

IDENTITY 1.0

Freeware program for the analysis of microsatellite data

Horst W. Wagner, Institute for General Physics, TU Vienna
and Kristina M. Sefc, Centre for Applied Genetics, BOKU Vienna

December 1999

Updated to identity4.exe; most important change: more data allowed
Otherwise, instructions for version 1.0 apply.

Overview

IDENTITY 1.0 is a command line based tool, which features a number of mainly statistical routines for the analysis of microsatellite data. The program was written for the analysis of microsatellite data obtained during a study of grapevine cultivars (see Sefc et al. 1997, 1998a, b, c, 1999a, b; Lopes et al. 1999; Maletic et al. 1999). In this context, emphasis was put on the evaluation of marker polymorphism, screens for identical genotypes (synonyms) and pedigree reconstruction.

After requests by colleagues, the program is now made available via the Internet. The authors do not accept responsibility for the accuracy of the results produced by this program. Users of IDENTITY are encouraged to cite this program in publications in which results of this program are employed (e.g.: Wagner H. W. and Sefc K. M., 1999. IDENTITY 1.0. Centre for Applied Genetics, University of Agricultural Sciences Vienna).

Features

IDENTITY 1.0 provides following functionality:

- Locus by locus calculation of
 - Number of alleles
 - Allele frequencies
 - Expected and observed heterozygosity
 - Paternity exclusion probability
 - Probability of identity
 - Frequency of null alleles
- Detection of identical genotypes
- Detection of possible parent-offspring combinations
- Likelihood ratio statistics for the detected, putative parent-offspring groups
- Conversion of data into MICROSAT (Minch et al. 1997) input file format

Obtaining IDENTITY

IDENTITY is free and can be obtained on the Internet from the World Wide Web site at <http://www-ang.kfunigraz.ac.at/~sefck>

Please email kristina.sefc@uni-graz.at if you have questions, or if the program refuses to run.

Input file format

Identity employs comma delimited files as input file format. Files in this format can conveniently be created by saving data from your favourite spreadsheet program (e.g. MS Excel) in .csv format. (Note: some country settings use ";" instead of "," in csv files. In this case, either change the setting – "USA" is OK – or substitute all ";" with ",".)

The first two columns of the input file contain sample identifier, followed by the microsatellite data in base-pairs (two columns per locus).

The first column must contain numbers. The second column may contain either numbers or strings or combinations of both.

Example for input file format in Microsoft Excel:

No	Cultivar	Locus1	Locus1	Locus2	Locus2
1	Grüner Veltliner	228	230	134	154
2	Cultivar 101	224	224	122	137
2	144			138	138

Saved in the .csv format, this looks like:

```
No,Cultivar,Locus1,Locus1,Locus2,Locus2
1,Grüner Veltliner,228,230,134,154
2,Cultivar 101,224,224,122,137
2,144,,138,138
```

Running the program

Make sure that the program file (identity.exe) and your input file are in the same folder. The program is started from a DOS window by selecting the correct folder and typing

```
>identity filename.csv
```

Output files

IDENTITY generates following output files:

Filename.stat : contains lists of loci and samples in the input file. Samples with missing data are indicated by an asterisk (*) appended to the end of the locus name.

For each locus, the following information is provided:

- Number of alleles
- Allele frequencies
- Standard deviation of allele frequencies
- Upper 95% confidence interval limit of allele frequencies (Sachs 1997; for use in the likelihood ratio statistic – see below)
- Sum of detected alleles (2x sample number, if no missing data)
- Expected heterozygosity / Gene diversity (Nei 1973)
- Observed heterozygosity (from direct counts)
- Frequency of null alleles (Brookfield 1996)
- Probability of identity (Paetkau et al. 1995)
- Paternity exclusion probability (Weir 1996)

Probability of identity and Paternity exclusion probability, combined over all loci (by multiplication).

Filename.syn: gives a list of samples with identical genotypes. Numbers in parentheses are those in the first column of your input file. Again clones with missing data are indicated by an '*’.

Filename.par: gives a list of possible parent-offspring combinations, assuming codominant Mendelian inheritance of alleles, in the form "offspring = parent 1 x parent 2", with numbers from the first input file column in parentheses. Samples with missing data are denoted by *. Including samples with identical genotypes leads to redundancy in this file.

Filename.lik: likelihood ratio statistics corresponding to the detected possible parent-offspring combinations (as described by Bowers and Meredith 1997).

The following likelihood ratios are provided (per locus and combined over loci):

1. The ratio of the probability that the proposed parents gave rise to the offspring's genotype versus the probability that two random individuals give rise to the offspring's genotype.
(Proposed parents) versus (two random cultivars)
= X x Y in output file
2. Likelihood ratio for: (Proposed parents) versus (random individual x proposed parent 1)
= X x (1)
3. (Proposed parents) versus (close relative of proposed parent 2 x proposed parent 1)
= rel(2) x (1)

4. (Proposed parents) versus (Proposed parent 2 x random cultivar)
= (2) x X
5. (Proposed parents) versus (Proposed parent 2 x close relative of proposed parent 1)
= (2) x rel(1)

Probabilities for the proposed parents are derived from Mendel's laws, while probabilities for random individuals are calculated from allele frequencies in the population.

Inheritance of rare alleles strongly strengthens conclusions on parentage. However, in small samples, certain alleles may be underrepresented and calculated allele frequencies may thus be smaller than actual population allele frequencies. In order to compensate for possible errors, the likelihood ratios are also calculated from the 95% upper confidence limits of the observed allele frequencies.

Example for output Filename.lik:

```

Offspring = Parent 1 x Parent 2  (a)
-----
(b) Locus1 (134 142) = (132 142) x (132 134) XxY, Xx(1), rel(2)x(1), (2)xX, (2)xrel(1)
      2.902902 3.642857 1.569231 1.593750 1.228916 (c)
      1.599820 2.522625 1.432242 1.268377 1.118312 (d)
Locus2 (188 200) = (188 194) x (190 200)  XxY, Xx(1), rel(2)x(1), (2)xX, (2)xrel(1)
      6.375000 8.500000 1.789474 1.500000 1.200000
      2.951418 4.899558 1.660992 1.204769 1.092875
Locus3 .....
-----

```

Combined over all loci
with observed allele frequencies: 1.49e+006 1.53e+004 2.09e+001 1.83e+003 1.34e+001 (e)
with 95% upper CI: 4.96e+003 5.50e+002 8.74e+000 1.65e+002 6.37e+000 (f)

- (a) Proposed parent – offspring group
- Per locus:
- (b) Locus with genotypes of involved individuals
- (c) Likelihood ratios including calculated allele frequencies
- (d) Likelihood ratios including 95% upper confidence limits of observed allele frequencies
- Combined over loci:
- (e) Likelihood ratios including calculated allele frequencies
- (f) Likelihood ratios including 95% upper confidence limits of observed allele frequencies

Microsat.dat: Data converted into MICROSAT (Minch 1997) input file format. The numbers in the first column of the IDENTITY input file are used as sample identifiers. Therefore, for analysis on “individual” level, all samples must be given different numbers in the first column of the IDENTITY input file, while groups of individuals (e.g. populations) can be defined by using identical numbers for their member individuals.

The usual disclaimer:

The software is provided “as-is” and without warranty of any kind, express, implied or otherwise, including without limitation, any warranty of merchantability or fitness for a particular purpose. In no event shall the authors or their institutions be liable for any special, incidental, indirect or consequential damages of any kind, or any damages whatsoever resulting from loss of use, data or profits, whether or not advised of the possibility of damage, and on any theory of liability, arising out of or in connection with the use or performance of this software.

References

- Bowers J.E., Meredith C.P., 1997: The parentage of a classic wine grape, Cabernet Sauvignon. *Nature Genetics* 16: 84-87
- Brookfield J.F.Y., 1996: A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol Ecol* 5, 453-455.
- Lopes M.S., Sefc K.M., Eiras Dias E., Steinkellner H., Laimer da Câmara Machado M., da Câmara Machado A., 1999: The use of microsatellites for germplasm management in a Portuguese grapevine collection. *Theor. Appl. Genet.* 99, 733-739.
- Maletic E., Sefc K.M., Steinkellner H., Kontic J.K., Pejic I., 1999: Genetic characterization of Croatian grapevine cultivars and the detection of synonymous cultivars in neighboring regions. *Vitis* 38, 79-83.
- Minch E. 1997: MICROSAT, Version 1.5b. Stanford University Medical Center, Stanford.
- Nei M. 1973: Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* 70, 3321-3323.
- Paetkau D., Calvert W., Stirling I., Strobeck C., 1995: Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol* 4 : 347-354
- Sachs L., 1997: *Angewandte Statistik*, 8th edition, Springer-Verlag, p 437 eq.(4.26a)
- Sefc K.M., Steinkellner H., Wagner H.W., Glössl J., Regner F., 1997: Application of microsatellite markers to parentage studies in grapevine. *Vitis* 36, 179-183.
- Sefc K.M., Regner F., Glössl J., Steinkellner H., 1998a: Genotyping of grapevine and rootstock cultivars using microsatellite markers. *Vitis* 37, 15-20.
- Sefc K.M., Steinkellner H., Glössl J., Kampfer S., Regner F., 1998b: Reconstruction of a grapevine pedigree by microsatellite analysis. *Theor. Appl. Genet.* 97, 227-231.
- Sefc K.M., Guggenberger S., Regner F., Lexer C., Glössl J., Steinkellner H., 1998c: Genetic analysis of grape berries and raisins using microsatellite markers. *Vitis* 37, 123-125.
- Sefc K.M., Regner F., Turetschek E., Glössl J., Steinkellner H., 1999a: Identification of microsatellite sequences in *Vitis riparia* and their applicability for genotyping of different *Vitis* species. *Genome* 42, 367-373.
- Sefc K.M., Lopes M.S., Lefort F., Botta R., Roubelakis-Angelakis K.A., Ibanez J., Pejic I., Wagner H.W., Glössl J., Steinkellner H., 1999b: Microsatellite variability in grapevine cultivars from different European regions and evaluation of assignment testing to assess the geographic origin of cultivars. *Theor. Appl. Genet.* in press.
- Weir B.S., 1996. *Genetic Data Analysis II*. Sinauer Associates, Inc. Publishers Sunderland, Massachusetts.