VCFPOP V1.0 User Manual

Performing population genetics analyses based on NGS data for haploids, diploids, polyploids and anisoploids.

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

1 System Requirement & Limitations

- CPU: x64 compatible
- OS: Windows, Linux and Mac OS X
- Memory: 8 Gb, more memory maybe required when large dataset is processed
- Hard drive: 4 Gb

- Maximum ploidy level: 10 (support anisoploids)
- Maximum number of loci: 4294967295
- Maximum number of individuals: 4294967295
- Maximum number of populations: 4294967295
- Maximum number of regions: 4294967295
- Maximum number of alleles at each locus: 65535
- Maximum number of genotypes at each locus: 4194303

2 Download, Setup, Compile & Uninstall

2.1 Download

The precompiled binary executables (for Windows, Ubuntu and Mac OS X) and the source code can be download via https://github.com/huangkang1987/vcfpop.

2.2 Compilation

For Windows, Ubuntu and Mac OS X, the precompiled binary executables have been provided, so it is unnecessary to compile VCFPOP. For other operation systems, users can compile VCFPOP on their own.

To compile the software in Linux, install the following libraries:

- GCC (including G++) https://gcc.gnu.org/
- Eigen https://eigen.tuxfamily.org/dox/
- Spectra https://spectralib.org/doc/
- Zlib http://zlib.net/

where the version of GCC must be above v8.2.0 to use the AVX-512 instructions. The latest versions for the other three libraries are also suggested.

Place Eigen and Spectra folders in the same folder as the source code. After that, use terminal to enter the `src` folder and run the command `make`, then wait a few minutes. The compilation may require a large amount of memory (8 Gib) to optimize these codes. Then run the command `make install` to copy the compiled binary executable `vcfpop` to the `bin` folder.

For Windows, by using the same method as Linux, the compilation can be performed under the following environments:

- Cygwin https://www.cygwin.com/
- MinGW http://www.mingw.org/
- MSYS2 https://www.msys2.org/
- Ubuntu subsystem of Windows 10

For Microsoft Visual Studio, the version must be 2017 or later because the AVX-512 instructions are used in VCFPOP. In order to compile the program '`vcfpop.exe`', install the following libraries:

- Eigen https://eigen.tuxfamily.org/dox/
- Spectra https://spectralib.org/doc/
- Zlib http://zlib.net/

Next, place Eigen, Spectra and Zlib in the same folder as the source code. After that, compile Zlib as the static library (`zlibstat.vcxproj`), and copy the `zlibstat.lib` to the same folder as the source code. Finally, by double-clicking the ‘`make.bat`’ and waiting a few minutes, the program ‘`vcfpop.exe`’ will be generated.

2.3 Setup

To setup the software, please create a folder on your disk, and then extract the files to the folder.

2.4 Launch and symbolic link

For Linux, open the command mode (bash). After that, switch to the `src` folder, and execute the command `` `./vcfpop' ``.

For Windows, press WINDOWS + R and run `cmd` to open the command mode. After that, switch to the install path of VCFPOP, and execute the command `vcfpop.exe`.

The user can also launch VCFPOP in other directory by adding the install path before the command, such as the path `` `~/Desktop/vcfpop/bin/vcfpop' `` or `` `c:\vcfpop\bin\vcfpop.exe' ``.

For Linux, the symbolic link enables the user to run VCFPOP at any directory without adding the install path, and the symbolic link can be added by:

```
ln -s -f vcfpop /usr/local/bin/vcfpop
```

For Windows, add the install path to the environment variable is the same as the symbolic link. These can be done by the following procedure:

1. open the menu `START’, right click the icon `This PC’ and click `Properties’;
2. select `Advanced system settings’ to open the dialog box `System properties’;
3. click the button `Environment Variables’ to open the dialog box `Environment Variables’;

4. select the item `PATH’ in the top list of `User Variable for XX’, and then click the button `Edit’;
5. click the button `New’ and paste the install path into the dialog box, and then click `OK’ three times to exit.

2.5 Uninstall

To uninstall VCFPOP, simply delete the corresponding folder.

For Linux, if the symbolic link is created, the command `` `rm -rf /usr/local/bin/vcfpop’ `` can be used to delete the symbolic link.

For Windows, to delete the environment variable, the former four steps in Section 2.4 need to be done. After that, select the install path, and click the button ‘Delete’. Finally, click ‘OK’ three times.

3 Usage

3.1 A brief introduction

VCFPOP performs population genetics analyses based on the NGS genotype data (with the VCF and BCF formats), which can also support other genotype formats, such as GENEPOP (Rousset 2008) and STRUCTURE (Pritchard *et al.* 2000). Existing methods for population genetics analyses are restricted to diploids. VCFPOP extends these methods to include haploids, polyploids and anisoploids, and supports a maximum ploidy level of 10. For even ploidy levels, the genotypic frequencies are obtained under the double-reduction equilibrium (Huang *et al.* 2019).

The functions of VCFPOP consist of:

1. Variant information, individual, genotype and locus filters;
2. Haplotype extraction from phased genotypes;
3. File format conversion;
4. Genetic diversity index estimation and genotypic equilibrium test;
5. Individual statistics estimation (inbreeding coefficient, kinship coefficient, heterozygosity);
6. Genetic differentiation index estimation and test;
7. Genetic distance estimation;
8. Analysis of molecular variances;
9. Population assignment;
10. Relatedness coefficient estimation;
11. Kinship coefficient estimation;
12. Principal coordinate analysis;
13. Hierarchical clustering;
14. Bayesian clustering.

VCFPOP has been optimized for memory expense and calculation speed for large datasets. For the Intel SkyLake processors (used on server or workstation, released on 2017) as well as the CannonLake processors (used on PC and laptop, released on 2018), the AVX-512 instructions can be used to accelerate the calculation speed, which are able to handle 512 bits data simultaneously. For Intel processors later than Haswell (released on 2013) and the AMD processors later than Excavator (released on 2015), the AVX2 instructions can be used to accelerate the calculating speed, which are able to handle 256 bits data simultaneously. These instruction sets can be freely switched without additional compilation.

For memory expense, bitwise storage is used to save the individual genotype indices. The memory usage for each genotype for SNPs is reduced from 4 bytes (e.g., `0/0 ') to 2 bits for VCF format (16-

folds). Because there are some extra costs for the information related to the locus and individual as well as the population, the typical compression ratio is 14.5-fold (e.g., if the size of Chr22 of 1000 genome data is 10.45 GiB, and VCFPOP use 1.1 Gib memory after loading the file). This method is a trade-off between access speed and memory expense, which can access the genotype data at a high rate and only requires several bit manipulations together with the integer arithmetic instructions. Although some advanced compressions can reach a compression ratio of 50-fold, the random access of genotype data requires a longer decompression time. Therefore, a typical laptop with a 16 GiB memory can load 100 GiB VCF files, and a regular workstation with 256 GiB memory can load 2 TiB VCF files.

Several methods are used in VCFPOP to optimize the calculation speed, which can completely exploit the potential of computers. These methods are listed below.

- Optimized algorithm for population genetics analyses, e.g., dynamic programing (to avoid repeated calculations).
- Advanced instruction set (SSE, AVX, FMA, AVX512).
- Lock-free multi-threading algorithm (to decrease the access conflict among threads).
- Virtual memory allocation (to avoid re-allocation and move memory when the bucket is full).
- Local memory management class (to allocate millions of small pieces of memory).
- Fast hash algorithm (to detect duplicate genotypes).
- O(1) hash table (to detect and access genotypes quickly).
- Indexing alleles and genotypes (to save memory expense).
- Memory cache for I/O (to reduce the frequency of disk I/O).

The loading speed of VCFPOP is also optimized, and uses a single thread to read the data from the disk and multiple threads to process the data. With a sample benchmark test of a 10.4 GiB uncompressed VCF file, VCFPOP can load data at 320 MiB/s and 560 MiB/s on a laptop (Intel i7-

8750H CPU with 2.2GHz and 6 cores, 16 GiB memory, 256 Gib nvme SSD) and a workstation (Intel Xeon E5-2696 V4 CPU with 2.2GHz and 44 cores, 64 GiB memory and 1 TiB nvme SSD), respectively. Restricted by the additional decompression process, the loading speed is reduced to 255 and 460 MiB/s on the laptop and the workstation for the compressed format (`vcf.gz`), respectively.

The load time is measured on the following machines:

Laptop: Dell G3 15G R 1765 / 2745

- Mainboard: Dell 00FK8Y
- CPU: Core i7-8750H (2.2 GHz, 6 cores)
- Memory: Micron DDR4 2667MHz (16 GiB) X 1
- SSD: SamSung 0MZVLB256HAHQ Nvme M.2 (256 GiB) X 1

Workstation:

- Mainboard: ASUS Z10PE-D16 WS
- CPU: Xeon E5-2696 V4 (2.2 GHz, 44 Cores)
- Memory: Micron PC4-2400T DDR4 ECC REG LRDIMM (64GB) X 1
- SSD: SamSung Evo 960 Nvme M.2 (1TiB) X 1

3.2 Input file format

VCFPOP supports VCF format V4.x and BCF format V2.x (Danecek *et al.* 2011). Because VCFPOP does not use the information that precedes the header row (‘`#Chrom`’), the obsolete or non-standard VCF/BCF files are also supported.

VCFPOP also supports some additional genotype formats but does not optimize their load speeds: GENEPOP (Rousset 2008), SPAGEDI (Hardy & Vekemans 2002), CERVUS (Kalinowski *et al.* 2007), ARLEQUIN (Excoffier & Lischer 2010), STRUCTURE (Pritchard *et al*. 2000), POLYGENE (Huang et al. 2020), and POLYRELATEDNESS (Huang *et al.* 2015a). All these input files can be compressed by gzip with the extension `*.gz`.

3.3 Command overviews

Use the command `` `vcfpop -h' `` to view the help information, or type the command `` `vcfpop -h -function_name' `` to view the detail information for specific functions. For example, the command `` `vcfpop -h -g' `` is able to show the detail command and description of general settings.

The names of the functions are listed below.

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

To perform a specific function, use the command `` `vcfpop -function_name -parameters' ``. For example, the following command convert the input file into GENEPOP format:

```
vcfpop -convert -convert_format=genepop
```

The parameters are separated by spaces. If there are spaces inside a parameter (e.g., path), use double quotes to embrace this parameter. Because there are too many parameters, it is troublesome to type them one by one in the console. Therefore, VCFPOP allows the user to write all parameters in a text file and use the command `-p=*parameter_file*' to run VCFPOP. For example, the command `vcfpop -p=pars.txt' uses all parameters configured in `pars.txt'.

The line break can be used as the parameter separator in the parameter set file.

The data types used in parameters are:

```
integer          e.g., -g_decimal=3
real             e.g., -popas_error=0
string           e.g., -g_scientific=yes
integer range    e.g., -haplotype_alleles=[2,10]
real range       e.g., -f_qual=[50,70]
```

Some parameters (or filters) are optional. For example, if the filter `-f_qual` is not configured, it cannot be applied. Some parameters have their default values. If they are not configured, the default values will be used. For example, the default option is `sse` for the parameter

```
-g_simd=mmx|sse|avx|avx512, integer, default:sse
```

For parameters with multiple selections, multiple values can be used and these values should be are separated by commas. For this situation, the calculation will be performed for each value. For example, the following command will convert the input file into both the GENEPOP format and the SPAGEDI format:

```
vcfpop -convert -convert_format=genepop,spagedi
```

3.4 General settings

General settings consist of the input and output files, the input and output styles, the temporary directory, the definitions of population, region and group, and also include various behaviors in calculations, such as the number of threads, the amount of memory used, the CPU instructions, and the random number generator seed. The parameters are listed below.

`-g_decimal=0~15, integer, default:5`

Decimal places of output real numbers.

`-g_scientific=yes|no, string, default:no`

Use scientific notation to output real numbers.

`-g_nthread=1~4096, integer, default:4`

Number of threads used in calculation.

`-g_simd=mmx|sse|avx|avx512, integer, default:sse`

SIMD (Single Instruction Multiple Data) instruction sets to accelerate float point or vector operations, where mmx, sse (using SSE1.0 to SSE4.2 instructions), avx (using AVX, AVX2 and FMA instructions) and avx512 (using AVX512 F & BW instructions) can handle 64, 128, 256, 512 bits simultaneously, respectively.

`-g_benchmark=yes|no, string, default:no`

Evaluate the performance of each SIMD instructions set from mmx to specified type before calculation.

`-g_seed=0~2147483647, integer, default:0`

Random number generator seed, 0 denotes using system time as the seed.

`-g_tmpdir=path, string, default:current_directory`

Directory placing the temporary files.

`-g_progress=10~100000, integer, default:80`

The number of characters used for the progress bar.

`-g_input=file_path, string`

Input file. Multiple VCF/BCF files using '|' and '&' as the column and row separators, respectively, e.g., `var1-4ind1-3.vcf|var1-4ind4-6.vcf&var5-9ind1-3.vcf|var5-9ind4-6.vcf`. The `FORMAT' field in each file should be equal. If there are spaces in the file path, use double quotes to embrace the whole parameter (e.g., "`-g_input=a 1.vcf&a 2.vcf`").

`-g_format=vcf|bcf|genepop|spagedi|cervus|arlequin|structure|polygene|polyrelatedness, string, default:vcf`

Input file format. The population and region should be defined in `g_indfile` or `g_indtext`, and the within input file will not be used. For GENEPOP format, VCFPOP do not support extra information; for CERVUS, at most one extra column (population or sex) is allowed; for SPAGEDI, multiple extra columns (population or coordinate) are allowed; for STRUCTURE, the number of extra columns can be specified.

`-g_extracol=0~4096, integer, default:0`

Regarding the number of extra columns between individuals and genotypes in the STRUCTURE input file. If there is a header row for locus, the number of extra columns can be automatically detected.

`-g_output=file_path, string, default:vcfpop.out`

Output file.

`-g_indfile=file_path, string, optional`

Assigns the population of each individual and region of each population (where regions can be nested to perform multi-level AMOVA). If `g_indfile` and `g_indtext` are not specified, all individuals are assigned to a default population. In the individual file, each line defines a population or a reg, start with an identifier and a colon, the individuals or populations are separated by commas. Spaces are not allowed. The individuals, populations, and regions not included in the contents are assigned to a default population or a default region. The example individual file is as follow:

```
pop1:ind1,ind2,ind3
pop2:#4-#6
pop3:ind7,ind8,ind9
#REG
reg1:#1-#2
reg2:pop3
```

`-g_indtext=text, string, optional`

Individual file in the text format, where #n, space or line break can be used to separate lines. Specifically, if spaces are used, a pair of double quotes should also be used to embrace the parameter. For example,

```
-g_indtext="pop1:ind1,ind2,ind3 pop2:#4-#6 pop3:ind7,ind8,ind9 #REG reg1:#1-#2 reg2:pop3"
```

`-g_delimitator=comma|tab, string, default:tab`

Column delimiter style, where a comma is used in the CSV format, and a tab is used in the text editors.

`-g_linebreak=unix|win, string, default:unix`

Line break style, where `\n' is used for unix, and `\r\n' is used for windows.

3.5 Filter

A filter can exclude variants of low quality, the individuals with a poor genotyping ratio, the genotypes of low quality and the variants of low genetic diversity. This step is performed before haplotype extraction and file conversion. The parameters of filters are listed as follows.

`-f`

Enable filters

Variant information filters: this filter is applied during loading variants.

`-f_qual=[min_val,max_val], real range, optional`

Range of variant quality. If multiple VCF/BCF files are used, the variant is filtered when at least one QUAL field is out of the range.

`-f_type=snp|indel|both, string, optional`

Type of variants used in calculations.

`-f_original=yes|no, string, optional`

Use original filter of VCF/BCF file. If multiple VCF/BCF files are used, the variant is filtered when at least one original filter is not a 'PASS'

Genotype filters: this filter is applied during loading genotypes, and an excluded genotype will be set as a missing genotype.

`-f_dp=[min_val,max_val], integer range, optional`

Range of sequencing depth.

`-f_gq=[min_val,max_val], integer range, optional`

Range of genotype quality.

`-f_ploidy=[min_val,max_val], integer range, optional`

Range of ploidy level for genotypes.

Individual filters: this filter is applied after the file is loaded.

`-f_ntype=[min_val,max_val], integer range, optional`

Range of number of called variants.

`-f_nploidy=[min_val,max_val], integer range, optional`

Range of ploidy level for individuals.

Locus diversity filters: this filter is applied after the individual filter. The diversity index of each locus will be calculated in a specific population or region.

`-f_pop=pop_identifier|total, string, default:total`

Target population used to calculate diversity and apply diversity filters.

`-f_region=region_identifier, string, optional`

Target region used to calculate diversity and apply diversity filters.

`-f_bmaf=[min_val,max_val], real range, optional`

Range of frequencies of minor alleles for biallelic loci.

`-f_k=[min_val,max_val], integer range, optional`

Range of number of alleles.

`-f_n=[min_val,max_val], integer range, optional`

Range of number of typed individuals.

`-f_ptype=[min_val,max_val], real range, optional`

Range of typed ratio.

`-f_pval=[min_val,max_val], real range, optional`

Range of P values in equilibrium tests.

`-f_model=rcs|prcs|ces|pes, string, default:rcs`

Double-reduction model to calculate genotypic frequencies for polyploids.

`-f_he=[min_val,max_val], real range, optional`

Range of expected heterozygosity.

`-f_ho=[min_val,max_val], real range, optional`

Range of observed heterozygosity.

`-f_pic=[min_val,max_val], real range, optional`

Range of polymorphic information content.

`-f_ae=[min_val,max_val], real range, optional`

Range of effective number of alleles.

3.6 Haplotype extraction

Haplotype extraction combines several adjacent variants into a single, highly polymorphic locus. The extracted haplotypes will be regarded as alleles, and the extracted locus can be further exported and further analysed.

`-haplotype`

Extracts haplotypes from phased genotypes, then use the haplotypes as alleles for further analysis. Note that all genotypes must be phased and only the variants genotyped in all individuals are used. The haplotype definitions are saved in *.haplotype.txt.

`-haplotype_typerate=[min_val,max_val], real range, default:[0.8,1]`

Range of genotype rate at extract loci.

`-haplotype_length=[min_val,max_val], integer range, default:[1,1000000]`

Range of haplotype size (in bp).

`-haplotype_variants=[min_val,max_val], integer range, default:[5,20]`

Range of number of variants in the haplotype.

`-haplotype_interval=0~100000000000, integer, default:0`

Minimum interval between adjacent loci (in bp).

`-haplotype_alleles=[min_val,max_val], integer range, default:[2,65535]`

Range of number of alleles at the extracted locus.

`-haplotype_genotypes=[min_val,max_val], integer range, default:[2,65535]`

Range of number of genotypes at the extracted locus.

The results are saved in `*.haplotype.txt`, including a heading of parameters used, CHROM, the number of variants, the number of haplotypes (i.e., the number of alleles), range (from the first position to the last position of variants), the length of haplotypes, the distance to the next extracted locus (i.e., the length of interval), the extracted haplotype and the corresponding alleles in each variant. An example is shown as follows.

```
Extracted haplotype information calculated by vcfpop v1.0
  Input: 1M.vcf
  Output: test.out
  Time: 2018-11-07 17:51:09
  Parameter: D:\vcfpop\vcfpop.exe
    -g_output=test.out
    -g_indtext=pop1:#1-#500 pop2:#501-#1000 pop3:#1001-#1500 pop4:#1501-#2000 pop5:#2001-#2504 #REG reg1:#1-#2 reg2:#3-#5
    -g_decimal=3
    -g_input=1M.vcf
    -g_format=vcf
    -haplotype
    -haplotype_variants=[2,3]
    -haplotype_alleles=[2,5]


Locus:Loc1
#CHROM:22
#Variants:2
#Haplotypes:3
Range:16050075-16050115
Length:41
Distance to next extracted locus:97
HapId   rs587697622     rs587755077
0       A       A
1       A       G
2       G       A

Locus:Loc2
#CHROM:22
#Variants:2
#Haplotypes:3
Range:16050213-16050319
Length:107
Distance to next extracted locus:207
HapId   rs587654921     rs587712275
0       C       C
1       T       C
2       C       T

Locus:Loc3
```

3.7 Conversion

This function converts genotypes into another genotype format, either GENEPOP (Rousset 2008), SPAGEDI (Hardy & Vekemans 2002), CERVUS (Kalinowski *et al*. 2007) , ARLEQUIN (Excoffier & Lischer 2010), STRUCTURE (Pritchard *et al*. 2000), POLYGENE (Huang *et al*. 2020), or POLYRELATEDNESS (Huang *et al*. 2015a). The results are saved in `*.convert.xxx.txt`, whose parameter is

`-convert`

Converts filtered data (and extracted haplotype) into the input format of other software. The result is saved in `*.convert.genepop.txt`.

`-convert_format=genepop|spagedi|cervus|arlequin|structure|polygene|polyrelatedness, string, multiple selections, default:genepop`

Target format, where GENEPOP, CERVUS and ARLEQUIN formats only support diploids.

3.8 Genetic diversity indices

This function estimates the genetic diversity indices and performs the genotypic equilibrium test (e.g., the HWE test). For the polyploid data, VCFPOP employs a Fisher's *G*-test to perform the genotypic distribution test. However, it is noteworthy that this test is performed when the ploidy levels all individuals in the population/region are equal. The parameters related to the genetic diversity are as follows.

`-diversity`

Estimates the genetic diversity indices. Results are saved in `*.diversity.txt`.

`-diversity_level=pop|reg|tot|popXloc|regXloc|totXloc, string, multiple selections, default:loc,pop`

Output mean diversity across all loci in each population, each region or in the total population, or output diversity for each locus in each population, in each region or in the total population.

`-diversity_model=rcs|prcs|ces|pes, string, multiple selections, default:rcs`

Double-reduction model to calculate genotypic frequencies for polyploids.

The results are saved in `*.diversity.txt`. An example is shown below.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
14	-diversity																			
15	-diversity_level=pop,reg,tot,popXloc,regXloc,totXloc																			
16	-diversity_model=hwe																			
17																				
18																				
19	Pop	Locus	Ploidy	k	n	#Hap	Ho	He	PIC	Ae	I	NE1P	NE2P	NEPP	NEID	NESID	Fis	G	d.f.	P-val
20	Total	(10)	4.000-4.0	4.000	120.000	480.000	0.696	0.702	0.648	3.358	1.284	0.040	0.003	0.000	0.000	0.000	0.008	-	-	-
21																				
22	Pop	Locus	Ploidy	k	n	#Hap	Ho	He	PIC	Ae	I	NE1P	NE2P	NEPP	NEID	NESID	Fis	G	d.f.	P-val
23																				
24	Pop	Locus	Ploidy	k	n	#Hap	Ho	He	PIC	Ae	I	NE1P	NE2P	NEPP	NEID	NESID	Fis	G	d.f.	P-val
25	pop1	(10)	4.000-4.0	4.000	30.000	120.000	0.707	0.712	0.660	3.483	1.306	0.034	0.002	0.000	0.000	0.000	0.007	-	-	-
26	pop2	(10)	4.000-4.0	4.000	30.000	120.000	0.693	0.691	0.634	3.241	1.254	0.047	0.004	0.000	0.000	0.000	-0.003	-	-	-
27	pop3	(10)	4.000-4.0	4.000	30.000	120.000	0.694	0.693	0.638	3.263	1.267	0.044	0.003	0.000	0.000	0.000	-0.002	-	-	-
28	pop4	(10)	4.000-4.0	4.000	30.000	120.000	0.692	0.696	0.641	3.299	1.272	0.043	0.003	0.000	0.000	0.000	0.007	-	-	-
29																				
30	Pop	Locus	Ploidy	k	n	#Hap	Ho	He	PIC	Ae	I	NE1P	NE2P	NEPP	NEID	NESID	Fis	G	d.f.	P-val
31	Total	Loc1	4-4	4	120	480	0.686	0.700	0.646	3.337	1.280	0.726	0.558	0.385	0.254	0.436	0.020	-0.845	0	nan
32	Total	Loc2	4-4	4	120	480	0.689	0.697	0.643	3.298	1.275	0.729	0.561	0.387	0.255	0.438	0.011	-2.788	0	nan
33	Total	Loc3	4-4	4	120	480	0.688	0.704	0.652	3.383	1.292	0.721	0.551	0.376	0.245	0.433	0.024	1.203	0	nan
34	Total	Loc4	4-4	4	120	480	0.703	0.706	0.654	3.400	1.299	0.718	0.547	0.372	0.241	0.432	0.004	-6.104	0	nan
35	Total	Loc5	4-4	4	120	480	0.707	0.719	0.665	3.560	1.314	0.709	0.539	0.368	0.240	0.424	0.017	9.124	0	nan
36	Total	Loc5	4-4	4	120	480	0.669	0.691	0.635	3.233	1.259	0.736	0.569	0.397	0.262	0.442	0.031	4.286	0	nan
37	Total	Loc7	4-4	4	120	480	0.708	0.702	0.646	3.356	1.280	0.725	0.558	0.386	0.257	0.435	-0.009	-4.250	0	nan
38	Total	Loc8	4-4	4	120	480	0.686	0.707	0.652	3.413	1.290	0.721	0.552	0.380	0.250	0.432	0.030	-0.946	0	nan
39	Total	Loc9	4-4	4	120	480	0.717	0.694	0.638	3.269	1.266	0.732	0.566	0.393	0.262	0.440	-0.033	13.770	0	nan
40	Total	Loc10	4-4	4	120	480	0.711	0.700	0.646	3.336	1.281	0.726	0.557	0.384	0.252	0.436	-0.016	2.304	0	nan

Here, the genetic diversity indices in the first five rows are divided into two parts. One part consists of k, n, #Hap, Ho, He, PIC, Ar and Fis, and the other part consists of NE1P, NE2P, NEPP, NEID and NESID. For the former (or the latter) part, the value of each index in a population is taken as the average (or the product) of those values across all locus.

The description of the table header is as follows (some modifications are made to be compatible with polyploids):

- Locus, number of loci (for the first five rows) or locus name (for the other rows);
- Ploidy, range of ploidy level;
- k, number of distinct alleles at this locus;
- n, total number of individuals genotyped at this locus;
- #Hap, number of haplotypes (or alleles copies);
- Ho, observed heterozygosity;
- He, expected heterozygosity;
- PIC, polymorphic information content;
- Ae, effective number of alleles;
- I, Shannon's information index;

- NE1P, average probability of not excluding a candidate parent from the parentage of an arbitrary offspring, given only the genotype of the offspring;
- NE2P, average probability of not excluding a candidate parent from the parentage of an arbitrary offspring, given the genotype of the offspring and of the known parent of the opposite sex;
- NEPP, average probability of not excluding a candidate parent pair from the parentage of an arbitrary offspring, given only the genotype of the offspring;
- NEID, average probability that the genotypes at a single locus does not differ between two unrelated individuals;
- NESID, average probability that the genotypes at a single locus do not differ between two full siblings;
- Fis, inbreeding coefficient;
- G, Fisher's G statistic of genotypic equilibrium test;
- d.f., degree of freedom of genotypic equilibrium test;
- P-val, significance of genotypic equilibrium test.

3.9 Individual statistics

This function calculates the individual statistics (e.g., inbreeding coefficient, heterozygosity, and likelihood, kinship coefficient). The parameters are as follows:

`-indstat`

Calculates individual statistics (e.g., inbreeding coefficient, heterozygosity, and etc). Results are saved in `*.indstat.txt`.

`-indstat_type=hidx|lnl|f|theta, string, multiple selections, default:hidx,lnl`

Output statistics: heterozygosity index, natural logarithm of genotype likelihood, inbreeding coefficient and kinship coefficient.

`-indstat_model=rcs|prcs|ces|pes, string, multiple selections, default:rcs`

Double-reduction model to calculate genotypic frequencies for polyploids.

`-indstat_estimator=Ritland1996|Loiselle1995|Weir1996, string, multiple selections, default:Ritland1996`

Inbreeding coefficient and kinship coefficient (within an individual itself) estimators.

`-indstat_ref=pop|reg|total, string, multiple selections, default:total`

Reference population: in the population, the region or the total population.

`-indstat_locus=all|each, string, multiple selections, default:all`

Output individual statistics for all loci or for each locus.

The results are saved in `*.indstat.txt`. An example is shown in the following.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
14	-indstat													
15	-indstat_ref=pop													
16	-indstat_model=prcs													
17	-indstat_locus=all													
18	-indstat_type=hidx,lnl,f,theta													
19	-indstat_estimator=Ritland1996,Loiselle1995,Weir1996													
20														
21														
22							All loci							
23	Ind	Pop	#typed	#miss	Ploidy	#Hap	H-idx	lnL_pop_p	F_pop_RI	F_pop_LO	F_pop_WE	Theta_pop	Theta_pop	Theta_pop_WE
24	Ind1	pop1	10	0	4-4	40	1.000	-31.783	-0.234	-0.263	-0.404	0.075	0.053	-0.053
25	Ind2	pop1	10	0	4-4	40	0.750	-31.606	-0.049	-0.024	-0.053	0.213	0.232	0.210
26	Ind3	pop1	10	0	4-4	40	0.700	-35.653	0.166	0.118	0.017	0.374	0.339	0.263
27	Ind4	pop1	10	0	4-4	40	0.717	-30.005	-0.043	-0.060	-0.006	0.218	0.205	0.246
28	Ind5	pop1	10	0	4-4	40	0.750	-31.822	-0.052	-0.044	-0.053	0.211	0.217	0.210
29	Ind6	pop1	10	0	4-4	40	0.683	-31.869	-0.004	0.004	0.041	0.247	0.253	0.281
30	Ind7	pop1	10	0	4-4	40	0.617	-30.207	0.012	0.054	0.134	0.259	0.290	0.351
31	Ind8	pop1	10	0	4-4	40	0.717	-30.996	-0.065	-0.046	-0.006	0.201	0.215	0.246

The description of table header is as follows:

- #typed, number of typed loci;
- #miss, number of missing genotypes;
- Ploidy, range of ploidy levels;
- #Hap, total number of haplotypes (i.e., allele copies);
- H-idx, heterozygosity index, taken from the average of heterozygosity across all loci;
- lnL_pop_prcs, natural logarithm of likelihood of genotypes in the current population under the PRCS double-reduction equilibrium model;
- F_pop_RI, inbreeding coefficient, using the allele frequencies in the current population as a reference and estimated by the Ritland1996 estimator;

- Theta_pop_RI, kinship coefficient, using the allele frequencies in the current population as a reference and estimated by the Ritland1996 estimator.

3.10 Genetic differentiation

For the genetic differentiation analysis, several F_{ST} analogous indices are calculated, and the differentiation between/among populations or regions are tested. The differentiation tests are performed by Fisher's *G*-tests, in which one is based on the genotype distribution, and the other is based on the allele distribution. The parameters related to the genetic differentiation are listed below.

`-fst`

Estimates the Fst statistics. Results are saved in `*.fst.txt`.

`-fst_level=regXtot|popXtot|popXreg|reg|pop, string, multiple selections, default:pop`

Estimates the Fst among all regions, among all populations, among populations in each region, between any two regions, and between any two populations.

`-fst_estimator=Nei1973|Weir1984|Hudson1992|Slatkin1995|Hedrick2005|Jost2008|Huang2021_homo|Huang2021_aniso, string, multiple selections, default:Nei1973`

Fst estimator, Nei1973 (Gst; Nei 1973, PNAS), Weir1984 (variance decomposition method, Weir & Cockerham 1984, Evolution), Hudson1992 (mean difference method, Hudson et al. 1992, Genetics), Slatkin1995 (Rst, Slatkin 1995, Genetics, for non-VCF/BCF input file only), Hedrick2005 (G'st; Hedrick 2005, Evolution), Jost2008 (D; Jost 2008, Molecular Ecology), Huang2021 (variance decomposition method for polyploid or anisoploid, Integrative Zoology).

`-fst_fmt=matrix|table, string, multiple selections, default:matrix`

Output format.

`-fst_locus=all|each, string, multiple selections, default:all`

Calculates Fst and perform test for all loci or for each locus.

`-fst_test=genotype|allele, string, multiple selections, default:no`

Tests the significance of differentiation by Fisher's G-test based on genotype distributions or allele distributions.

The results are saved in `*.fst.txt`. An example is shown as follows.

	A	B	C	D	E	F	G	H	I
14	-fst								
15	-fst_estimator=Nei1973								
16	-fst_level=popXtot,pop								
17	-fst_locus=all								
18	-fst_test=genotype								
19	-fst_fmt=matrix								
20									
21									
22	Locus	A	Among all	pop1	pop1	pop1	pop2	pop2	pop3
23		B		pop2	pop3	pop4	pop3	pop4	pop4
24	All loci	Nei1973	0.006	0.003	0.004	0.003	0.004	0.005	0.005
25		Genotype	103.228	40.301	39.051	50.986	28.089	37.836	31.812
26		d.f.	90.000	30.000	30.000	30.000	30.000	30.000	31.000
27		P	0.161	0.099	0.125	0.010	0.566	0.154	0.426
28									
29									
30	Nei1973	pop1	pop2	pop3	pop4				
31	pop1	0.000	0.003	0.004	0.003				
32	pop2	0.003	0.000	0.004	0.005				
33	pop3	0.004	0.004	0.000	0.005				
34	pop4	0.003	0.005	0.005	0.000				

In the first table (Line 22-37), the header row shows four estimators and two Fisher's G statistics together with two degrees of freedom and two P-values (significances of differentiation by the Fisher's G-test); Line 24 shows various results related to the global F_{ST} among all populations; each of Lines 25-27 shows various results related to the differentiation test between a pair of populations. Next, the output format for Lines 30-34 is a matrix, where each element in this matrix is the F_{ST} estimate between a pair of populations for the Nei1973 estimator.

3.11 Genetic distance

This function calculates the genetic distance between individuals, populations or regions. Specifically, Reynolds *et al.*'s (1983) distance D and the Slatkin's (1995) linearized distance F_{ST} are obtained by transforming the corresponding F_{ST} estimators. The parameters of this function are listed below.

`-gdist`

Estimates the genetic distance. Results are saved in `*.gdist.txt`.

`-gdist_weightmissing=yes|no, string, default:yes`

Use population/region allele frequency for missing data.

`-gdist_level=ind|pop|reg, string, multiple selections, default:pop`

Estimates the genetic distance between individuals, populations or regions.

`-gdist_estimator=Nei1972|Cavalli-Sforza1967|Reynolds1983|Nei1983|Euclidean|Goldstein1995|Nei1974|Roger1972|Slatkin_Nei1973|Slatkin_Weir1984|Slatkin_Hudson1992|Slatkin_Slatkin1995|Slatkin_Hedrick2005|Slatkin_Jost2008|Slatkin_Huang2021_homo|Slatkin_Huang2021_aniso|Reynolds_Nei1973|Reynolds_Weir1984|Reynolds_Hudson1992|Reynolds_Slatkin1995|Reynolds_Hedrick2005|Reynolds_Jost2008|Reynolds_Huang2021_homo|Reynolds_Huang2021_aniso, string, multiple selections, default:Nei1972`

Genetic distance estimators: Nei1972 (Ds, Nei 1972, Am Nat), Cavalli-Sforza1967 (Cavalli-Sforza & Edwards 1967, Am J Human Genet), Reynolds1983 (thetaW, Reynolds et al. Genetics, 1983), Nei1983 (Da, Nei 1983, J Mol Evol), Euclidean, Goldstein1995 (dmu2, Goldstein 1995, PNAS), Nei1974 (Dm, Nei & Roychoudhury 1974, Am J Human Genet), Roger1972 (Rogers 1972, Studies in Genetics), the Slatkin's transform d = Fst/(1-Fst) converts the range of Fst from [0,1] to [0, infinity), and the Reynolds's transformation d = -ln(1 - Fst).

`-gdist_fmt=matrix|table, string, multiple selections, default:matrix`

Output format.

The results are saved `*.gdist.txt`. An example is shown as follows.

	A	B	C	D	E
14	-gdist				
15	-gdist_level=pop				
16	-gdist_estimator=Nei1972,Euclidean				
17	-gdist_fmt=matrix,table				
18					
19					
20	A	B	Nei1972	Euclidean	
21	pop1	pop1	0.000	0.000	
22	pop1	pop2	0.012	0.274	
23	pop1	pop3	0.017	0.320	
24	pop1	pop4	0.016	0.307	
25	pop2	pop2	0.000	0.000	
26	pop2	pop3	0.020	0.351	
27	pop2	pop4	0.022	0.362	
28	pop3	pop3	0.000	0.000	
29	pop3	pop4	0.021	0.360	
30	pop4	pop4	0.000	0.000	
31					
32	Nei1972	pop1	pop2	pop3	pop4
33	pop1	0.000	0.012	0.017	0.016
34	pop2	0.012	0.000	0.020	0.022
35	pop3	0.017	0.020	0.000	0.021
36	pop4	0.016	0.022	0.021	0.000

3.12 Analysis of molecular variance

The analysis of molecular variance (AMOVA) partitions the genetic variance into several hierarchies and tests the significance of each variance component. The procedure of AMOVA follows Excoffier *et al*. (1992) and Weir & Cockerham (1984), but some modifications are made so as to accommodate polyploids and anisoploids. The parameters related to the AMOVA are listed below.

`-amova`

Performs analysis of molecular variance. Results are saved in `*.amova.txt`.

`-amova_method=homoploid|anisoploid|likelihood, string, multiple selections, default:homoploid`

The homoploid method requires that all individuals are homoploids, and performs AMOVA and tests by extracting and permuting the dummy haplotypes. The anisoploid method supports anisoploids and permutes the alleles at each locus. The likelihood method also supports anisoploids, and uses the maximum-likelihood estimator to estimate F-statistics (Fis, Fic, Fit).

`-amova_mutation=iam|smm, string, multiple selections, default:iam`

Allele mutation model, iam denotes infinity alleles model (Fst like, distance between alleles is a binary variable) and smm denotes stepwise mutation model (Rst like, distance between alleles is the absoulte value of their difference in sizes). The smm model can only be applied for non-vcf input file and should use size as the allele identifier.

`-amova_ind=yes|no, string, multiple selections, default:yes`

Includes the individual level during AMOVA.

`-amova_test=yes|no, string, default:yes`

Evaluates the significance of each variance component and F-statistics (Fis, Fic, Fit, Fsc, Fst, Fct).

`-amova_nperm=99~99999999, integer, default:9999`

Number of permutations.

`-amova_pseudo=0~9999, integer, default:50\n");`

Number of pseudo-permutations for the anisoploid method. Zero-value disables the pseudo-permutation.

`-amova_printss=yes|no, string, default:no`

Prints SS within individuals, populations and regions.

The results are saved in `*.amova.txt`. An example is shown as follows.

	A	B	C	D	E	F
14	-amova					
15	-amova_method=anisoploid					
16	-amova_mutation=iam					
17	-amova_ind=yes					
18	-amova_test=yes					
19	-amova_nperm=999					
20	-amova_printss=no					
21						
22						
23	AMOVA Summary, method: anisoploid, mutation model: IAM, ind-level=yes					
24	Source	d.f.	SS	MS	Var	Percentage
25	Within Individual	3600	1253.500	0.348	0.348	99.000
26	Among Individual	1160	421.600	0.363	0.004	1.084
27	Among Population	30	9.831	0.328	0.000	-0.085
28	Total	4790	1684.931	0.352	0.352	100.000
29						
30	F-statistics					
31	Statistics	Value	Permute M	Permute V	Pr(rand>ol	Pr(rand=obs)
32	FIS	0.011	-0.001	0.000	0.036	0.002
33	FIT	0.010	0.000	0.000	0.074	0.000
34	FST	-0.001	0.000	0.000	0.736	0.005

The first table (Lines 23-28) shows the AMOVA summary, and the results of the following four sources are shown in Line 25 to Line 28 in turn: within individuals, among individuals, among populations, among regions and the total population. The second table (Lines 28-33) shows six *F*-statistics. The meanings of column headers in Line 19 are as follows.

- D.f., degrees of freedom;
- SS, sum of squares;
- MS, mean squares, i.e., SS/d.f.;
- Var, Variance component.

The meanings of column headers in Line 27 are as follows.

- Permute Mean, Mean of permuted F-statistics;
- Permute Var, Variance of permuted F-statistics;
- Pr(rand>obs), Probability that permuted F-statistics is greater than the original value;
- Pr(rand=obs), Probability that permuted F-statistics is equal to the original value, note that if the difference is smaller than 10^{-7}, then they are considered equal.

3.13 Population assignment

Population assignment assigns each individual to the population/region with the maximum likelihood, and facilitates the identification of the natal population of each individual. The likelihood is the product of genotype frequencies across loci using the allele frequencies of the target population/region (Paetkau *et al.* 2004). The parameters related to the population assignment are listed below:

`-popas`

Assigns individuals to their natal population according to their genotypic frequencies in each population. Results are saved in `*.popas.txt`.

`-popas_model=rcs|prcs|ces|pes, string, multiple selections, default:rcs`

Double-reduction model to calculate genotypic frequencies for polyploids.

`-popas_level=pop|reg, string, multiple selections, default:pop`

Assigns individuals to populations or regions.

`-popas_error=0~0.2, real, default:0.01`

Mistype rate, used to avoid the probability of being zero.

The results are saved in `*.popas.txt`. An example is shown in the following.

	A	B	C	D	E	F	G	H	I	J	K	L	M
14	-popas												
15	-popas_model=prcs												
16	-popas_level=pop,reg												
17	-popas_error=0.05												
18													
19													
20	Ind	Pop	#typed	#miss	Ploidy	#Hap	assign_	lnL_pop1_	lnL_pop2_	lnL_pop3_	lnL_pop4_	assign_reg	lnL_Total_prcs
21	Ind1	pop1	10	0	4 - 4	40	pop1	-31.85	-34.714	-33.548	-33.419	Total	-32.894
22	Ind2	pop1	10	0	4 - 4	40	pop2	-31.619	-31.207	-31.504	-32.51	Total	-31.475
23	Ind3	pop1	10	0	4 - 4	40	pop1	-35.714	-39.213	-37.239	-37.617	Total	-37.119
24	Ind4	pop1	10	0	4 - 4	40	pop3	-30.048	-31.974	-29.674	-31.564	Total	-30.52
25	Ind5	pop1	10	0	4 - 4	40	pop1	-31.856	-33.475	-32.252	-32.73	Total	-32.237
26	Ind6	pop1	10	0	4 - 4	40	pop1	-31.898	-32.315	-33.333	-32.682	Total	-32.217
27	Ind7	pop1	10	0	4 - 4	40	pop4	-30.225	-30.146	-31.371	-30.112	Total	-30.219
28	Ind8	pop1	10	0	4 - 4	40	pop4	-31.015	-31.858	-31.487	-30.806	Total	-30.959
29	Ind9	pop1	10	0	4 - 4	40	pop2	-30.077	-29.083	-29.85	-30.163	Total	-29.595
30	Ind10	pop1	10	0	4 - 4	40	pop1	-32.977	-34.492	-33.634	-34.693	Total	-33.583

Where each identifier in the column assign_pop_prcs is an assigned population under the PRCS double-reduction model, and each value in the subsequent columns is the natural logarithm of the genotype likelihood with the hypothesis that the individual originates from a target population.

3.14 Relatedness coefficient

Relatedness is the correlation between the means of allele frequencies between individuals (correlation definition), or the probability that an allele sampled from one individual is identical-by-descent to one of the alleles from the other individual (IBD definition). For diploids, VCFPOP provides nine relatedness estimators (Anderson & Weir 2007; Huang *et al.* 2016; Li *et al.* 1993; Lynch & Ritland 1999; Milligan 2003; Queller & Goodnight 1989; Thomas 2010; Wang 2002).

For polyploids and anisoploids, VCFPOP provides two relatedness estimators, in which one is a method-of-moment estimator (Huang *et al.* 2014), and the other is a maximum-likelihood estimator (Huang *et al.* 2015a). Moreover, VCFPOP also provides three kinship estimators, which can be used to estimate the relatedness coefficient between two individuals (Loiselle *et al.* 1995; Ritland 1996; Weir 1996).

The maximum ploidy level supported is eight for Huang *et al.* (2014) and Huang *et al.* (2015a), or ten for Ritland (1996), Loiselle *et al.* (1995) and Weir (1996). Huang *et al.* (2014) is designed for multi-allelic locus with number of alleles greater than ploidy level. Otherwise, it may encounter a singular matrix problem. For two individuals with distinct ploidy levels, the relatedness coefficient between them from the higher ploidy individual to the lower ploidy individual can also be calculated (see Huang *et al.* 2015b). There are two methods (one is the original version and the other is the modified version) to convert the kinship coefficients to the relatedness coefficients.

The parameters associated with the relatedness coefficients are listed below.

`-relatedness`

Estimates pairwise relatedness between individuals. Results are saved in `*.relatedness.txt`.

`-relatedness_range=pop|reg|total, string, multiple selections, default:total`

Estimates pairwise relatedness between members within the same population, the same region or the total population.

`-relatedness_fmt=matrix|table, string, multiple selections, default:matrix`

Output format.

```
-relatedness_estimator=Lynch1999|Wang2002|Thomas2010|Li1993|Queller1989|Huang2
 016A|Huang2016B|Milligan2003|Anderson2007|Huang2014|Huang2015|Ritland1996_mod
 ified|Loiselle1995_modified|Ritland1996|Loiselle1995|Weir1996, string, multip
 le selections, default=Lynch1999
```

Relatedness estimators: Huang2014 and Huang2015 support ploidy level <= 8, Ritland1996, Loiselle1995 and Weir1996 estimators support ploidy level <= 10, and other estimators only support diploids. Milligan2003, Anderson2007 and Huang2015 are maximum-likelihood estimators, and other estimators are method-of-moment estimators. Unbiased Ritland1996 and Loiselle1995 relatedness estimates are converted from kinship coefficient by eqn (8) of Huang et al. (2015, Heredity).

The results are saved in `*.relatedness.txt`. The table format results come first, then the matrix format results. An example is shown in the following.

	A	B	C	D	E	F	G	H	I
14	-relatedness								
15	-relatedness_range=pop								
16	-relatedness_fmt=matrix,table								
17	-relatedness_estimator=Huang2015								
18									
19									
20	pop1								
21	A	pop	B	pop	AB_typed	A_typed	B_typed	Huang2015	
22	Ind1	pop1	Ind1	pop1	10	10	10	1.000	
23	Ind1	pop1	Ind2	pop1	10	10	10	0.101	
24	Ind1	pop1	Ind3	pop1	10	10	10	0.250	
25	Ind1	pop1	Ind4	pop1	10	10	10	0.250	
26	Ind1	pop1	Ind5	pop1	10	10	10	0.228	
487									
488	pop1								
489	Huang201!	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8
490	Ind1	1.000	0.101	0.250	0.250	0.228	0.250	0.000	0.121
491	Ind2	0.101	1.000	0.250	0.088	0.006	0.000	0.000	0.000
492	Ind3	0.250	0.250	1.000	0.000	0.000	0.250	0.000	0.000
493	Ind4	0.250	0.088	0.000	1.000	0.000	0.000	0.330	0.000
494	Ind5	0.228	0.006	0.000	0.000	1.000	0.250	0.000	0.310
495	Ind6	0.250	0.000	0.250	0.000	0.250	1.000	0.181	0.250
496	Ind7	0.000	0.000	0.000	0.330	0.000	0.181	1.000	0.063
497	Ind8	0.121	0.000	0.000	0.000	0.310	0.250	0.063	1.000

Where each value in the column AB_typed (A_typed or B_typed) is the number of loci genotyped in the individuals A and B (the individual A or the individual B), and each value in the column Thomas2010 is the relatedness estimate.

3.15 Kinship coefficient

The kinship coefficient (also known as the coancestry coefficient) is the probability that two alleles, one randomly drawn from each individual with replacement, are identical-by-descent. VCFPOP provides three method-of-moment estimators to estimate the kinship coefficient between individuals. The parameters related to kinship coefficients are listed in the following.

`-kinship`

Estimates kinship coefficient between individuals. Results are saved in `*.kinship.txt`.

`-kinship_range=pop|reg|total, string, multiple selections, default:total`

Estimates the kinship coefficient between members within the same population, the same region or the total population.

`-kinship_fmt=matrix|table, string, multiple selections, default:matrix`

Output format.

```
-kinship_estimator=Ritland1996|Loiselle1995|Weir1996, string, multiple selecti
 ons, default:Ritland1996
```

Kinship estimators. Supports a maximum level of ploidy of 10.

The results are saved in `*.kinship.txt`, with the same format as the file of previous section. An example is shown as follows.

	A	B	C	D	E	F	G	H	I	J	
14	-kinship										
15	-kinship_fmt=matrix,table										
16	-kinship_range=total										
17	-kinship_estimator=Ritland1996										
18											
19											
20	Total										
21	A	pop	regL1	B	pop	regL1	AB_typed	A_typed	B_typed	Ritland1996	
22	Ind1	pop1	reg1	Ind1	pop1	reg1	10	10	10	0.099	
23	Ind1	pop1	reg1	Ind2	pop1	reg1	10	10	10	0.003	
24	Ind1	pop1	reg1	Ind3	pop1	reg1	10	10	10	0.079	
25	Ind1	pop1	reg1	Ind4	pop1	reg1	10	10	10	0.004	
26	Ind1	pop1	reg1	Ind5	pop1	reg1	10	10	10	0.023	
7282											
7283	Total										
7284	Ritland199	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8	Ind9	In
7285	Ind1	0.099	0.003	0.079	0.004	0.023	0.018	-0.033	0.004	-0.055	
7286	Ind2	0.003	0.211	0.066	0.022	-0.01	-0.037	-0.006	-0.062	-0.045	
7287	Ind3	0.079	0.066	0.472	-0.006	-0.077	0.031	-0.003	0.006	-0.004	
7288	Ind4	0.004	0.022	-0.006	0.242	0.012	-0.018	0.048	-0.011	-0.038	
7289	Ind5	0.023	-0.01	-0.077	0.012	0.231	0.077	-0.051	0.033	-0.081	
7290	Ind6	0.018	-0.037	0.031	-0.018	0.077	0.268	0.023	0.051	0.003	
7291	Ind7	-0.033	-0.006	-0.003	0.048	-0.051	0.023	0.257	-0.004	0.03	
7292	Ind8	0.004	-0.062	0.006	-0.011	0.033	0.051	-0.004	0.205	0.043	
7293	Ind9	-0.055	-0.045	-0.004	-0.038	-0.081	0.003	0.03	0.043	0.265	

3.16 Principal coordinates analysis

The principal coordinates analysis (PCoA) is an ordination technique to coordinate individuals into a multi-dimensional space by their dissimilarity to visualize the data, reduce the dimension of data with the least loss of information, and eliminate the correlation among coordinates. This technique is performed based on the genetic distances. The results using Euclidean distance are equivalent to principal components analysis (PCA). The parameters related to the PCoA are listed as follows.

`-pcoa`

Performs principal coordinate analysis for individuals, populations or regions. Results are saved in `*.pcoa.txt`.

`-pcoa_level=ind|pop|reg, string, multiple selections, default:ind`

Ordinate individuals, populations or regions.

`-pcoa_dim=1~4096, default:3`

Number of dimensions to output.

`-pcoa_estimator=Nei1972|Cavalli-Sforza1967|Reynolds1983|Nei1983|Euclidean|Goldstein1995|Nei1974|Roger1972|Slatkin_Nei1973|Slatkin_Weir1984|Slatkin_Hudson1992|Slatkin_Slatkin1995|Slatkin_Hedrick2005|Slatkin_Jost2008|Slatkin_Huang2021_homo|Slatkin_Huang2021_aniso|Reynolds_Nei1973|Reynolds_Weir1984|Reynolds_Hudson1992|Reynolds_Slatkin1995|Reynolds_Hedrick2005|Reynolds_Jost2008|Reynolds_Huang2021_homo|Reynolds_Huang2021_aniso, string, multiple selections, default:Nei1972`

Genetic distance estimators. Results of Euclidean distance are equivalent to PCA.

The results are saved in `*.pcoa.txt`. An example is shown as follows.

	A	B	C	D
14	-pcoa			
15	-pcoa_level=pop			
16	-pcoa_estimator=Euclidean			
17				
18				
19	Euclidean			
20	Total variance	0.055		
21	Variance	0.022	0.021	0.011
22	Pop	PC1	PC2	PC3
23	pop1	0.023	0.059	-0.152
24	pop2	0.137	0.121	0.090
25	pop3	0.052	-0.212	0.016
26	pop4	-0.211	0.032	0.046

Row 19 shows the total variance, and Row 20 shows the variance in each extracted principal coordinates. Rows 22-26 form a table showing the results of a PCoA by using the Nei1972 estimator of genetic distances, in which PC1, PC2 and PC3 are three principal coordinates of an individual, and each eigenvalue is the variance of the corresponding principal coordinates. Using a figure-plotting software, these individuals (populations or regions) can be shown in a 2D or 3D scatter-plot.

3.17 Hierarchical clustering

VCFPOP can also perform the hierarchical clustering based on the genetic distances. The parameters related to the hierarchical clustering are listed as follows.

`-cluster`

Perform hierarchical clustering for individuals, populations or regions. Results are saved in `*.cluster.txt` in standard tree format.

`-cluster_level=ind|pop|reg, string, multiple selections, default:ind`

Level of object in clustering: individuals, populations or regions.

`-cluster_method=NEAREST|FURTHEST|UPGMA|WPGMA|UPGMC|WPGMC|WARD, string, multiple selections, default:UPGMA`

Clustering methods.

`-cluster_estimator=Nei1972|Cavalli-Sforza1967|Reynolds1983|Nei1983|Euclidean|Goldstein1995|Nei1974|Roger1972|Slatkin_Nei1973|Slatkin_Weir1984|Slatkin_Hudson1992|Slatkin_Slatkin1995|Slatkin_Hedrick2005|Slatkin_Jost2008|Slatkin_Huang2021_homo|Slatkin_Huang2021_aniso|Reynolds_Nei1973|Reynolds_Weir1984|Reynolds_Hudson1992|Reynolds_Slatkin1995|Reynolds_Hedrick2005|Reynolds_Jost2008|Reynolds_Huang2021_homo|Reynolds_Huang2021_aniso, string, multiple selections, default:Nei1972`

Genetic distance estimators.

The results are saved in `*.cluster.txt` in the standard tree format. Such a file can be loaded by other tree viewing software. An example of tree files is shown below.

```
*polygene.out.cluster.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
##Hierarchical clustering calculated by vcfpop v1.0
#  Input: polygene.txt
#  Output: polygene.out
#  Time: 2022-03-25 16:01:22
#  Parameter: C:\vcfpop.exe
#    -g_output=polygene.out
#    -g_decimal=3
#    -g_progress=80
#    -g_nthread=5
#    -g_seed=1
#    -g_indtext=pop1:#1-#30 pop2:#31-#60 pop3:#61-#90 pop4:#91-#120
#    -g_input=polygene.txt
#    -g_format=polygene
#    -cluster
#    -cluster_method=NEAREST
#    -cluster_estimator=Cavalli-Sforza1967
#    -cluster_level=pop
#Level: population, genetic distance: Cavalli-Sforza1967, method: NEAREST
(((pop1:0.032,pop4:0.032):0.001,pop3:0.033):0.001,pop2:0.034);
第 3 行, 第 12 列   50%   Unix (LF)   ANSI
```

3.18 Bayesian clustering

Bayesian clustering estimates ancestral proportions of each individual by the Markov Chain Monte Carlo (MCMC) method. Following STRUCTURE (Pritchard *et al*. 2000), three models are included: ADMIXTURE (Pritchard *et al*. 2000), LOCPRIORI (Hubisz *et al*. 2009) and F model (Falush *et al*. 2003). The parameters related to Bayesian clustering are listed below.

`-structure`

Perform Bayesian clustering. Results are saved in `*.structure.txt` and `*.structure.k=5.r=1.txt`. The former is the summary and the latter is the result of each run.

Models:

In the non-ADMIXTURE model, each individual is assumed to only originate from one cluster, whilst in the ADMIXTURE model, each allele copy within an individual is assumed to originate from one cluster. The LOCPRIORI model takes advantage of the sample population of each individual to help infer weak population structure. The F model assumes that the allele frequencies in each cluster are correlated with those in the ancestral cluster. The parameters used to configure these models are as follows.

`-structure_admix=yes|no, string, defaul:no`

ADMIX model assumes each allele copy at each locus within the same individual can be drawn from the different clusters. Otherwise, all allele copies within the same individual are drawn from a cluster in each iteration.

`-structure_locpriori=yes|no, string, defaul:no`

LOCPRIORI model uses the sample population to cluster individuals.

`-structure_f=yes|no, string, defaul:no`

F model assumes the allele frequencies in each cluster are correlated with that in the ancestral population.

MCMC parameters:

Configures the parameters used for the MCMC algorithm.

`-structure_krange=[min_val,max_val], integer range, default:[1,5]`

Range of K (number of clusters).

`-structure_nburnin=1000~10000000, integer, default:10000`

Number of burn-in cycles.

`-structure_nreps=1000~10000000, integer, default:100000`

Number of iterations after burn-in.

`-structure_nthinning=1~10000, integer, default:1`

Sampling interval to dememorize.

`-structure_nruns=1~1000, integer, default:1`

Number of independent runs for each value of K.

`-structure_nadmburnin=100~10000000, integer, default:500`

Number of admixture burn-in cycles. This parameter is used for the non-ADMIXTURE and non-LOCPRIORI models, which generates a proper initial state to prevent the Markov chain to be blocked in the local maxima.

Misc:

λ is the Dirichlet parameter used to update the allele frequencies in each cluster. The value of λ can be updated during iterations if the option `-structure_inferlambda=yes` is used. The two options `-structure_stdlambda` and `-structure_maxlambda` configure the generation of a new λ. The option `-structure_difflambda=yes` allow using a different λ in different clusters. The parameters related to allele frequency and admixture burnin are as follows:

`-structure_lambda=0~10000, real, default:1`

The initial value of lambda.

`-structure_inferlambda=yes|no, string, default:no`

Updated lambda in each iteration.

`-structure_stdlambda=0~10000, real, default:0.3`

Standard deviation of new lambda.

`-structure_maxlambda=0~10000, real, default:10`

Maximum of new lambda.

`-structure_difflambda=yes|no, string, default:yes`

Use separate lambda for each cluster.

`-structure_diversity=yes|no, default:no`

Output diversity parameters for each cluster.

ADMIX:

In the ADMIXTURE model (Pritchard *et al*. 2000), α is the Dirichlet parameter and is used to update the admixture proportions of individuals, whose initial value is configured by `-structure_alpha`. The value of α will be updated according to `-structure_stdalpha` or `-structure_maxalpha` if `-structure_inferalpha=yes` is used. 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. The parameters related to the ADMIXTURE model are as follows:

`-structure_alpha=0~10000, real, default:1`

The initial alpha, the priori Dirichlet parameter of admixture proportions Q.

`-structure_inferalpha=yes|no, string, default:no`

Update alpha in ADMIX model.

`-structure_diffalpha=yes|no, string, default:no`

Use separate alpha for each cluster.

`-structure_uniformalpha=yes|no, string, default:yes`

Priori distribution for alpha, yes for uniform distribution and no for gamma distribution.

`-structure_stdalpha=0~10000, real, default:0.025`

Standard deviation of uniform priori distribution of alpha.

`-structure_maxalpha=0~10000, real, default:10`

Maximum of uniform priori distribution of alpha.

`-structure_alphapriora=0~10000, real, default:0.05`

One gamma priori distribution parameter.

`-structure_alphapriorb=0~10000, real, default:0.001`

The other gamma priori distribution parameter.

`-structure_metrofreq=0~1000000, integer, default:10`

Frequency of Metropolis-Hastings update of admixture proportions Q, set 0 to disable Metropolis-Hastings update.

LOCPRIORI:

In the LOCPRIORI model, a weak population structure can be inferred under the assistance of sample group information. There are three parameters r, η and γ in this mode. The parameter r is used to estimate the informativeness of the data for the sampling location. The parameter η and γ are used in the non-ADMIXTURE model to reflect the relative proportions of individuals assigned to a cluster. The values of η or γ is updated by drawing a new value from the uniform distribution $\eta \pm \text{eps}(\eta)$ or $\gamma \pm \text{eps}(\gamma)$. For the ADMIXTURE model, LOCPRIORI model adds one additional Dirichlet parameter, called the local α, and the original α is called the global α to distinguish from the local α. The global α and the local α are updated by drawing two new values from two normal distributions, respectively. The parameters related to the LOCPRIORI model are listed as follows:

`-structure_r=0~10000, real, default:1`

Initial value of r, where r evaluates the informativeness of data for the sampling location.

`-structure_maxr=0~10000, real, default:20`

Maximum of new r.

`-structure_epsr=0~10000, real, default:0.1`

Max step value of new r.

`-structure_epseta=0~10000, real, default:0.025`

Max step value of new eta for the non-ADMIXTURE model, where eta reflects the relative proportion of individuals assigned to a cluster.

`-structure_epsgamma=0~10000, real, default:0.025`

Max step value of new gamma for the non-ADMIXTURE model, where gamma reflects the relative proportion of individuals sampled from a location and assigned to a cluster.

FMODEL:

In the F model (Falush *et al.* 2003), the parameter F is a measure analogous to the Wright's F_{ST}, which is evaluated by the correlation between the allele frequencies in a cluster and those in the ancestral cluster. In the update of F, the new value of F is drawn from the normal distribution $\mathrm{N}(F, \mathrm{std}^2(F))$, and is evaluated according to the priori gamma distribution $\Gamma(A, B)$. The frequencies of alleles in the ancestral cluster are also updated.

`-structure_pmeanf=0~10000, real, default:0.01`

Priori mean F, where F is the amount of drift from the ancestral population to the cluster k in the F model.

`-structure_pstdf=0~10000, real, default:0.05`

Priori standard deviation of F.

`-structure_stdf=0~10000, real, default:0.05`

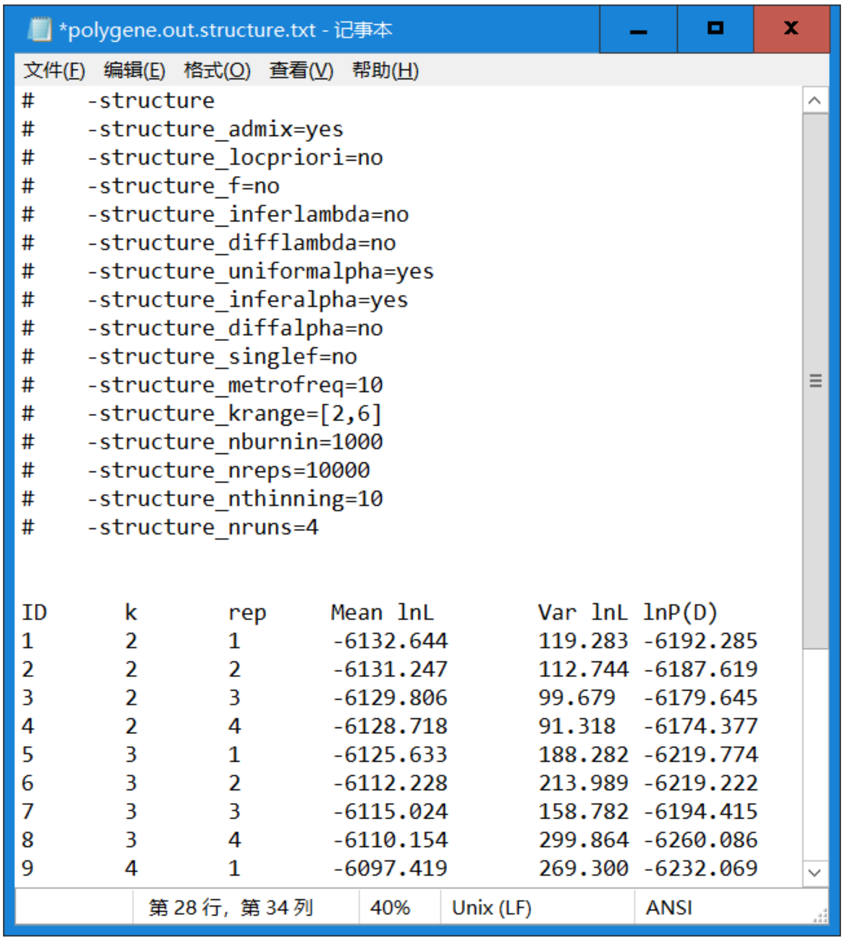
Standard deviation of new F.

`-structure_singlef=yes|no, string, default:no`

Use the same F in all clusters.

The results are saved in `*.structure.txt` and `*.structure.k=5.r=1.txt`. The former is a summary and the latter is the result of each run. An example of summary file is shown below.

3 Usage



*polygene.out.structure.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
#     -structure
#     -structure_admix=yes
#     -structure_locpriori=no
#     -structure_f=no
#     -structure_inferlambda=no
#     -structure_difflambda=no
#     -structure_uniformalpha=yes
#     -structure_inferalpha=yes
#     -structure_diffalpha=no
#     -structure_singlef=no
#     -structure_metrofreq=10
#     -structure_krange=[2,6]
#     -structure_nburnin=1000
#     -structure_nreps=10000
#     -structure_nthinning=10
#     -structure_nruns=4


ID       k        rep      Mean lnL         Var lnL  lnP(D)
1        2        1        -6132.644        119.283  -6192.285
2        2        2        -6131.247        112.744  -6187.619
3        2        3        -6129.806        99.679   -6179.645
4        2        4        -6128.718        91.318   -6174.377
5        3        1        -6125.633        188.282  -6219.774
6        3        2        -6112.228        213.989  -6219.222
7        3        3        -6115.024        158.782  -6194.415
8        3        4        -6110.154        299.864  -6260.086
9        4        1        -6097.419        269.300  -6232.069
```

第 28 行，第 34 列 40% Unix (LF) ANSI

The out data consist of the mean and variance of natural logarithms of likelihoods, the lnP(D) and the estimated parameters used in a specified model, where the estimated parameters are as follows: α for the ADMIXTURE model; r, η and γ for the LOCPRIORI model; r, the global α and the local α for the LOCPRIORI + ADMIXTURE model. Moreover, for the F model, the value of F for each cluster or the value of the global F is also included in the output. The next output consists of four tables: (i) proportion of membership of each pre-defined population in each cluster; (ii) inferred ancestry of individuals; (iii) allele-frequency divergence among populations (net nucleotide distance); and (iv) estimated heterozygosity in each cluster (if the option `-structure_diversity=yes` is used, otherwise the results will not be outputted).

An example of results of a run is shown in the following.

```
Parameters:
Seed=313261304
```

```
Model=NOADM,LOC,F
Number of cluster=4
Replicate=1
Run id=1
Mean value of ln likelihood=-9841.448123
Variance of ln likelihood=2628.160010
Estimated Ln Prob of Data=-11155.528129


Mean value of r=4.325618
Mean value of global eta
          Cluster
          1          2         3            4
          0.292969   0.148320  0.222200     0.336511
Mean value of local gamma for each location
          Cluster
Pop       1          2         3            4
DefPop    0.327675   0.196635  0.228096     0.247593
pop1      0.344244   0.060885  0.298596     0.296274
pop2      0.249133   0.191707  0.069353     0.489806
pop4      0.434370   0.039163  0.170569     0.355898
Mean value of Fst
          Cluster
          1          2         3            4
          0.660690   0.803997  0.668524     0.726568

Proportion of membership of each pre-defined population in each of the 4
clusters
          Cluster
Pop       1          2         3            4
DefPop    0.346233   0.179909  0.235900     0.237958
pop1      0.360000   0.000000  0.320000     0.320000
pop2      0.320000   0.120000  0.040000     0.520000
pop4      0.560000   0.000000  0.160000     0.280000

Inferred ancestry of individuals
                     Cluster
Ind       Pop        1         2            3 4
HG00096   pop1       0.845900  0.082100     0.065500        0.006500
HG00097   pop1       0.000000  0.000000     1.000000        0.000000
HG00099   pop1       0.008800  0.000200     0.986400        0.004600
...

Allele-frequency divergence among pops (Net nucleotide distance)
          Cluster
Cluster   1          2         3            4
1         0.000000   0.001296  0.012116     0.009320
2         0.001296   0.000000  0.012108     0.006990
3         0.012116   0.012108  0.000000     0.014451
4         0.009320   0.006990  0.014451     0.000000
```

4 Methodology

4.1 Haplotype extraction

There are some preparations to be conducted before the haplotypes are extracted: (i) exclude individuals with varying ploidy level in the same chromosome (or contig); (ii) sort the loci according to their chromosomes and positions; (iii) exclude the variants that are not typed in all individuals.

In the following description, we will use the ordinal numbers of variants in a chromosome to be their identifiers, and we let the variant i be ahead of the variant j (i.e., $i \leq j$). The procedures of our search algorithm for each chromosome are described as follows.

1. Set $i = 1$ and $j = 1$ in the beginning.
2. If the value of j exceeds the number of variants in this chromosome or contig, then terminate this algorithm.
3. Calculate the values of several parameters (the length of each haplotype, the number of variants, the number of alleles and the number of genotypes, the rate of missing data).
4. If any parameter exceeds the upper bound then increase i and go to step 2.
5. If any parameter exceeds the lower bound then increase j and go to step 2.
6. Combine all variants between the variants i and j, and extract the haplotypes.
7. Set i and j to the next applicable variant according to `-haplotype_interval`, and do to Step 2.

4.2 Genetic diversity indices

The definitions of genetic diversity indices and their calculating formulas are as follows.

- Ho, the observed heterozygosity, which is extended to polysomic inheritance. There are four hierarchical observed heterozygosities: the Ho for an individual at a locus is defined as the probability of randomly choosing two non-IBS (identical-by-state) alleles within this individual at this locus without replacement; the Ho for a population (a region or the total population) at a locus is the weighted average of those Ho for all individuals in this population (this region or the total population) at this locus. The expression of Ho for the individual i is

$$H_{Oi} = \frac{1}{\binom{v_i}{2}} \sum_{1 \le a < b \le v_i} \mathcal{B}_{A_{ila}A_{ilb}},$$

where v_i is the ploidy level of individual i at the target locus, A_{ia} and A_{ib} are respectively the a^{th} and b^{th} allele copies in the genotype of individual i at the target locus, and $\mathcal{B}_{A_{ia}A_{ib}}$ is a binary variable, such that $\mathcal{B}_{A_{ia}A_{ib}} = 0$ if $A_{ia} = A_{ib}$, or $\mathcal{B}_{A_{ia}A_{ib}} = 1$ if $A_{ia} \ne A_{ib}$. The expression of Ho for a population (a region or the total population) is

$$H_O = \frac{\sum_i \binom{v_i}{2} H_{Oi}}{\sum_i \binom{v_i}{2}},$$

where i is taken from all individuals in this population (this region or the total population).

- He, the expected heterozygosity, the same as the disomic inheritance, whose value at a locus is $1 - \sum_j^J P_j^2$, where J is the number of alleles at this locus, and P_j is the frequency of the j^{th} allele A_j at this locus.
- PIC, the polymorphic information content, the same as the disomic inheritance, whose value at a locus is $2\sum_{i=1}^{J}\sum_{j=i+1}^{J} P_i P_j (1 - P_i P_j)$.
- Ae, the effective number of alleles, like the disomic inheritance, whose value at a locus is $1/\sum_j P_j^2$.
- I, Shannon's Information Index, like the disomic inheritance, whose value at a locus is $-\sum P_j \ln P_j$.
- NE1P, the average probability of not excluding a candidate parent from the parentage of an arbitrary offspring, given only the genotype of this offspring. The same as the disomic inheritance, whose value at a locus is $4a_2 - 2a_2^2 - 4a_3 + 3a_4$, where $a_i = \sum_j P_j^i$, $i = 1, 2, 3, 4$.

- NE2P, 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. The same as the disomic inheritance, whose value at a locus is
$$2a_2^2 + 2a_2 - a_3 - 3a_2a_3 + 3a_5 - 2a_4.$$
- NEPP, the average probability of not excluding a candidate parent pair from the parentage of an arbitrary offspring, given only the genotype of this offspring. The same as the disomic inheritance, whose value at a locus is $4\mathrm{a}_5 - 4a_4 + 3a_6 + 8a_2^2 - 8a_2a_3 - 2a_3^2$.
- NEI, the average probability that the genotypes at a single locus do not differ between two unrelated individuals. This is the same as the disomic inheritance, whose value is $4a_2^2 - 3a_4$.
- NESI, the average probability that the genotypes at a single locus do not differ between two full siblings. This is the same as the disomic inheritance, whose value is $\frac{1}{4} + \frac{1}{2}a_2 + \frac{1}{2}a_2^2 - \frac{1}{4}a_4$.
- Fis, the inbreeding coefficient, which is defined as the probability of sampling two identical-by-descent (IBD) alleles from an individual without replacement, whose value at this locus is $1 - H_O/H_E$.

Fisher's G-test is used to test the genotypic distribution. The null hypothesis is that this distribution accords with the expected frequencies under a specific double-reduction model (e.g., RCS, PRCS, CES or PES). Here, if the ploidy level at a locus varies, Fisher's G-test will not be performed.

Assuming there are J alleles at the target locus, and the ploidy level is v. Then the number of possible genotypes at this locus is $\binom{v+J-1}{v}$. Because the numbers of allele copies are fixed, the distribution of genotypes is subject to the following J constraints:
$$f_j = \sum_i v\,\Pr(A_j\,|G_i), \qquad j = 1, 2, \cdots, J,$$
where $\Pr(A_j\,|G_i)$ is the frequency of j^{th} allele A_j in the genotype G_i of individual i. Therefore, the degrees-of-freedom are $\binom{v+J-1}{v} - J$. If the expected occurrence of any genotype is less than 5, then two minor alleles are collapsed and treated as one allele.

The expected occurrence of each genotype can be calculated under a specific double-reduction model, such as the model of RCS, PRCS, CES or PES (Huang *et al.* 2019). Fisher's G statistic is then calculated by

$$G = \sum_i 2O_i \ln(O_i/E_i),$$

where O_i and E_i are the values of the observed and the expected occurrences of the genotype G_i. Finally, the P-value can be obtained by the right-tail probability of the Chi-squared distribution.

4.3 Individual statistics

In diploids, the heterozygosity-index (H-index) for a genotype is a binary variable, whose value is equal to one for a heterozygote, or equal to zero for a homozygote. In polyploids, the H-index for a genotype G is defined as the probability of randomly sampling two different IBS alleles within G without replacement. For example, the H-index for the tetraploid genotype $AABB$ is 2/3. The H-index for an individual is defined as the arithmetic mean of H-indices for the genotypes of this individual across all loci.

The likelihood $\mathcal{L}$ of an individual is the product of frequencies of genotypes of this individual in this target population across all loci, symbolically

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

where L is the number of loci. The logarithm form of $\mathcal{L}$ is

$$\ln \mathcal{L} = \sum_{l=1}^{L} \ln \Pr(G_l).$$

The inbreeding coefficient of an individual is defined as the probability of sampling two IBD alleles from this individual without replacement, denoted by f. The kinship coefficient within an individual itself is defined as the probability of sampling two IBD alleles from this individual with replacement, denoted by θ. According to these definitions, the following relationship holds:

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

where v is the ploidy level of the individual. Therefore, the inbreeding coefficient f can be converted from the kinship coefficient θ so long as θ is estimated.

VCFPOP provides three method-of-moment estimators to estimate the kinship coefficient θ:

- Ritland's (1996) estimator

$$\hat{\theta}_{\mathrm{RI}} = \frac{\sum_l \left[\left(\sum_j P_{alj} P_{blj} / P_{lj}\right) - 1\right]}{\sum_l (J_l - 1)};$$

- Loiselle's (1995) estimator

$$\hat{\theta}_{\mathrm{LO}} = \frac{\sum_l \sum_j \left(P_{alj} - P_{lj}\right)\left(P_{blj} - P_{lj}\right)}{\sum_l \sum_j P_{lj}\left(1 - P_{lj}\right)};$$

- Weir's (1996) estimator

$$\hat{\theta}_{\mathrm{WE}} = \frac{\sum_l \sum_j \left(P_{alj} P_{blj} - P_{lj}^2\right)}{L - \sum_l \sum_j P_{lj}^2},$$

where J_l is the number of alleles at locus l, A_{lj} is the j^{th} allele at locus l, P_{alj} (or P_{blj}) is the frequency of A_{lj} in the individual a (or b), and P_{lj} is the frequency of A_{lj} in the reference population.

4.4 Genetic differentiation

The F_{ST} estimators are Nei's (1973) G_{ST}, Weir & Cockerham's (1984) θ, Hudson *et al.*'s (1992) F_{ST}, Slatkin's (1995) R_{ST}, Hedrick's (2005) G'_{ST}, Jost's (2008) D and Huang *et al.*'s (2021) estimators. Their formulas are listed as follows.

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

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

where L is the number of loci, H_{Tl} is the expected heterozygosity in the total population at locus l, and H_{Sl} is the weighted average of expected heterozygosities in all populations at locus l, with the number of haplotypes in each population as a weight, whose expressions are as follows:

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

$$H_{Sl} = \frac{\sum_p v_p\left(1 - \sum_j^{J_l} P_{plj}^2\right)}{v_t},$$

in which J_l is the number of alleles at locus l, p is taken from all populations, P_{lj} (or P_{plj}) is the frequency of allele A_{lj} in the total population (or the population p), v_p (or v_t) is the number of haplotypes in the population p (or the total population). Note that any monomorphic loci are excluded from the calculation.

- Weir & Cockerham's (1984) θ can only be used for diploids, and is calculated by

$$\theta = \frac{\sum_l^L \sum_j^{J_l} a_{lj}}{\sum_l^L \sum_j^{J_l}\left(a_{lj} + b_{lj} + c_{lj}\right)},$$

where a_{lj}, b_{lj} and c_{lj} are respectively the variance components among populations, among individuals and within individuals for the allele A_{lj}, whose expressions are

$$a_{lj} = \frac{\bar{n}_l}{n_{cl}}\left\{s_{lj}^2 - \frac{1}{\bar{n}_l - 1}\left[\bar{P}_{lj}\left(1 - \bar{P}_{lj}\right) - \frac{r-1}{r}s_{lj}^2 - \frac{\bar{h}_{lj}}{4}\right]\right\},$$

$$b_{lj} = \frac{\bar{n}_l}{\bar{n}_l - 1}\left[\bar{P}_{lj}\left(1 - \bar{P}_{lj}\right) - \frac{r-1}{r}s_{lj}^2 - \frac{2\bar{n}_l - 1}{4\bar{n}_l}\bar{h}_{lj}\right],$$

$$c_{lj} = \bar{h}_{lj}/2,$$

in which the symbol r denotes the number of populations, and the other symbols are explained as follows:

$\bar{n}_l$ is the average sample size per population at locus l, i.e., $\bar{n}_l = \frac{\sum_p n_{pl}}{r}$, in which p is taken from all populations, n_{pl} is the sample size of population p at locus l;

n_{cl} is an intermediate variable, whose expression is $n_{cl} = \frac{r\bar{n}_l - \sum_p n_{pl}^2/r\bar{n}_l}{r-1}$;

$\bar{P}_{lj}$ is the weighted average of frequencies of A_{lj}, i.e., $\bar{P}_{lj} = \frac{\sum_p n_{pl}P_{plj}}{\sum_p n_{pl}}$;

s_{lj}^2 is the sampling variance of frequencies of A_{lj}, i.e., $s_{lj}^2 = \frac{r\left(\overline{P^2}_{lj} - \bar{P}_{lj}^2\right)}{r-1}$, in which $\overline{P^2}_{lj}$ is the weighted average of frequency squares of A_{lj}, i.e., $\overline{P^2}_{lj} = \frac{\sum_p n_{pl} P_{plj}^2}{\sum_p n_{pl}}$;

$\bar{h}_{lj}$ is the average heterozygosity for A_{lj}, i.e., $\bar{h}_{lj} = 2\left(\bar{P}_{lj} - \overline{P^2}_{lj}\right)$.

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

$$H_{wl} = \frac{\sum_p \sum_{j_1<j_2} v_{pl}^2 P_{plj_1} P_{plj_2} {}^{\mathrm{IAM}}d_{lj_1lj_2}^2}{\sum_p v_{pl}(v_{pl}-1)/2},$$

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

$$H_{bl} = \frac{\sum_{j_1<j_2} v_{tl}^2 P_{lj_1} P_{lj_2} {}^{\mathrm{IAM}}d_{lj_1lj_2}^2 - \sum_p \sum_{j_1<j_2} v_{pl}^2 P_{plj_1} P_{plj_2} {}^{\mathrm{IAM}}d_{lj_1lj_2}^2}{v_{tl}(v_{tl}-1)/2 - \sum_p v_{pl}(v_{pl}-1)/2}.$$

- 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

$$S_{Tl} = \frac{\sum_{j_1<j_2} v_{tl}^2 P_{lj_1} P_{lj_2} {}^{\mathrm{SMM}}d_{lj_1lj_2}^2}{v_{tl}(v_{tl}-1)/2},$$

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

$$S_{Wl} = \frac{\sum_p \sum_{j_1<j_2} v_{pl}^2 P_{plj_1} P_{plj_2} {}^{\mathrm{SMM}}d_{lj_1lj_2}^2}{\sum_p v_{pl}(v_{pl}-1)/2},$$

in v_{tl} (or v_{pl}) is the number of allele copies at the l^{th} locus and in the total population (or in the p^{th} 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 p^{th} population), and ${}^{\mathrm{SMM}}d_{lj_1lj_2}$ is the SMM distance between the alleles A_{lj_1} and A_{lj_2}.

- Hedrick's (2005) estimator G'_{ST} is the standardization of Nei's (1973) estimator G_{ST}, obtained by dividing the theoretical maximum of G_{ST}, whose expression is

$$G'_{ST} = \frac{\sum_l^L (H_{Tl} - H_{Sl})(S - 1 + H_{Sl})}{(S - 1)\sum_l^L (1 - H_{Sl}) H_{Tl}},$$

  where S is the number of populations.

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

$$D = \frac{S \sum_l^L (H_{Tl} - H_{Sl})}{(S - 1)\sum_l^L (1 - H_{Sl})}.$$

- Huang *et al.* (2021) homo and aniso estimators are based on AMOVA, where

$$F_{ST} = \frac{\hat{\sigma}_{AP}^2}{\hat{\sigma}_{AP}^2 + \hat{\sigma}_{AI}^2 + \hat{\sigma}_{WI}^2},$$

  The variance components for homoploid and anisoploid models are calculated by different weighting scheme, see Section 4.6 for detail.

The test of genetic differentiation is performed by Fisher's *G*-test. This test can be applied to either multiple loci or a single locus. For multiple loci, the values of the statistic G (or the values of the degree of freedom d. f.) will be summed.

4.5 Genetic distance

VCFPOP can calculate various genetic distances between individuals, populations, and regions. These measures are all based on allele frequencies. The following items show these measures together with the corresponding references and calculation methods.

- Nei's (1972) standard genetic distance D_S. This measure assumes that the genetic differences are caused by both mutation and genetic drift. If the mutation rate is constant, the distance D_S between the populations x and y is proportional to divergence time, whose calculating formula is

$$D_S = -\ln \frac{J_{xy}}{\sqrt{J_x J_y}},$$

where $J_x = \sum_l^L \sum_j^{J_l} P_{xlj}/L$, $J_y = \sum_l^L \sum_j^{J_l} P_{ylj}/L$, $P_{xy} = \sum_l^L \sum_j^{J_l} P_{xlj} P_{ylj}/L$ (J_l is the number of alleles at locus l), and P_{xlj} and P_{ylj} are the frequencies of the j^{th} alleles in the individuals/populations/regions x and y at locus l, respectively.

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

$$D_{CH} = \frac{2}{\pi} \sqrt{2\left(1 - \frac{1}{L}\sum_l^L \sum_j^{J_l} \sqrt{P_{xlj} P_{ylj}}\right)}.$$

- Reynolds *et al.*'s (1983) distance θ_w. This measure assumes that genetic differences occur only due to genetic drift without any mutations. This measure is used to estimate the kinship coefficient θ which provides a measure of the genetic divergence, where the expression of θ is

$$\theta_w = \sqrt{\frac{\sum_l^L \sum_j^{J_l} \left(P_{xlj} - P_{ylj}\right)^2}{2\sum_l^L \left(1 - \sum_j^{J_l} P_{xlj} P_{ylj}\right)}}.$$

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

$$D_A = 1 - \frac{1}{L}\sum_l^L \sum_j^{J_l} \sqrt{P_{xlj} P_{ylj}}.$$

- Euclidean distance D_{EU}. This is a usual measure based on Euclidean space, where the frequencies of alleles in a population at a locus form a vector in this space. The distance D_{EU} is calculated by

$$D_{EU} = \sqrt{\sum_{l}^{L}\sum_{j}^{J_l}\left(P_{xlj} - P_{ylj}\right)^2}.$$

- Goldstein's (1995) distance $(\delta\mu)^2$. This measure was developed based on a stepwise mutation model (SMM), used specifically for the microsatellite markers, the formula of which is:

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

where μ_{xl} (or μ_{yl}) is the average allele size in the population x (or y) at locus l.

- Nei's (1974) minimum genetic distance D_M. This measure assumes that the genetic differences arise due to the mutation and the genetic drift, the formula being:

$$D_M = \frac{J_x + J_y}{2} - J_{xy},$$

where the meanings of J_x, J_y and J_{xy} are as indicated in the first item of this section.

- 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

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

- Reynolds *et al.*'s (1983) distance D_{RA}. This measure is defined as the divergence time. If two diploid populations of a constant size N diverged t generations ago, then the divergence time is $t/2N$. In this scenario, Wright's F_{ST} between these two populations can be expressed as $F_{ST} = 1 - \left(1 - \frac{1}{2N}\right)^t$. Because $1 - \left(1 - \frac{1}{2N}\right)^t \approx 1 - e^{-t/2N}$, it is easy to derive that $t/2N \approx -\ln\left(1 - F_{ST}\right)$. Therefore, the distance D_{RA} can be calculated as:

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

- Slatkin's (1995) linearized distance D_{Sl}. This measure is also defined as the divergence time. Slatkin considered a simple demographic model, in which two diploid populations of size N diverged at τ generations ago from a population of identical size, and have remained isolated ever since, without exchanging any migrants. The divergence time is thus $\tau/2N$. 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 $\tau/2N = F_{ST}/(1 - F_{ST})$. Therefore, the distance D_{Sl} can be calculated as:

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

4.6 Analysis of molecular variance

Homoploid method

The homoploid method follows that of GENALEX (Peakall & Smouse 2006). In this method, all loci are treated as one dummy locus, and the j^{th} haplotype of an individual at this dummy locus consists of the j^{th} allele copies within this individual across all loci. In this method, the missing data of each individual are weighted according to the allele frequencies in the sampling population of this individual.

The square of the genetic distance $d_{hh'}$ between two haplotypes h and h' is the sum of squares of either the IAM distances or the SMM distances of allele pairs across all loci, which is:

$$d_{hh'}^2 = \sum_l d_{A_{hl}A_{h'l}}^2,$$

where A_{hl} and $A_{h'l}$ are a pair of allele copies at locus l with the former in h and the latter in h'. Note that the SMM distance can only be applied for non-VCF/BCF input files, and the identifier of each allele must be its size.

We use the symbol SS_{WI} to denote the sum $\sum_{i\in\mathbf{I}} \bar{S}_i$, where $\mathbf{I}$ is the collection of all individuals, and $\bar{S}_i$ is the arithmetic mean of the squares of genetic distances between haplotypes within the individual i. For convenience, we call roughly SS_{WI} as the sum of squares of haplotype distances within individuals. Similarly, the symbol SS_{WP} (SS_{WR} or SS_{TOT}) is roughly called the sum of squares of haplotype distances within populations (regions or the total population), denoted by:

$$\mathrm{SS_{WI}} = \frac{1}{2}\sum_{i\in\mathbf{I}}\sum_{h,h'\in i}\frac{d_{hh'}^2}{v_i},$$

$$\mathrm{SS_{WP}} = \frac{1}{2}\sum_{p\in\mathbf{P}}\sum_{h,h'\in p}\frac{d_{hh'}^2}{v_p},$$

$$\mathrm{SS_{WR}} = \frac{1}{2}\sum_{r\in\mathbf{R}}\sum_{h,h'\in r}\frac{d_{hh'}^2}{v_r},$$

$$\mathrm{SS_{TOT}} = \frac{1}{2}\sum_{h,h'}\frac{d_{hh'}^2}{v_t},$$

where v_i, v_p, v_r and v_t are the numbers of haplotypes within the individual i, the population p, the region r and the total population in turn, and **P** and **R** are the collection of all populations and all regions, respectively.

The degrees-of-freedom and the sum of squares at each hierarchy level are listed in the next table.

Source	d.f.	SS
Within individuals	$N_\mathrm{H} - N_\mathrm{I}$	$\mathrm{SS_{WI}}$
Within populations	$N_\mathrm{H} - N_\mathrm{P}$	$\mathrm{SS_{WP}}$
Within regions	$N_\mathrm{H} - N_\mathrm{R}$	$\mathrm{SS_{WR}}$
Total population	$N_\mathrm{H} - 1$	$\mathrm{SS_{TOT}}$

Here, N_H, N_I, N_P and N_R are the total number of haplotypes, individuals, populations and regions, respectively.

The expressions of expected SS at each hierarchy level are listed in the following equations:

$$\mathrm{E(SS_{WI})} = \sigma_\mathrm{WI}^2(N_\mathrm{H} - N_\mathrm{I}),$$

$$\mathrm{E(SS_{WP})} = \sigma_\mathrm{WI}^2(N_\mathrm{H} - N_\mathrm{P}) + \sigma_\mathrm{AI}^2\left(N_\mathrm{H} - \sum_{p\in\mathbf{P}}\sum_{i\in p}\frac{v_i^2}{v_p}\right),$$

$$\mathrm{E(SS_{WR})} = \sigma_\mathrm{WI}^2(N_\mathrm{H} - N_\mathrm{R}) + \sigma_\mathrm{AI}^2\left(N_\mathrm{H} - \sum_{r\in\mathbf{R}}\sum_{i\in r}\frac{v_i^2}{v_r}\right) + \sigma_\mathrm{AP}^2\left(N_\mathrm{H} - \sum_{r\in\mathbf{R}}\sum_{p\in r}\frac{v_p^2}{v_r}\right),$$

$$\mathrm{E(SS_{TOT})} = \sigma_\mathrm{WI}^2(N_\mathrm{H} - 1) + \sigma_\mathrm{AI}^2\left(N_\mathrm{H} - \sum_{i\in\mathbf{I}}\frac{v_i^2}{v_t}\right) + \sigma_\mathrm{AP}^2\left(N_\mathrm{H} - \sum_{p\in\mathbf{P}}\frac{v_p^2}{v_t}\right) + \sigma_\mathrm{AR}^2\left(N_\mathrm{H} - \sum_{r\in\mathbf{R}}\frac{v_r^2}{v_t}\right),$$

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

If we regard these variances as unknowns, the above expressions form a linear equation set. Therefore, we can obtain the values of these variances by solving this equation set. It is worth noting that because this variance estimator is unbiased, the estimated value of a variance may be negative when the true value of this variance is close to zero.

Imitating the above method, for any positive integer M, we 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 i^{th} 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 related expressions can be expressed as the form of matrices as follows:

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

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

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

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

$$\hat{F}_{ij}=1-\frac{\sum_{k=1}^{i}\hat{\sigma}_k^2}{\sum_{k=1}^{j}\hat{\sigma}_k^2},\qquad 1\le i<j\le M.$$

Following Excoffier *et al.* (1992), the differentiation test is performed for each *F*-statistic independently. The null hypothesis is that each *F*-statistic (e.g. F_{ij}) is zero, which is equivalent to that the hierarchy j is real but hierarchy i is artificial. To obtain the null distribution of $\hat{F}_{ij}$, we randomly permute hierarchy $i-1$ within hierarchy j 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.

Anisoploid method

The anisoploid method is similar to the locus-by-locus AMOVA in ARLEQUIN (Excoffier & Lischer 2010), where the SS and equations of E(SS) are calculated at each locus and summed across loci. This method does not extract the dummy haplotypes from each individual, and supports anisoploids. This method also supports both the IAM and the SMM distances between a pair of alleles at the same locus without any additional restrictions.

Following the homoploid method, the matrices $\mathbf{C}_l$ and $\hat{\mathbf{S}}_l$ at each locus can be obtained, and VCFPOP sums these matrices across loci, and next solves the linear equation set to obtain the variance components $\mathbf{\Sigma}$. If these variances have been obtained, the *F*-statistics can be calculated by using the relational expressions described above. Moreover, the hierarchies at each locus will be permuted independently in the test for *F*-statistics. Note that because any missing data are not considered in the anisoploid method, the two values of $N_{\mathrm{A}l}$ ($N_{\mathrm{I}l}$, $N_{\mathrm{P}l}$ or $N_{\mathrm{R}l}$) before and after a permutation may not be identical.

Because this method permutes vessels for each locus and is extreme slow. VCFPOP uses a pseudo-permutation method to solve this problem, which first perform a small number of permutations (e.g., 100) for each locus, then subsamples one permutation at each locus to generate results for each permutation. The number of permutations is suggested to be greater than 100.

M**aximum-likelihood method**

A reverse procedure is used, in which the F-statistics are estimated first, 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-statistics can be obtained from F_{12}, F_{13}, …, F_{1M}.

The global likelihood for individuals at the i^{th} hierarchy is the product of frequencies of all 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 i^{th} hierarchy, N_{V_i} is the number of individuals in V_i, $\mathcal{P}_{V_i l j}$ is the phenotype of j^{th} individual in V_i and at the l^{th} locus, $\mathbf{p}_{V_i l}$ is the vector consisting of the frequencies of all alleles in V_i and at the l^{th} 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

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

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

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

where c_j in the multinomial coefficient is the number of the j^{th} allele copies in G ($j = 1, 2, \cdots, J_l$), $\alpha_{1i} = 1/F_{1i} - 1$ and $\alpha_{1ij} = \alpha_{1i} P_{V_i l j}$ ($P_{V_i l j}$ is the j^{th} 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 using the following additional constraint:

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

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

$$\mathrm{SS}_M = |V_M| \sum_{1 \le i < j \le J} P_{Mi} P_{Mj} d_{A_i A_j}^2,$$

where P_{Mi} (or P_{Mj}) is the frequency of A_i (or A_j) in V_M. The permutation test is the same as the homoploid method, but the F_I. (e.g., F_{IS}, F_{IC} and F_{IT}) are not tested because the haplotype is not permuted if there are anisoploids.

4.7 Population assignment

For the population assignment, the likelihood $\mathcal{L}$ of an individual for a target population is the probability of observing the genotypic data of this individual under the hypothesis that this individual originated from the target population. Such a probability is defined as the product of the probabilities of observing the genotypic data of this individual across all loci under the above hypothesis, symbolically $\mathcal{L} = \prod_l \Pr(G_l)$, in which G_l is the genotype at locus l. A higher likelihood implies that the hypothesis is more reliable, i.e., the individual is more likely to originate from the target population.

Each individual is assigned to the population with the highest likelihood. The difference between the highest and the second highest common logarithm of likelihoods of an individual for all target populations is called the LOD score of this individual. Such a LOD score can be used to evaluate the accuracy of highest likelihood, and the greater the LOD score, the higher the accuracy of highest likelihood. For example, if the LOD score is up to 3, it means that $\mathcal{L}_{\mathrm{most}}$ is 1000 times of $\mathcal{L}_{\mathrm{second}}$, where $\mathcal{L}_{\mathrm{most}}$ (or $\mathcal{L}_{\mathrm{second}}$) is the likelihood of this individual for the most (or the second most) possible target population.

The likelihood of an individual for a target population will be equal to zero if this individual carries any alleles absent in 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 VCFPOP, the mistyping rate e is added into the calculation of genotypic frequency, where e is the probability of a genotype being mistyped. When a genotype is mistyped, the observed genotype is randomly drawn according to the genotypic frequency in the total population. Therefore, the posterior probability $\Pr(G)$ that a given genotype G is observed from the population p is

$$\Pr(G) = \Pr(G|p)(1-e) + [1 - \Pr(G|p)]\, e \Pr(G|t),$$

where $\Pr(G|p)$ (or $\Pr(G|t)$) is the frequency of G in the population p (or the total population).

4.8 Kinship coefficient

There are two kinds of kinship coefficients. One is within an individual, denoted by θ, whose definition has been given in Section 4.3. The other is between two individuals, which is defined as the probability of sampling two IBD alleles from these two individuals (each sampled from one individual), still denoted by θ.

For these two kinds of kinship coefficients, their estimators have the same calculating formulas. VCFPOP provides three estimators of method-of-moment to estimate these two kinds of kinship coefficients, and their calculating formulas have been listed in Section 4.3. After a kinship coefficient is calculated, it can be converted as either an inbreeding coefficient (see Section 4.3) or a relatedness coefficient (see Section 4.9).

4.9 Relatedness coefficient

The relatedness coefficient r is the correlation of allele frequencies between two individuals. The relatedness estimators can be roughly divided into two categories: one is the estimators of method-of-moment, and the other is the estimators of maximum-likelihood. The former category of estimators is unbiased, while the latter has a slight RMSE (the root mean square error).

In diploids, if the inbreeding is absent, the relatedness coefficient r can be expressed as

$$r = \phi/2 + \Delta,$$

where ϕ is the probability that two individuals share one pair of IBD alleles, called the two-gene coefficient, and Δ is the probability that two individuals share two pairs of IBD alleles, called the four-gene coefficient. Therefore, in the absence of inbreeding, the relatedness coefficient r can be converted from ϕ and Δ by using the above relational expression.

In the absence of inbreeding, the calculating formula of each relatedness estimator of method- of-moment is a linear equation set with ϕ and Δ as the unknowns. VCFPOP provides several such estimators, whose calculating formulas are listed below.

- Lynch & Ritland's (1999) estimator: if G_x is a homozygote, it can be calculated by

$$\begin{bmatrix} \Pr(G_y = A_iA_i|G_x = A_iA_i) \\ \Pr(G_y = A_iA_x|G_x = A_iA_i) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ 2p_ip_x \end{bmatrix} + \begin{bmatrix} 1 - p_i^2 & p_i - p_i^2 \\ -2p_ip_x & p_x - 2p_ip_x \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix};$$

if G_x is a heterozygote, it can be calculated by

$$\begin{bmatrix} \Pr(G_y = A_iA_i|G_x = A_iA_j) \\ \Pr(G_y = A_jA_j|G_x = A_iA_j) \\ \Pr(G_y = A_iA_j|G_x = A_iA_j) \\ \Pr(G_y = A_iA_x|G_x = A_iA_j) \\ \Pr(G_y = A_jA_x|G_x = A_iA_j) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ p_j^2 \\ 2p_ip_j \\ 2p_ip_x \\ 2p_jp_x \end{bmatrix} + \begin{bmatrix} -p_i^2 & \frac{1}{2}p_i - p_i^2 \\ -p_j^2 & \frac{1}{2}p_j - p_j^2 \\ 1 - 2p_ip_j & \frac{1}{2}p_i + \frac{1}{2}p_j - 2p_ip_j \\ -2p_ip_x & \frac{1}{2}p_x - 2p_ip_x \\ -2p_jp_x & \frac{1}{2}p_x - 2p_jp_x \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where A_x is an allele not appearing in G_x, p_i (or p_j) is the frequency of allele A_i (or A_j), and $p_x = 1 - p_i$ if $G_x = A_iA_i$, or $p_x = 1 - p_i - p_j$ if $G_x = A_iA_j$ ($i \neq j$).

- Wang's (2002) estimator is calculated by

$$\begin{bmatrix} \Pr(S=1) \\ \Pr(S=3/4) \\ \Pr(S=1/2) \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & a_2-\lambda_1 \\ -\lambda_2 & 2a_2 - 2a^3 - \lambda_2 \\ -\lambda_3 & 1 - 3a_2 + 2a_3 - \lambda_3 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = 2a_2^2 - a_4$, $\lambda_2 = 4a_3 - 4a_4$, $\lambda_3 = 4a_2 - 4a_2^2 - 8a_3 + 8a_4$, $a_k = \sum_j p_j^k$ ($k = 1, 2, 3, 4$), and S is the similarity index of the allelic pattern of G_x and G_y, i.e.,

$$S = \begin{cases} 1 & \text{if } G_x = A_iA_i, G_y = A_iA_i \text{ or if } G_x = A_iA_j, G_y = A_iA_j, \\ 3/4 & \text{if } G_x = A_iA_i, G_y = A_iA_j, \\ 1/2 & \text{if } G_x = A_iA_j, G_y = A_iA_k, \\ 0 & \text{otherwise,} \end{cases}$$

in which i, j and k are different with each other.

- Thomas's (2010) estimator is calculated by

$$\begin{bmatrix} \Pr(S=1) \\ \Pr(S=3/4 \text{ or } 1/2) \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & a_2-\lambda_1 \\ -\lambda_2 & 1-a_2-\lambda_2 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = 2a_2^2 - a_4$, $\lambda_2 = 4a_2 - 4a_2^2 - 4a_3 + 4a_4$.

- Huang *et al.* (2016a) estimator: if G_x is homozygous, it can be calculated by

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \end{bmatrix} = \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \left(\begin{bmatrix} p_i^2 \\ 2p_ip_x \\ 0 \end{bmatrix} + \begin{bmatrix} 1-p_i^2 & p_i - p_i^2 \\ -2p_ip_x & p_x - 2p_ip_x \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix} \right);$$

if G_x is heterozygous, it can be calculated by

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \end{bmatrix} = \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \left(\begin{bmatrix} 2p_ip_j \\ p_i^2 + p_j^2 \\ 2p_x(p_i+p_j) \end{bmatrix} + \begin{bmatrix} 1-2p_ip_j & \frac{1}{2}(p_i+p_j) - 2p_ip_j \\ -p_i^2 - p_j^2 & \frac{1}{2}(p_i+p_j) - p_i^2 - p_j^2 \\ -2p_x(p_i+p_j) & p_x - 2p_x(p_i+p_j) \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix} \right).$$

- Huang *et al.* (2016b) estimator is calculated by

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \end{bmatrix} = \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \left(\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & a_2-\lambda_1 \\ -\lambda_2 & 2a_2-a_3-\lambda_2 \\ -\lambda_3 & 1-3a_2+2a_3-\lambda_3 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix} \right),$$

Where $\lambda_1 = 2a_2^2 - a_4$, $\lambda_2 = 4a_3 - 4a_4$ and $\lambda_3 = 4a_2 - 4a_2^2 - 8a_3 + 8a_4$.

In the next two estimators, we omit the expressions of linear equation sets for simplicity.

- Queller & Goodnight's (1989) estimator is calculated by $\hat{r} = (\hat{r}_{xy} + \hat{r}_{yx})/2$, where

$$\hat{r}_{xy} = \frac{K_{ac} + K_{ad} + K_{bc} + K_{bd}}{2(1 + K_{ab} - p_a - p_b)},$$

$$\hat{r}_{yx} = \frac{K_{ac} + K_{ad} + K_{bc} + K_{bd}}{2(1 + K_{cd} - p_c - p_b)},$$

  in which $\hat{r}_{xy}$ (or $\hat{r}_{yx}$) is the estimated value with G_x (or G_y) as the reference individual, a and b are the two alleles in G_x, c and d are those in G_y, and $K_{a_1a_2}$ is a Kronecker operator, such that $K_{a_1a_2} = 1$ if $a_1 = a_2$, or $K_{a_1a_2} = 0$ if $a_1 \neq a_2$. Note that this estimator cannot be applied for a biallelic marker. That is because at least one of the denominators of the above fractions is equal to zero if G_x or G_y is heterozygous.
- Li *et al.*'s (1993) estimator is calculated by

$$r = \frac{S - S_0}{1 - S_0},$$

  where $S_0 = \sum_j^J p_j^2 (2 - p_j)$.

For diploids, VCFPOP also provides two relatedness estimators of maximum likelihood: one is Milligan's (2003) estimator, and the other is Anderson & Weir's (2007) estimator. The likelihood $\mathcal{L}(G_x, G_y | \phi, \Delta)$ of observing the genotypic data of a pair of individuals G_x and G_y conditional on ϕ and Δ is defined in these two estimators, whose expression for all patterns of genotypic pairs is

$$\mathcal{L}(G_x, G_y \mid \phi, \Delta) = \begin{cases} p_i^2\Delta + p_i^3\phi + p_i^4(1-\phi-\Delta) & \text{if } G_x = A_iA_i, G_y = A_iA_i, \\ p_i^2p_j^2(1-\phi-\Delta) & \text{if } G_x = A_iA_i, G_y = A_jA_j, \\ p_i^2p_j\phi + 2p_i^3p_j(1-\phi-\Delta) & \text{if } G_x = A_iA_i, G_y = A_iA_j, \\ 2p_i^2p_jp_k(1-\phi-\Delta) & \text{if } G_x = A_iA_i, G_y = A_jA_k, \\ p_i^2p_j\phi + 2p_i^3p_j(1-\phi-\Delta) & \text{if } G_x = A_iA_j, G_y = A_iA_i, \\ 2p_i^2p_jp_k(1-\phi-\Delta) & \text{if } G_x = A_jA_k, G_y = A_iA_i, \\ 2p_ip_j\Delta + p_ip_j(p_i+p_j)\phi + 4p_i^2p_j^2(1-\phi-\Delta) & \text{if } G_x = A_iA_j, G_y = A_iA_j, \\ p_ip_jp_k\phi + 4p_i^2p_jp_k(1-\phi-\Delta) & \text{if } G_x = A_iA_j, G_y = A_iA_k, \\ 4p_ip_jp_kp_l(1-\phi-\Delta) & \text{if } G_x = A_iA_j, G_y = A_kA_l. \end{cases}$$

A numerical algorithm (e.g., Nelder-Mead simplex algorithm) is used to search the optimal ordered couple (ϕ, Δ) that maximize this likelihood, and next the relatedness coefficient r will be converted from this ordered couple (ϕ, Δ), i.e., $r = \phi/2 + \Delta$, which is the maximum likelihood estimate. For Anderson & Weir's (2007) estimator, the higher-order relatedness coefficients ϕ and Δ are subject to an additional constraint: $4\Delta(1-\Delta) < \phi^2$.

For polyploids, VCFPOP provides two relatedness estimators (Huang *et al*. 2015a; Huang *et al*. 2014) and three kinship estimators (Loiselle *et al*. 1995; Ritland 1996; Weir 1996).

The polyploid method-of-moment estimator is a modification of Huang *et al*.'s (2016a) estimator. In this estimator, all possible genotype patterns are enumerated (there are 5, 11 and 22 reference genotype modes for tetraploids, hexaploids and octoploids, respectively). The definition of similarity index S is modified as the number of alleles that are identical-by-state between two individuals, with each allele being counted only once. Therefore, there are $v + 1$ possible values of S, i.e., the range of S is $1, \frac{v-1}{v}, \frac{v-2}{v}, \cdots, 0$, where v is the ploidy level. Let $\mathrm{E}(S^k)$ be the expected value of the k^{th} moment of S, and let Δ_k be the probability that two individuals at any given locus share k pairs of IBD alleles, $k = 1, 2, \cdots, v$. If we denote $\mathbf{E}$ for the vector $[\mathrm{E}(S), \mathrm{E}(S^2), \cdots, \mathrm{E}(S^v)]^\mathrm{T}$, and $\mathbf{\Delta}$ for the vector $[\Delta_1, \Delta_2, \cdots, \Delta_v]^\mathrm{T}$, then the relation between $\mathbf{E}$ and $\mathbf{\Delta}$ can be expressed as

$$\mathbf{E} = \mathbf{A} + \mathbf{M}\mathbf{\Delta},$$

where $\mathbf{A}$ is a column vector containing v elements, and $\mathbf{M}$ is a square matrix with order v. Now, if we regard the elements in $\boldsymbol{\Delta}$ as unknowns, the above expression is a linear equation set, whose solution is $\widehat{\boldsymbol{\Delta}} = \mathbf{M}^{-1}(\hat{\mathbf{E}} - \mathbf{A})$. For example, in tetraploids, if $G_x = A_iA_iA_iA_i$, then

$$\mathbf{A} = \begin{bmatrix} 1 & 0.75 & 0.5 & 0.25 & 0 \\ 1 & 0.75^2 & 0.5^2 & 0.25^2 & 0 \\ 1 & 0.75^3 & 0.5^3 & 0.25^3 & 0 \\ 1 & 0.75^4 & 0.5^4 & 0.25^4 & 0 \end{bmatrix} \begin{bmatrix} p_i^4 \\ 4p_i^3p_x \\ 6p_i^2p_x^2 \\ 4p_ip_x^3 \\ p_x^4 \end{bmatrix},$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0.75 & 0.5 & 0.25 & 0 \\ 1 & 0.75^2 & 0.5^2 & 0.25^2 & 0 \\ 1 & 0.75^3 & 0.5^3 & 0.25^3 & 0 \\ 1 & 0.75^4 & 0.5^4 & 0.25^4 & 0 \end{bmatrix} \begin{bmatrix} p_i^3 - p_i^4 & p_i^2 - p_i^4 & p_i - p_i^4 & 1 - p_i^4 \\ (3p_i^2 - 4p_i^3)p_x & (2p_i - 4p_i^3)p_x & (1 - 4p_i^3)p_x & -4p_i^3p_x \\ (3p_i - 6p_i^2)p_x^2 & (1 - 6p_i^2)p_x^2 & -6p_i^2p_x^2 & -6p_i^2p_x^2 \\ (1 - 4p_i)p_x^3 & -4p_ip_x^3 & -4p_ip_x^3 & -4p_ip_x^3 \\ -p_x^4 & -p_x^4 & -p_x^4 & -p_x^4 \end{bmatrix},$$

and so the equation set $\mathbf{E} = \mathbf{A} + \mathbf{M}\boldsymbol{\Delta}$ becomes

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \\ \mathrm{E}(S^3) \\ \mathrm{E}(S^4) \end{bmatrix} = \begin{bmatrix} 1 & 0.75 & 0.5 & 0.25 & 0 \\ 1 & 0.75^2 & 0.5^2 & 0.25^2 & 0 \\ 1 & 0.75^3 & 0.5^3 & 0.25^3 & 0 \\ 1 & 0.75^4 & 0.5^4 & 0.25^4 & 0 \end{bmatrix} \left(\begin{bmatrix} p_i^4 \\ 4p_i^3p_x \\ 6p_i^2p_x^2 \\ 4p_ip_x^3 \\ p_x^4 \end{bmatrix} + \begin{bmatrix} p_i^3 - p_i^4 & p_i^2 - p_i^4 & p_i - p_i^4 & 1 - p_i^4 \\ (3p_i^2 - 4p_i^3)p_x & (2p_i - 4p_i^3)p_x & (1 - 4p_i^3)p_x & -4p_i^3p_x \\ (3p_i - 6p_i^2)p_x^2 & (1 - 6p_i^2)p_x^2 & -6p_i^2p_x^2 & -6p_i^2p_x^2 \\ (1 - 4p_i)p_x^3 & -4p_ip_x^3 & -4p_ip_x^3 & -4p_ip_x^3 \\ -p_x^4 & -p_x^4 & -p_x^4 & -p_x^4 \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_3 \\ \Delta_3 \\ \Delta_4 \end{bmatrix} \right).$$

The relatedness estimator of maximum-likelihood for polyploid is a modification of Milligan's (2003) estimator. In this estimator, all patterns of genotypic pairs between two individuals are enumerated. There are 9 patterns in diploids (see the expression of $\mathcal{L}(G_x, G_y \mid \phi, \Delta)$ mentioned above). There are 109, 1043, 8405 patterns in tetraploids, hexaploids and octoploids, respectively. For example, if $G_x = A_iA_iA_iA_i$ in tetraploids, then the expression of $\mathcal{L}(G_x, G_y \mid \phi, \Delta)$ for the first several patterns is as follows:

$$\mathcal{L}(G_x, G_y|\,\phi, \Delta) = \begin{cases} p_i^4\Delta_4 + p_i^5\Delta_3 + p_i^6\Delta_2 + p_i^7\Delta_1 + p_i^8\Delta_0 & \text{if } G_y = A_iA_iA_iA_i, \\ 4p_i^4p_j^4\Delta_0 & \text{if } G_y = A_jA_jA_jA_j, \\ p_i^4p_j\Delta_3 + 2p_i^5p_j\Delta_2 + 3p_i^6p_j\Delta_1 + 4p_i^7p_j\Delta_0 & \text{if } G_y = A_iA_iA_iA_j, \\ p_i^4p_j^3\Delta_1 + 4p_i^5p_j^3\Delta_0 & \text{if } G_y = A_iA_jA_jA_j, \\ 4p_i^6p_jp_k^3\Delta_0 & \text{if } G_y = A_jA_kA_kA_k, \\ p_i^4p_j^2\Delta_2 + 3p_i^5p_j^2\Delta_1 + 6p_i^6p_j^2\Delta_0 & \text{if } G_y = A_iA_iA_jA_j, \\ 6p_i^4p_j^2p_k^2\Delta_0 & \text{if } G_y = A_jA_jA_kA_k, \\ 2p_i^4p_jp_k\Delta_2 + 6p_i^5p_jp_k\Delta_1 + 12p_i^6p_jp_k\Delta_0 & \text{if } G_y = A_iA_iA_jA_k, \\ 3p_i^4p_j^2p_k\Delta_1 + 12p_i^5p_j^2p_k\Delta_0 & \text{if } G_y = A_iA_jA_jA_k, \\ 12p_i^4p_j^2p_kp_l\Delta_0 & \text{if } G_y = A_jA_jA_kA_l, \\ 6p_i^4p_jp_kp_l\Delta_1 + 24p_i^5p_jp_kp_l\Delta_0 & \text{if } G_y = A_iA_jA_kA_l, \\ 24p_i^4p_jp_kp_lp_m\Delta_0 & \text{if } G_y = A_jA_kA_lA_m, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{cases}$$

For different ploidy levels, VCFPOP 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 equal, then $\hat{r}_{HL}$ is denoted by $\hat{r}$.

The relatedness coefficients can also be converted from the kinship coefficient except the above methods. There are two methods to achieve this conversion. One method is the original conversion, whose formula is

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

where $v_{\min}$ is the ploidy level of the lower ploidy individual, and $\hat{\theta}$ is the kinship coefficient between these two individuals. It is worth noting that this conversion can only be used for outbred populations, and $\hat{r}_{HL}$ may be greater than one. The other method is provided by Huang *et al*. (2015a), whose converting 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 kinship coefficient between two individuals, $\hat{\theta}_{xx}$ and $\hat{\theta}_{yy}$ are the kinship coefficients within the individuals x and y, respectively, and $v_{\max}$ is the ploidy 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. Such method can be applied to Ritland's (1996) and Loiselle *et al.*'s (1995) estimators, but not to Weir's (1996) estimator because its $\hat{\theta}_{xx}$ may be zero or negative.

4.10 Principal coordinates analysis

Assume that $\mathbf{D}$ is the genetic distance matrix $[d_{ij}]$ of order n, and $\mathbf{A}$ is the matrix $[a_{ij}]$, in which $a_{ij} = -\frac{1}{2}d_{ij}^2$, $i, j = 1, 2, \cdots, n$. Then the Gower's (1966) centered matrix $\mathbf{G}$ can be calculated by centering the elements of $\mathbf{A}$, whose calculating formula is $\mathbf{G} = \mathbf{EAE}$, where $\mathbf{E}$ is a square matrix of order n, whose diagonal elements are all $1 - 1/n$, and other elements are all $-1/n$.

Because the distance matrix $\mathbf{D}$ is symmetric, so is $\mathbf{G}$. Therefore, there are exactly n real eigenvalues of $\mathbf{G}$, denoted by $\lambda_1, \lambda_2, \cdots, \lambda_n$, and there are n orthonormal eigenvectors of $\mathbf{G}$, denoted by $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_n$, where each eigenvector is a column vector with n components. Let $\mathbf{U}$ be the matrix $[\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_n]$. Then $\mathbf{U}$ is an orthogonal matrix. Assume that the corresponding eigenvalues of $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_n$ are $\lambda_1, \lambda_2, \cdots, \lambda_n$ in turn, then $\mathbf{G}$ can be decomposed as $\mathbf{G} = \mathbf{UVU}^{\mathrm{T}}$, where $\mathbf{V} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$.

It is noteworthy that each eigenvalue λ_i represents variance in the data that is projected into the eigenvector $\boldsymbol{\xi}_i$, and each variance is a non-negative number. Therefore, any negative eigenvalues should be excluded. It does not affect the hypothesis if the first m eigenvalues are whole non-negative eigenvalues of $\mathbf{G}$. If so, then the output matrix $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \left[\sqrt{\lambda_1}\boldsymbol{\xi}_1, \sqrt{\lambda_2}\boldsymbol{\xi}_2, \cdots, \sqrt{\lambda_m}\boldsymbol{\xi}_m\right].$$

Finally, if all eigenvalues of $\mathbf{G}$ are negative, an output matrix will not be generated.

4.11 Hierarchical clustering

The hierarchical clustering analysis is based on the distance matrix $\mathbf{D}$ of order n, where n is the number of clusters (assuming the identifiers of clusters are from 1 to n), and 'distance' has various indicators between two clusters, such as the genetic distance, phenotypic dissimilarity, geographical distance and so on. Initially, each individual, each population or each region is defined as a cluster, and the element d_{ij} in $\mathbf{D}$ is the distance between the clusters i and j. We will repeatedly update the matrix $\mathbf{D}$. The updated procedure is described as follows.

Assume that the minimum non-diagonal element in $\mathbf{D}$ is d_{ab} ($a \neq b$). Then, for the output dendrogram, the two nodes representing the clusters a and b are merged to the node with the coordinate $d_{ab}/2$. Second, update the elements in the a^{th} row and the a^{th} column of $\mathbf{D}$ 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 b^{th} row and the b^{th} column of $\mathbf{D}$, such that the order of $\mathbf{D}$ is reduced to $n - 1$. Such a procedure needs to be repeated $n - 1$ times, until the order of $\mathbf{D}$ is reduced to 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 below (N_a, N_b and N_c denote the numbers of the members of the clusters a, b and c in turn).

- 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

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

4.12 Bayesian clustering

MCMC algorithm. Pritchard *et al*. (2000) used a Markov Chain Monte-Carlo (MCMC) algorithm with Gibbs sampling to infer population structure. In the MCMC algorithm, various parameters are repeatedly updated according to the allele frequencies in both 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_{ila}, 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_{ila} will be explained in this section, and the definitions of the other parameters are shown in Section 4.12. The state will be updated by iterations. 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 Gibbs sampling. The Markov chain starts from a randomly generated initial state, and the state after several iterations will become stable and independent to the initial state. This iteration period is the 'burnin'. After that, this algorithm begins to record the number of times that an allele or an individual is assigned to each cluster. Such iterations are the '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' to eliminate any autocorrelation. Some independent runs (with different random number generator seeds) can be performed to repeat the results or to avoid the Markov chain being blocked at local maxima. The number of independent runs is `number of runs'.

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

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

for the non-F model; or from the Dirichlet distribution

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

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 cluster k, ε_{lj} is the frequency of A_{lj} in the ancestral cluster, and $f_k = (1 - F_k)/F_k$. The F model together with the value F_k of parameter F will be described in the penultimate paragraph of this section. The parameter λ can be used to prevent the allele frequencies to be 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=yes` is checked, then λ will be updated by the Metropolis-Hastings approach. The updated value λ' is randomly drawn from the normal distribution $\mathrm{N}(\lambda, \mathrm{std}^2(\lambda))$. If λ' is below zero or above $\max(\lambda)$, it is rejected directly. Otherwise, it is accepted at the probability of $\max(1, E)$, where E is the probability of accepting a worse value of λ', whose calculating formula is as follows:

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

for the non-F model, or

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

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

Moreover, if the option `Diff λ' is checked, this enables this 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 cluster k will be randomly drawn from the normal distribution $\mathrm{N}(\lambda_k, \mathrm{std}^2(\lambda))$, whose accepted probability E_k is as follows:

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

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 individual i. The updated cluster Z_i' of Z_i is randomly drawn from all K clusters, and the accepting probability $\Pr(Z_i' = k)$ that Z_i' is equal to the cluster k is

$$\Pr(Z_i' = k) = \frac{\prod_l^L \prod_a^{v_i} P_{kA_{la}}}{\sum_{k'}^K \prod_l^L \prod_a^{v_i} P_{k'A_{la}}}, \qquad k = 1, 2, \cdots, K,$$

where v_i is the ploidy level of the individual i, A_{la} is the a^{th} allele copy in the individual i at locus l, and $P_{kA_{la}}$ is the frequency of A_{la} in the cluster k. For the ADMIXTURE model, it is assumed that different alleles in the same individual can originate from different clusters. Let Z_{ila} be the current originating cluster of the allele copy A_{la}. Like the situation of non-admixture model, the updated cluster Z_{ila}' of Z_{ila} is randomly drawn from all K clusters. Using the elements in the vector **Q**, the accepting probability that Z_{ila}' is equal to the cluster k is

$$\Pr(Z_{ila}' = k) = \frac{Q_{ik}P_{kA_{la}}}{\sum_{k'}^K Q_{ik'}P_{k'A_{la}}},$$

where $l = 1, 2, \cdots, L$, $a = 1, 2, \cdots, v_i$, $k = 1, 2, \cdots, K$ and the meanings of $\mathbf{Q}$ together with its elements are explained in the next paragraph.

Vector Q of admixture proportions. Assume that Q_{ik} denotes the admixture proportion of genome of the individual i that originates from the cluster k ($k = 1, 2, \cdots, K$), and $\mathbf{Q}$ denotes the vector $[Q_{i1}, Q_{i2}, \cdots, Q_{iK}]$. As stated previously, $\mathbf{Q}$ is used in the ADMIXTURE model (Pritchard *et al.* 2000) to update the originating cluster Z_{ila}. The elements in $\mathbf{Q}$ 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 individual i that is assigned to the cluster k, and α_k is the value of Dirichlet parameter α for the cluster k ($k = 1, 2, \cdots, K$). These alphas can be used to prevent the proportions becoming fixed. The higher the values of these alphas, the higher the mixture level. Additionally, the elements in $\mathbf{Q}$ are also updated via a Metropolis-Hastings approach in the sense that 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 that is used for sampling the individuals from both the location s and the cluster k ($k = 1, 2, \cdots, K$). These updated values $Q'_{i1}, Q'_{i2}, \cdots, Q'_{iK}$ are accepted at the probability of min $(1, E)$, where

$$E = \prod_a^{v_i} \prod_l^{L} \frac{\sum_k^K Q'_{ik} P_{kA_{la}}}{\sum_k^K Q_{ik} P_{kA_{la}}}.$$

Such an update will improve the mixing when these alphas are small, and it will shuffle the individuals to prevent the Markov chain becoming blocked at local maxima.

Dirichlet parameter $\boldsymbol{\alpha}$ 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 in the non-LOCPRIORI model. For the non-LOCPRIORI model, if the values of α for the whole clusters are assumed to be all equal, the updated value α' is randomly drawn from the normal distribution $\mathrm{N}\big(\alpha, \mathrm{std}^2(\alpha)\big)$, and α' is accepted at the probability of $\min(1, E)$, where

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

in which N is the number of individuals, $\Pr(\alpha)$ is the priori probability of the parameter α, and $\frac{\Pr(\alpha')}{\Pr(\alpha)}$ 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 the gamma distribution $\Gamma(A, B)$. Moreover, if the values of α for the whole clusters are assumed to be not all equal, the updated value α_k' for the cluster k is randomly drawn from the normal distribution $\mathrm{N}\big(\alpha_k, \mathrm{std}^2(\alpha)\big)$, and α_k' is accepted at the probability of $\min(1, E_k)$, where

$$E_k = \frac{\Pr(\alpha_k')}{\Pr(\alpha_k)} \prod_i^N \frac{\Gamma\big(\alpha_k' - \alpha_k + \sum_{k'}^K \alpha_{k'}\big)\Gamma(\alpha_k)}{\Gamma\big(\sum_{k'}^K \alpha_{k'}\big)\Gamma(\alpha_k')} Q_{ik}^{\alpha_k'-\alpha_k}, \qquad k = 1, 2, \cdots, K,$$

in which 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 *a priori* information to assist 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 k^{th} element η_k in $\boldsymbol{\eta}$ is the priori probability of individuals assigned to the cluster k, and the k^{th} element γ_{sk} in $\boldsymbol{\gamma}_s$ is the priori probability of individuals sampled from the location s and assigned to the cluster k ($k = 1, 2, \cdots, K$). For the ADMIXTURE model, the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_{\text{local}}$ are used, which consist of the global alphas and the local alphas, respectively, where the k^{th} element α_k in $\boldsymbol{\alpha}$ (or $\alpha_{\text{local},sk}$ in $\boldsymbol{\alpha}_{\text{local}}$) reflects reflect the relative levels of admixture from each cluster over all individuals (individuals sampled from the location s), $k = 1, 2, \cdots, K$. Moreover, the parameter r

related to the informativeness of data parameterizes the extent to which the ancestral proportions at the locations of the sampled individuals can deviate from the overall proportion. If r is high ($\gg 1$), then the priori ancestry proportions across 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 different locations, implying that the location data are more informative about ancestry.

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

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

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

$$\eta'_a=\eta_a+d,$$

$$\eta'_b=\eta_b-d.$$

Now, if η'_a or η'_b is not in the acceptable range $[0,1]$, both are rejected, and the original η_a and η_b are regarded as the updated values; if ε'_{la} and ε'_{lb} are in the acceptable range $[0,1]$, both are accepted at the acceptance rate E, where

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

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

$$\gamma'_{sk}=\gamma_{sk}+d,$$

$$\gamma'_{sl} = \gamma_{sl} - d.$$

Like the situation of $\boldsymbol{\eta}$, γ'_{sk} and γ'_{sl} are checked within the range [0,1], and both are accepted at the acceptance rate E_s, where

$$E_s = \left(\frac{\gamma'_{sk}}{\gamma_{sk}}\right)^{r\eta_k - 1 - N_{sk}} \left(\frac{\gamma'_{sl}}{\gamma_{sl}}\right)^{r\eta_l - 1 - N_{sl}}, \qquad s = 1, 2, \cdots, S,$$

in which N_{sk} (or N_{sl}) is the number of the individuals sampled from the location s and assigned to the cluster k (or l).

Update of LOCPRIORI model + ADMIXTURE model. For these two models, the new value r' of the parameter r is randomly drawn from the same uniform distribution as described in the preceding paragraph, whose acceptance rate E is

$$E = \prod_k^K \prod_s^S \left[\frac{r'^{r'\alpha_k}}{r^{r\alpha_k}} \frac{\Gamma(r\alpha_k)}{\Gamma(r'\alpha_k)} \alpha_{\mathrm{local},sk}^{\alpha_k d} \exp\left(-d\alpha_{\mathrm{local},sk}\right) \right],$$

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

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

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

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

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

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

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

where $f_k = (1 - F_k)/F_k$, and $\varepsilon_{l1},\ \varepsilon_{l2}, \dots,\ \varepsilon_{lJ_l}$ are the frequencies of alleles in the ancestor cluster at locus l. The values of F are also updated. If different value F in each cluster is used, the updated value F'_k of F_k is randomly drawn from the normal distribution $\mathrm{N}\big(F_k, \mathrm{std}^2(F)\big)$, which is accepted at the probability of $\min(1, E_k)$, where

$$E_k = \left(\frac{F'_k}{F_k}\right)^{\mu^2/\sigma^2 - 1} \exp\left(\frac{\mu(F_k - F'_k)}{\sigma^2}\right) \prod_l^L \frac{\Gamma(f'_k) \prod_j^{J_l} \Gamma\big(f_k \varepsilon_{lj}\big) P_{klj}^{f'_k \varepsilon_{lj}}}{\Gamma(f_k) \prod_j^{J_l} \Gamma\big(f'_k \varepsilon_{lj}\big) P_{klj}^{f_k \varepsilon_{lj}}}, \qquad k = 1, 2, \cdots, K,$$

in which μ and σ are respectively the priori mean and the priori standard deviation of F, P_{klj} is an element in $\mathbf{P}$, and $f'_k = (1 - F'_k)/F'_k$. Moreover, if the values of F for the all clusters are assumed to be all equal, the updated value F' of F will be randomly drawn from the normal distribution $\mathrm{N}\big(F, \mathrm{std}^2(F)\big)$, whose acceptance rate E is

$$E = \left(\frac{F'}{F}\right)^{\mu^2/\sigma^2 - 1} \exp\left(\frac{\mu(F - F')}{\sigma^2}\right) \prod_k^K \prod_l^L \frac{\Gamma(f') \prod_j^{J_l} \Gamma\big(f \varepsilon_{lj}\big) P_{klj}^{f' \varepsilon_{lj}}}{\Gamma(f) \prod_j^{J_l} \Gamma\big(f' \varepsilon_{lj}\big) P_{klj}^{f \varepsilon_{lj}}}.$$

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 drawn from the Dirichlet distribution

$$\mathcal{D}\left(\lambda + \sum\nolimits_k^K P_{kl1} f_k, \lambda + \sum\nolimits_k^K P_{kl2} f_k, \cdots, \lambda + \sum\nolimits_k^K P_{klJ_l} f_k\right), \qquad l = 1, 2, \cdots, L,$$

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

$$E_{lj'} = \prod_j^{J_l} \left[\left(\frac{\varepsilon_{lj}}{\varepsilon'_{lj}}\right)^{\sum_k^K P_{klj} f_k} \prod_k^K \frac{\Gamma\big(f_k \varepsilon_{lj}\big)}{\Gamma\big(f_k \varepsilon'_{lj}\big)} P_{klj}^{f(\varepsilon'_{lj} - \varepsilon_{lj})}\right], \qquad l = 1, 2, \cdots, L, j' = 1, 2, \cdots, J_l.$$

For the second approach, this only needs to be updated to the frequencies of 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 $\mathrm{U}(0, N^{-1/2})$, 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 $[0,1]$, 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 $[0,1]$, both are accepted at the acceptance rate E, where

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

5 Update history

2021/4/12 V1.02

* Support C++20 standard.
* Optimize calculation speed for genetic distance.
* Use `shared_mutex` class to realize read/write lock.
* Update linear algebra library to Eigen 3.4.0 and Spectra 1.0.1.

2021/4/12 V1.01

* Fix a bug in calculating genetic distance using multiple threads.
+ Replace genetic distance with average value for pairs with missing data.

2021/3/31 V1.0

* Reduce memory expense for locus and genotype

* Optimize sequential genotype index accession speed

* Reduce temporary file use

+ Add collapse allele function in testing genotype distribution

+ Add variable length arrays to optimize speed

+ Add `-haplotype_ptype` option

- Remove `-g_buffer` and `-g_memory` option and reduce memory during loading

2020/9/22 V0.9

* Fix a bug in calculating hash values

* Optimize AMOVA speed

* Optimize memory expense

* Optimize loading and calculation speed

2020/5/21 V0.8

* Optimize AMOVA code

+ Add multi-level region definition

+ Modify source code to support LLVM/Clang

+ Add Huang (2021) Fst estimators

2019/12/8 V0.7

* Optimize calculating speed

* Optimize memory expense

* Fix bugs

* Optimize Bayesian clustering code

* Optimize calculating speed

+ Add support for AVX-512 instructions

2018/12/8 V0.6

+ Add support for BCF format and .vcf.gz compression

+ Add support for other genotype formats

+ Add file format converter for other genotype formats

2018/9/6 V0.5

* Optimize loading speed

+ Add haplotype extraction function

+ Add population differentiation estimator

+ Add PCoA function

+ Add hierarchical clustering function

2018/6/19 V0.4

+ Add support for different ploidy level

+ Add Bayesian clustering function

2018/4/9 V0.3

+ Add relatedness coefficient estimators

+ Add AMOVA function

2018/2/19 V0.2

+ Add genotype distribution test

+ Add population differentiation estimator

+ Add genetic differentiation tests

+ Add individual H-index estimator

+ Add population assignment

+ Add genetic distance estimators

2018/2/6 V0.1

+ Loading from VCF files

+ Add population genetic diversity estimators

+ Add double-reduction model: RCS, PRCS, CES, PES

+ Add file format converter for GENEPOP and STRUCTURE

6 References

Anderson AD, Weir BS (2007) A maximum likelihood method for estimation of pairwise relatedness in structured populations. *Genetics*.

Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis: Models and estimation procedures. *Evolution* **21**, 550-570.

Danecek P, Auton A, Abecasis G, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

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.

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.

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325-338.

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**, 618-620.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* **59**, 1633-1638.

Huang K, Dunn DW, Ritland K, Li B (2020) polygene: Population genetics analyses for autopolyploids based on allelic phenotypes. *Methods in Ecology and Evolution* **11**, 448-456.

Huang K, Guo ST, Shattuck MR, *et al.* (2015a) A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity* **114**, 133-142.

Huang K, Ritland K, Dunn DW, *et al.* (2016) Estimating relatedness in the presence of null alleles. *Genetics* **202**, 247-260.

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**.

Huang K, Wang T, Dunn DW, *et al.* (2021) A generalized framework for AMOVA with multiple hierarchies and ploidies. *Integrative Zoology* **16**, 33-52.

Huang K, Wang TC, Dunn DW, *et al.* (2019) Genotypic frequencies at equilibrium for polysomic inheritance under double-reduction. *G3: Genes, Genomes, Genetics* **9**, 1693-1706.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**, 1322-1332.

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

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

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* **16**, 1099-1106.

Kimura M (1954) Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. *Genetics* **39**, 280.

Li C, Weeks D, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. *Human heredity* **43**, 45-52.

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.

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

Milligan BG (2003) Maximum-likelihood estimation of relatedness. *Genetics* **163**, 1153-1167.

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.

Paetkau D, Slade R, Burden M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Molecular Ecology* **13**, 55-65.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. *Evolution* **43**, 258-275.

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**, 767-779.

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research* **67**, 175-185.

Rogers JS (1972) Measures of similarity and genetic distance. In: *Studies in Genetics VII* (ed. Wheeler MR), pp. 145-153. University of Texas Publication, Austin.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

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

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 (2010) A simplified estimator of two and four gene relationship coefficients. *Molecular Ecology Resources* **10**, 986-994.

Wang J (2002) An estimator for pairwise relatedness using molecular markers. *Genetics* **160**, 1203-1215.

Weir BS (1996) *Genetic data analysis II: methods for discrete population genetic data* Sinauer Associates, Sunderland.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.