PolyRelatedness V1.8 User Manual

This software is designed to estimate relatedness among either polyploids, aneuploids, or between different levels of ploidy. It and can also use genetic markers with null alleles.

Table of contents

Developed by Huang Kang
Lecturer, PhD of Zoology
College of Life Sciences, Northwest University
No. 299, Taibai North Avenue
Xi'an, Shaanxi Province, China
Zipcode: 710069
E-mail: huangkang@nwu.edu.cn
Comments and suggestions are welcome.

System Requirements

- CPU: x64 compatible
- OS: Microsoft Windows, Ubuntu, or Mac OS X
- Memory: 128 Mib
- Hard drive: 100 Mib

Limitations

- Maximum number of loci: 65536
- Maximum number of alleles: 65536
- Maximum number of individuals: 65536
- Maximum number of populations: 65536

Citation

Polyploid MOM estimator:

Huang K, Ritland K, Guo ST, et al. (2014) A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources*, 14, 734-744.

Polyploid Maximum-likelihood estimator:

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

Estimating relatedness in different levels of ploidy:

Huang K, Ritland K, Guo ST, et al. (2015) Estimating pairwise relatedness between individuals with different levels of ploidy. *Molecular Ecology Resources*, 15, 772-784.

Null alleles adjustment for diploids:

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

Download, setup & uninstall of the program

The software can be downloaded from the following URL:

https://github.com/huangkang1987/polyrelatedness

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

To uninstall, just delete this folder.

Functions with the program

Polyploid estimators

This software implements two relatedness estimators for polyploids:

1. Huang et al. 2014 MOM estimator (ploidy <= 8);
2. Huang et al. 2015 ML estimator (ploidy <= 8);

The former is a method-of-moment estimator, and the latter is a maximum-likelihood estimator. These two estimators support any levels of ploidy, even between different level of ploidy. But the latter do not support aneuploid (different chromosomes have different numbers of copies).

Methods-of-moment (MOM) estimators equate sample moments for the unknown population moment to expected moments under a genetic model. The MOM method is able to generate unbiased estimations of relatedness coefficients. However, MOM estimators can produce values outside of this range (i.e. negative or larger than one). One way to address this is to truncate the estimators, forcing the estimate to lie within the expected range, although this produces bias.

The Huang et al. 2014 MOM estimator encounters the ill-conditioned matrix problem due to the lack of genetic information in specific dyads at some loci, this problem will introduce an extra bias. But the bias can be eliminated by increase the number of loci, beyond about 60.

Maximum-likelihood (ML) models the probability of observing the pairwise allelic pattern, given the prior of 'higher order' coefficients and the allele frequencies. By searching the parameter space for values of 'higher order' coefficients that maximize the probability of observed genotypes, ML estimates can be obtained. Because maximization can be limited to the parameter space as defined by probabilities of IBD, invalid values for the parameters are avoided.

For ML estimation, there are another three coefficients of coancestry estimators, they can also estimate the relatedness between different level of ploidy and aneuploid.

3. Ritland 1996 estimator (corrected by Eqn 8 in Huang et al. 2015 Heredity, ploidy <= 8);
4. Loiselle et al. 1995 estimator (corrected by Eqn 8 in Huang et al. 2015 Heredity, ploidy <= 8);
5. Ritland 1996 estimator (ploidy <= 8);
6. Loiselle et al. 1995 estimator (ploidy <= 8)
7. Weir 1996 estimator (ploidy <= 8);

The first two estimators are corrected by Eqn 8 in Huang et al. 2015 Heredity to convert the coefficient of ancestry to relatedness coefficient, and the next three estimators are multiplied by ploidy level to convert to relatedness coefficient. The latter method may lead to relatedness estimates above one.

The coefficient of ancestry (a.k.a. kinship coefficient) is the probability that two alleles, one randomly drawn from each individual, are identical-by-descent. If the ploidy of the two individuals are equal, the relatedness coefficient is the coefficient of coancestry multiply by ploidy. Note that, this software output relatedness coefficient.

In polyploids, allele dosage of heterozygotes cannot be determined as the PCR method is insensitive to copy number “ambiguous genotype” (this is in contrast to isozymes, where allele copy number can be determined; earlier methods assumed that allele copy number can be determined by inspection of gels, hence new methods are needed in the age of PCR). All these 5 estimators can handle ambiguous genotypes. In general, ambiguous genotypes cause a negative bias for relatives, because the true genotype-pair (which has a high relatedness estimate) is

“diluted“ by many less related genotypes.

Diploid estimators used in the program

This software also imbeds many estimators for diploids, including:

1. Lynch & Ritland 1999 *;
2. Wang 2002 *;
3. Thomas 2010 *;
4. Thomas 2010 weighted by Wang 2002 *;
5. Li et al. 1993;
6. Queller & Goodnight 1989;
7. Huang 2016 A for diploid *;
8. Huang 2016 B for diploid *;
9. Anderson & Weir 2007 *;
10. Milligan 2003 *;
11. Milligan 2003 for inbreeding *.

The first eight estimators are based upon method-of-moments (MOM), while the last three are based upon maximum-likelihood (ML).

The asterisks denote estimators corrected for null alleles. If null alleles are present, the relatedness can be either overestimated or underestimated. Null alleles can bias estimation in two ways: (i) the observed homozygote may be a heterozygote carrying a null allele; (ii) the frequency of observed alleles is overestimated due to null alleles. The correction for MOM estimators cannot eliminate bias, but the bias is reduced to an acceptable level: the bias is less than 0.05 if the frequency of null allele is less than 0.5. The potency for a locus with a frequency of null alleles larger than 0.5 is about 1/10 that of locus without null alleles, so is suggested leave out loci with null alleles of frequency of 0.5 or above.

Simulation functions of the program

The simulation function is able to measure the performance of estimators; this function can help researchers to choose the best estimators under specific conditions. The information of loci is defined the input file including: the number of loci, the number of alleles of each locus, the

frequency of each allele, the typed rated of each locus, the simulation types are listed as follows:

```
1.  Diploid;
2.  Tetraploid;
3.  Tetraploid with ambiguous genotype;
4.  Hexaploid;
5.  Hexaploid with ambiguous genotype;
6.  Octoploid;
7.  Octoploid with ambiguous genotype;
8.  Haplo-diploid;
9.  Diplo-tetraploid;
10. Diplo-tetraploid with ambiguous genotype;
11. Triplo-hexaploid;
12. Triplo-hexaploid with ambiguous genotype;
13. Tetraplo-octoploid;
14. Tetraplo-octoploid with ambiguous genotype;
```

Each estimator has two modes: (1) estimate relatedness between any two individuals, or (2) estimate average relatedness between individuals within a population. Function (8) above is also able to switch these two modes. The input and output file are set to 'in.txt' and 'out.txt' as default, they can be changed by function (9) above.

Null alleles accommodated by the program

This software also has 7 estimators for estimating frequency of null alleles:

```
1.  Chakraborty et al. (1992, MOM);
2.  Summers & Amos (1997, MOM);
3.  Kalinowski & Taper (2006, *EM);
4.  Chybicki & Burczyk (2008, #EM);
5.  van Oosterhout et al. (2006, *#EM);
6.  van Oosterhout et al. (2004, MOM);
* estimators consider PCR fail (beta)
# for consider inbreeding (f).
```

The first two estimators and van Oosterhout et al (2004) estimators are MOM estimators, they assuming no inbreeding and no PCR failures are only because of null alleles, they are both moment estimators. Kalinowski & Taper (2006) estimator consider the probability of amplification failures, and apply an expectation maximization (EM) algorithm to maximum the likelihood. Chybicki & Burczyk (2006) estimator models the effect of inbreeding without the negative amplifications.

Although we can model both inbreeding and PCR failures together, and write the likelihood function, there are multiple solutions for f (inbreeding coefficient) and β (negative amplifications due to other reasons) that maximize the likelihood. The f and β cannot be solved simultaneously, I must determine either of them first. So van Oosterhout et al. (2006) estimator require you to input the f for each population.

Usage of the program

Launching

- Windows:

double click the `PolyRelatedness.exe`

- Mac OS X:

double click the `PolyRelatedness`

- Ubuntu:

Press Ctrl+Alt+T to open the console and drag the PolyRelatedness.out into the window, then press Enter. Or use the command

```
$ cd folder
```

to open the folder where the software located. Then use the following command to launch the program:

```
$ ./PolyRelatedness.out
```

Input file for the program

The input genotypic data are saved in 'in.txt', where the encoding of characters should be ASCII-compatible (i.e. UTF-8, GBK, BIG5, Shift-JIS, KSC, etc), do not use UTF-16 or UTF-32 format. The encoding for output file is same to the input file, so double byte chatacters like Chinese/Japanese/Korean are supported. In the input file, you can configure the format of genotypes. After the software is launched, select the function you desire. The results will be saved in 'out.txt' after the program is finished.

The configuration section of input file is as follow:

```
1  //configuration
2  //#alleledigits(1~4)     #outputdigits(0~10)      #missingallele
         #ambiguousallele   #nthreads(1~64)
3  3     8     0     999    4
```

The main component of this configuration file is line 3. In this example "3" means that there are 3 digits per genotype, the "8" means that eight digits are produced in the output, the "0" means that zero is the missing data value, the "999" is for polyploidy data to indicate ambiguous genotypes (see **Unknown balance of heterozygotes in polyploids below**), and the "4" is the number of threads in the computer processer (specifying this option allows faster computation).

Missing alleles and ambiguous allele labels should be zero or positive integer, and be different from each other, and should not be appeared in the allele frequency section. Columns are separated by tabs, you can edit the input file with a spreadsheet software like Microsoft Excel, and paste it into the text editor.

Allele frequency in the reference population

This part of the program is optional. If not chosen, allele frequency will be estimated from genotypes assuming no missing alleles by the EM algorithm.

```
4  //allele frequency
5  D6S4935     D7S8204
6  212   0.19491525   208   0.44915254
7  216   0.28813559   204   0.16949153
8  200   0.16101695   200   0.24576271
9  220   0.26271186   212   0.13559322
10 224   0.09322035
```

Each locus requires two columns, one for allele identity; another for allele frequency. The first row is the locus identity and number of alleles. The software will divided the frequency of each allele by their sum to make sure that the sum of allele frequencies at a locus is strictly one. But if the diploid estimator corrected for null alleles is used, the sum of allele frequencies is allowed to be

less than one. The frequency of null alleles is defined by one minus the sum of allele frequencies. However, this software cannot estimate the frequency of null alleles currently. The Queller & Goodnight (1989) estimator is invalid for diallelic loci, and diallelic loci with null alleles may introduce bias in diploid estimators.

The locus weights provide a method for a secondary weighting for locus which can be used to reduce the contribution of some linked loci to the final estimate. For example, if the one locus is linked with n loci, its secondary weight can be $(n + 1)^{-1}$. It is the lower bound of weight, based on the assumption that if two identical loci are used; the genotypes at these two loci of each individuals are equal, and these two loci should be treated as one.

For MOM estimators, there is a locus specific weight, and it is multiplied by the secondary weight. For ML estimators, the overall likelihood is the product of the likelihoods for each locus. In calculation, the summation of the logarithm of the likelihood of each locus is used, and it is also multiplied by the secondary weight. This section is also **optional**.

```
11 //locus weight
12 1     1
```

The genotype section is as follow:

```
13 //genotype
14 Ind   Pop   D6S493 D7S820
15 BBAF1 BB    212212212212 208212208212
16 BBAF2 BB    212220212220 200200200200
17 BBAF4 BB    212216212216 208208208208
18 //end of file
```

In the genotype section, the header row defines the names of loci from the third column; the order must be same as the previous section. The first column is the individual identity, followed by a column define the population of the individual. From the third column, there are number of columns defines the genotypes, the order of locus should meet that defined in the allele frequency

section.

The procedure of handing missing alleles is simple: if a genotype contains at least one missing allele, it is called a ”half-genotype”. In estimating pairwise relatedness, if either or both genotypes is half-genotype, the estimators will not calculated the relatedness for this locus. The weight of the loci is set to zero. The final relatedness is a weighted average across loci, and if no locus gives a valid estimates, the final estimate of relatedness is zero (sum of estimated relatedness across loci) divided by zero (sum of weights).

Unknown balance of heterozygotes in polyploids

The procedure of handling unknown allele dosage in heterozygous genotypes (“ambiguous genotype”) is as follows. The ambiguous allele code is used to tell the program what are the possible underlying genotypes as follows. If the tetraploid genotype is 12XX (only bands 1 and two observed on the gel, and we use here used one digits per allele), there are three possible genotypes: 1112, 1122 and 1222. The data can be encoded as 1299 when “9” is specified as the ambiguous genotype code. Also the order does not matter, for example: 1299 is the same as 1129 and 1229. Finally, the order of allele can be any one that you like, i.e. 9129 or 9921 are both the same.

Note that, only Huang et al. MOM, Huang et al. ML estimators, Ritland 1996, Loiselle 1995, and Weir 1996 estimators are able to handle ambiguous genotypes.

Output file

The output is an n-by-n matrix, where the column header and row header list the identity of individual, as the order in the input file. If you choose to calculated relatedness with in populations, where will be many matrices in the output file. You can open the 'out.txt' with an spread sheet

software.

```
Relatedness coefficient calculated by PolyRelatedness V1.6
    Estimator: Huang et al. 2015 ML
    Type: between all individuals
    Input: example1.txt
    Output: eout1.txt
    Time: 2015-5-12 02:01:57
      BBAF1 BBAF2 BBAF4
BBAF1 1.00000000  0.17618605  0.50000000
BBAF2 0.17618605  1.00000000  0.00000000
BBAF4 0.50000000  0.00000000  1.00000000
```

Command mode to run program

The software can be launched by command mode by adding parameters after the executable file name. You can use shell commands to run this software, or use another program to call it automatelly. For example:

```
PolyRelatedness in.txt out.txt i
PolyRelatedness in.txt out.txt e 1 0
PolyRelatedness in.txt out.txt s 1 100000
PolyRelatedness in.txt out.txt n 5 0.1 0.05 0.2
```

The first two parameters are the input and output files, if the file name (or full path) contains space, you should use double quote to embrace it. Note that the working directory is the directory the executable file located, so that the file name should be relative to this directory, or use absolute path.

The third parameter is the type of functions, "`i`" for estimating the individual inbreeding coefficient, "`e`" for estimating the relatedness, "`s`" for simulation, and "`n`" for estimating the frequency of null alleles.

In relatedness estimation, the fourth and fifth parameters are the estimator identifier and the calculation mode (0 for between all individuals, 1 for within population), the estimator identifiers

are as follow:

```
1      Huang et al. 2014 MOM estimator;
2      Huang et al. 2015 ML estimator;
3      Ritland 1996 estimator (corrected by Eqn 8 in Huang et al. 2015
Heredity, ploidy <= 8);
4      Loiselle et al. 1995 estimator (corrected by Eqn 8 in Huang et al.
2015 Heredity, ploidy <= 8);
5      Ritland 1996 estimator;
6      Loiselle et al. 1995 estimator;
7      Weir 1996 estimator;
8      Lynch & Ritland 1999;
9      Wang 2002;
10     Thomas 2010;
11     Thomas 2010 weighted by Wang 2002;
12     Li et al. 1993;
13     Queller & Goodnight 1989;
14     Huang 2016 A for diploid;
15     Huang 2016 B for diploid;
16     Anderson & Weir 2007;
17     Milligan 2003;
18     Milligan 2003 for inbreeding;
```

In simulation, the fourth and fifth parameters are the simulaion type and replications, the simulaion types are as follow:

```
1      Diploid with null alleles;
2      Tetraploid;
3      Tetraploid with ambiguous genotype;
4      Hexaploid;
5      Hexaploid with ambiguous genotype;
6      Octoploid;
7      Octoploid with ambiguous genotype;
8      Haplo-diploid;
9      Diplo-tetraploid;
10     Diplo-tetraploid with ambiguous genotype;
11     Triplo-hexaploid;
12     Triplo-hexaploid with ambiguous genotype;
13     Tetraplo-octoploid;
14     Tetraplo-octoploid with ambiguous genotype.
```

In null allele estimation, the fourth and following parameters are the estimator and inbreeding coefficient for each populations for van Oosterhout et al. (2006) estimator. For other estimators, the inbreeding coefficient can be omitted.

```
1      Chakraborty et al. (1992, MOM);
2      Summers & Amos (1997, MOM);
3      Kalinowski & Taper (2006, *EM);
```

```
4      Chybicki & Burczyk (2008, #EM);
5      van Oosterhout et al. (2006, *#EM);
6      van Oosterhout et al. (2004, MOM);
```

Updates

V1.8 2018/10/31

* Upgrade to 64 bit executable to allow to allocate more memory for large dataset.
* Accelerate the loading speed.
* Reduce the memory expense for ML estimator.
* Reset MOM estimator to its original version, which do not check the condition number of matrix M.

V1.7 2018/9/28

* Fix a bug in polyploid moment estimator to solve ambiguous genotype.
* Fix a stack bug crash the software after multiple runs.
* Optimize polyploid likelihood estimator for ridge likelihood surface.
+ Distinguish two kinds of corrections for Loiselle (1995) and Ritland (1996) estimator.

V1.6 2016/3/16

* Fix a bug in Lynch & Ritland (1999) estimator 's correction for null alleles.
* Fix a bug that prevent read genotypes with huge number of loci.
+ Estimate the individual inbreeding coefficient.
+ Output relatedness distribution (9 percentiles) in simulation function.

V1.5 2014/9/7

* Fix a bug that cannot estimate the allele frequency correctly.

V1.4 2014/6/1

* Revise the Ritland (1996) estimator.
* Fix a bug that cannot estimate the relatedness between different levels of ploidy.
* Modify the input file, the allele frequency and locus weight sections are optional.
* The allele frequency can be estimated from the genotypes without missing alleles, support different ploidy and ambiguous alleles. But the null alleles is not considered in the polyploid estimators, so I do not consider null alleles.

V1.3 2014/3/18

* Accelerate maximum-likelihood estimators.
* Update input file format, add secondary locus weight.
* Update maximum-likelihood searching algorithm.
* Fix maximum-likelihood estimator, they can get accurate results in Ubuntu now.
* Update error message prompt, and check whether alleles are inside the allele frequency section.
* Return to main menu after one calculation is done.
+ Add estimating relatedness with null alleles for diploid estimators.
+ Add null allele frequency estimators.
+ Add Genepop V4 format converter.
+ Using Microsoft word to write the manual, and add methodology section.

V1.2 2013/11/15

* Modify the format of input file.
* Moidfy the format of output file, add input file, output file, functions, and time.
* Switch compiler to GCC, and accelerate the program.
* Fix a bug that cannot recognize UTF-8 encoding text file.
* Fix a bug that misassign alleles for some IBS modes in Maximum-likelihood estimator.
+ Extend Huang et al. MOM estimator, Ritland 1996, Loiselle 1996 and Weir 1996 estimators for different levels of ploidy and aneuploid.
+ Extend Huang et al. ML estimator for higher and different levels of ploidy.
+ Modify all polyploid estimators to support ambiguous genotypes.
+ Add Weir 1996 estimator.
+ Add diploid estimators.
+ Add simulation function.
+ Add command line mode.
+ Add support for multi-threads acceleration.
+ Support for MAC OS X.

V1.1 2013/8/23

* Change the name of the software from TetraRelatedness to PolyRelatedness.
* Modify the format of input file to support missing allele and ambiguous allele.
* Update the Huang et al. MOM estimator, the valid range of output relatedness at each locus is changed to (-16, 1.001), where the 1.001 is to endure the error of floating-point number (IEEE-754).
+ Add support for missing allele.
+ Extend the Huang et al. MOM estimator to diploid, hexaploid and octoploid.

+ Add a maximum-likelihood estimator: Huang et al. ML estimator.
+ Add a coefficient of coancestry estimator: Ritland 1996.
+ Add a coefficient of coancestry estimator: Loiselle et al. 1995.
+ Support for Ubuntu.

V1.0 2013/6/22

Basic method-of-moment function, support for autotetraploids.

Reference

Anderson AD, Weir BS (2007) A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics, 176, 421-440.

Brookfield JFY (1996) A simple new method for estimating null allele frequency from heterozygote deficiency. Molecular Ecology, 5: 453-455.

Chakraborty R, Andrade M, Daiger SP, Budowle B (1992) Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Annals of human genetics, 56: 45-57.

Chybicki IJ, Burczyk J (2009) Simultaneous estimation of null alleles and inbreeding coefficients. Journal of Heredity, 100: 106-113.

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.

Li CC, Weeks DE, 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 Botany, 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.

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

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

Summers K, Amos W (1997) Behavioral, ecological, and molecular genetic analyses of reproductive strategies in the Amazonian dart-poison frog, *Dendrobates ventrimaculatus*. Behavioral Ecology, 8: 260-267.

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

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes, 4: 535-538.

van Oosterhout C, Weetman D, Hutchinson WF (2006) Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. Molecular Ecology Notes, 6: 255-256.

Wang JL (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. Sunderland: Sinauer Associates.