times n$ can be used to control the effect of $\mathbf{X}$ on $\mathbf{Y}$, where $\mathbf{X}$ is a distance matrix of type $n \times n$, and $\mathbf{Y} = \beta_0 \mathbf{X} + \sum_{i=1}^{k} \beta_i \mathbf{Z}_i + \boldsymbol{\varepsilon}$, in which those betas are the regression coefficients, and $\boldsymbol{\varepsilon}$ is a random error matrix. A recursion method is used in the calculating process, such that one of these partial matrices will be reduced at each step. The kth partial correlation coefficient $r(\mathbf{X}, \mathbf{Y}|\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k)$ between $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k$ can be calculated from the following recursion formula:

$$r(\mathbf{X}, \mathbf{Y}|\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k) = \frac{r(\mathbf{X}, \mathbf{Y}|\mathbf{Z}_2, \cdots, \mathbf{Z}_k) - r(\mathbf{X}, \mathbf{Z}_1|\mathbf{Z}_2, \cdots, \mathbf{Z}_k) r(\mathbf{Y}, \mathbf{Z}_1|\mathbf{Z}_2, \cdots, \mathbf{Z}_k)}{\sqrt{1 - r^2(\mathbf{X}, \mathbf{Z}_1|\mathbf{Z}_2, \cdots, \mathbf{Z}_k)} \sqrt{1 - r^2(\mathbf{Y}, \mathbf{Z}_1|\mathbf{Z}_2, \cdots, \mathbf{Z}_k)}}.$$

A Monte-Carlo algorithm is used to permute the distance matrix $\mathbf{X}$, and calculate the probability that the partial correlation coefficient $r(\mathbf{X}', \mathbf{Y}|\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k)$ after $\mathbf{X}$ being permuted as $\mathbf{X}'$ is greater than $r(\mathbf{X}, \mathbf{Y}|\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k)$ in the Mantel test. This probability is used as the single-tailed P value.

6 Update history

2021/7/18 V1.5 TODO

    + xxx Adjust allele frequency model, AICc and BIC equations; xxx

6 Update history

+ Add alternative method to obtain thresholds of Delta for parentage analysis for extreme cases.

+ Add spatial pattern analysis;

+ Add heritability estimation;

+ Add Qst estimation;

+ Add Huang (unpublished) individual inbreeding coefficient estimator;

+ Add Hardy & Vekemans (1999) relatedness estimator and Huang (unpublished) kinship estimator;

+ Add migration rate estimation (BayesAss) function;

+ Add estimates of genotyping error rate and sample rate for all applications;

+ Add likelihood convergency line chart for Structure and BayesAss analyses;

+ Add Jackknife SE estimator for a various of methods;

+ Add save figure function for PCoA and Hierarchical clustering;

+ Add export EMF/WMF vector image function for PCoA, Hierarchical clustering, Structure and BayesAss;

* Revise genetic distance calculation method: use population allele frequency for missing data;

* Revise Ritland (1996) and Loiselle (1995) relatedness estimator to improve accuracy;

* Revise Euclidean genetic distance equation to ensure the results of PCoA based on Euclidean distance is identical to that of PCA.

* Fixed a bug in random seed selection for multiple threads;

* Fixed a bug in import VCF files;

* Fixed a bug causing crash in parentage analysis;

* Fixed a bug in estimate genotyping error rate in parentage analysis;

* Fixed a bug in estimate sample rate in parentage analysis;

* Fixed a bug in Nomura 2008 Ne estimator in handing ties;

6 Update history

* Fixed a bug in calculate genotypic/phenotypic frequencies under selfing;

* Fixed a bug in calculating kinship coefficient in Weir 1996 estimator;

* Fixed a bug in calculating pairwise F_{ST};

* Fixed a bug in calculating genetic distance from F_{ST};

* Fixed a bug in reading unfinished Bayesian clustering results;

* Fixed a bug in calculating individual inbreeding coefficient;

* Fixed a bug in performing genotypic and phenotypic distribution test that cause false positive when the number of alleles is high (Thanks Prof. Olivier Hardy);.

2021/5/1 V1.4

+ Add Model Test function;

+ Add Hardy's (2016) Fz-based and g2z-based selfing rate estimators;

+ Add haplotype sampling (HS) mating system to Waples & Do's (2010) effective population size estimator;

+ Add export dummy-genotype function;

+ Add support for one-digit genotype format;

+ Add loading genotypes from VCF files;

+ Add loading input genotype/allelic phenotype data from file;

+ Add warning for duplicate allele copies for allelic phenotype data;

* Fix finite sample correction for Waples & Do's (2010) linkage disequilibrium estimator, from $1/n$ to $1/(n-1)$;

* Fix a bug in counting genotypes/allelic phenotypes during diversity calculation;

* Fix a bug in Mantel test;

* Fix a bug in linkage disequilibrium test (FDR correction).

2020/10/8 V1.3

+ Add three effective population size estimators;

+ Add Burrow's Δ, corresponding squared correlation coefficient and linkage disequilibrium test;

+ Add Huang *et al*. unpublished *F*-statistics estimator to perform correlated samples correction;

+ Add collapse alleles function in genotype distribution test;

+ Add AICc and BIC estimates for the whole dataset;

- Disable performing genotype and allelic phenotype analyses simultaneously;

- Disable performing linkage disequilibrium test for regions with population with different ploidy levels;

* Adjust UI controls to both high- and low-resolution screens;

* Adjust scale in kinship estimators;

* Fix a bug in calculate d.f. in allelic phenotype distribution test;

* Fix a bug in calculate average observed heterozygosity in genetic diversity;

* Fix a bug in estimating null allele frequency with varying ploidy levels;

* Fix a bug in allele frequency summary in Bayesian results (Thanks Prof. Olivier Hardy);

* Fix a bug in weight genotype method in AMOVA;

* Revising weighing scheme in Slatkin 1996, Nei 1973 and Hudson 1992 Fst estimators.

2020/5/28 V1.2

+ Add Shannon's information index, allelic richness, number of private alleles and *F*-statistics (F_{IX}, F_{SX}, F_{C1X}, …) in genetic diversity results;

+ Add F_{IX} unbiased correction for missing data in AMOVA (homoploid);

* Fix a bug in counting individuals in genetic diversity results;

* Fix a bug in loading individuals and populations using multiple threads;

* Fix a bug in SS calculation in AMOVA (weight genotype) using multiple threads;

* Fix a bug causes crash in AMOVA (likelihood + SMM);

6 Update history

* Fix some display bugs in high- or low-resolution screen.

2020/5/16 V1.1

+ Add two parentage analysis methods (exclusion and dominant);

+ Add the genotyping error rate estimator in parentage analysis;

+ Add the sample rate estimator in parentage analysis;

* Change a parentage analysis method (allelic phenotype) to improve performance;

* Fix Mono compilation for Linux and Mac OS X;

* Fix blurry fonts at high resolutions.

2019/7/7 V1.0b

+ Add cross platform support;

* Fix a bug in AMOVA;

* Fix a bug in parentage analysis (identifying father and mother jointly).

2019/1/7 V1.0

+ Support multi-level region definition

+ Support multi-level AMOVA

+ Add anisoploid, weight genotype and maximum-likelihood AMOVA methods

+ Support for both genotypes and allelic phenotypes

+ Incorporate self-fertilization into allele frequency estimation and parentage analysis

* Optimize calculation speed in AMOVA

* Reduce memory expense

* Revise Bayesian clustering function: rewrite codes and add details parameters

* Revise population simulator: use double-reduction equilibrium genotype generator

6 Update history

2018/9/13 V0.9

+ Add parentage analysis function

+ Add Huang *et al.* (2014, 2015a) relatedness coefficient estimators

+ Apply 'Region' definition of AMOVA for all other relevant functions

* Fix a bug in parentage analysis function that overestimate the mismatch errors

* Revise population simulator: support genotype output

* Apply multi-threading technique for all analysis methods

2018/6/28 V0.8

+ Add AMOVA function

+ Add Population differentiation estimator: Cockerham 1973, Slatkin 1993's Rst, Hudson *et al.* 1992, Hedrick 2005's G'st, and Jost 2008's D

* Optimize observed heterozygosity and genotype pattern calculations

* Optimize genotype and allelic phenotype data structure to accelerate analyses

* Fix a bug in drawing the dendrogram and scatter plot for hierarchical clustering and PCoA results that cause the program to crash

2018/4/2 V0.7

+ Add 2D scatter plot preview for PCoA results

+ Add dendrogram plot for hierarchical clustering results

+ Add bar plot preview for Bayesian clustering results

* Change allelic phenotype distribution test to Fisher's G-test

* Change linkage dis-equilibrium test to Fisher's G-test and Raymond & Rousset's Markov Chain test

* Change genetic differentiation test to Fisher's G-test and Raymond & Rousset's Markov Chain test

* Apply FDR correction for multiple tests

6 Update history

2018/2/19 V0.6

+ Add Population differentiation estimator: Nei 1972's Gst

+ Add PCoA function

+ Add hierarchical clustering function

+ Add corrections for Ritland 1996 and Loiselle 1995 relatedness estimator using Huang *et al.* 2015a Eqn8

+ Add support for different ploidy level for relatedness estimators

* Apply Gst estimator for population simulator

* Fix a bug in allele frequency estimator cause same allelic phenotypes only count once

2017/8/20 V0.5

+ Add Bayesian clustering function

+ Add Mantel test function

* Fix a bug in calculating Ho in total population

- Remove permutation test for genetic differentiation

2017/4/9 V0.4

+ Add relatedness coefficient estimators: Ritland 1996, Loiselle 1995, Weir 1996

+ Add inbreeding coefficient estimators: Ritland 1996, Loiselle 1995, Weir 1996

+ Add genetic differentiation tests: permutation test

* Use symbolic expressions to obtain genotypic and allelic phenotypic frequencies

2017/1/5 V0.3

+ Add allelic phenotype distribution test: Chi-squared test

+ Add linkage dis-equilibrium test: Chi-squared test

+ Add genetic differentiation tests: Chi-squared test,

+ Add support for null alleles, which is implemented in all analyses

2016/8/32 V0.2

    + Add population genetic diversity

    + Add individual H-index estimator

    + Add population assignment

    + Add genetic distance estimators

2016/5/20 V0.1

    + Add population simulator

    + Add inheritance model: RCS, PRCS, CES, PES

    + Add allele frequency estimator

    + Add file format converter for genepop and structure

7 Reference

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* **57**, 289-300.

Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis: Models and estimation procedures. *Evolution* **21**, 550-570.

Cockerham CC, Weir BS (1977) Digenic descent measures for finite populations. *Genetics Research* **30**, 121-147.

Excoffier L, Smouse PE, Quattro JM (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 JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.

Gerber S, Mariette S, Streiff R, Bodenes C, Kremer A (2000) Comparison of microsatellites and

7 Reference

amplified fragment length polymorphism markers for parentage analysis. *Molecular Ecology* **9**, 1037-1048.

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.

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361-372.

Haldane JB (1930) Theoretical genetics of autopolyploids. *Journal of Genetics* **22**, 359-372.

Hamilton M (2009) *Population genetics* John Wiley & Sons, Oxford.

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 (1999) Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**, 145-154.

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.

Hill WG, Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics* **54**, 705.

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 ST, *et al.* (2015b) Estimating pairwise relatedness between individuals with different levels of ploidy. *Molecular Ecology Resources* **15**, 772–784.

Huang K, Ritland K, Guo ST, Shattuck M, Li BG (2014) A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources* **14**, 734-744.

Huang K, Wang T, Dunn DW, *et al.* (2021) A generalized framework for AMOVA with multiple hierarchies and ploidies. *Integrative Zoology* **16**, 33-52.

7 Reference

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 (2006) Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics* **7**, 991-995.

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.

Kalinowski ST, Wagner AP, Taper ML (2006) ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Notes* **6**, 576-579.

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.

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

Marshall TC, Slate JBKE, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood‐based paternity inference in natural populations. *Molecular Ecology* **7**, 639-655.

Mather K (1935) Reductional and equational separation of the chromosomes in bivalents and multivalents. *Journal of Genetics* **30**, 53-78.

Mousseau TA, Ritland K, Heath DD (1998) A novel method for estimating heritability using molecular markers. *Heredity* **80**, 218-224.

Muller HJ (1914) A new mode of segregation in Gregory's tetraploid primulas. *The American Naturalist* **48**, 508-512.

7 Reference

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.

Nelder JA, Mead R (1965) A simplex method for function minimization. *The computer journal* **7**, 308-313.

Nomura T (2008) Estimation of effective number of breeders from molecular coancestry of single cohort sample. *Evolutionary Applications* **1**, 462-474.

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.

Parisod C, Holderegger R, Brochmann C (2010) Evolutionary consequences of autopolyploidy. *New phytologist* **186**, 5-17.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Pudovkin AI, Zaykin DV, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics* **144**, 383-387.

Raymond M, Rousset F (1995) An exact test for population differentiation. *Evolution* **49**, 1280-1283.

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 (1996a) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research* **67**, 175-185.

Ritland K (1996b) A marker‐based method for inferences about quantitative inheritance in natural populations. *Evolution* **50**, 1062-1073.

7 Reference

Rodzen JA, Famula TR, May B (2004) Estimation of parentage and relatedness in the polyploid white sturgeon (Acipenser transmontanus) using a dominant marker approach for duplicated microsatellite loci. *Aquaculture* **232**, 165-182.

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.

Selander RK (1970) Behavior and genetic variation in natural populations. *American Zoologist* **10**, 53-66.

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

Smouse PE, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology* **35**, 627-632.

Taberlet P, Griffin S, Goossens B, *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189-3194.

Thomas SC, Pemberton JM, Hill WG (2000) Estimating variance components in natural populations using inferred relationships. *Heredity* **84**, 427-436.

Waples RS, Chi D (2010) Linkage disequilibrium estimates of contemporary *Ne* using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications* **3**, 244-262.

Weir BS (1996) *Genetic data analysis II: methods for discrete population genetic data* Sinauer Associates, Sunderland.

Weir BS, Cockerham CC (1979) Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **42**, 105.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Wilson GA, Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177-1191.

Zwart AB, Elliott C, Hopley T, Lovell D, Young A (2016) Polypatex: an R package for paternity

7 Reference

exclusion in autopolyploids. *Molecular Ecology Resources* **16**, 694-700.