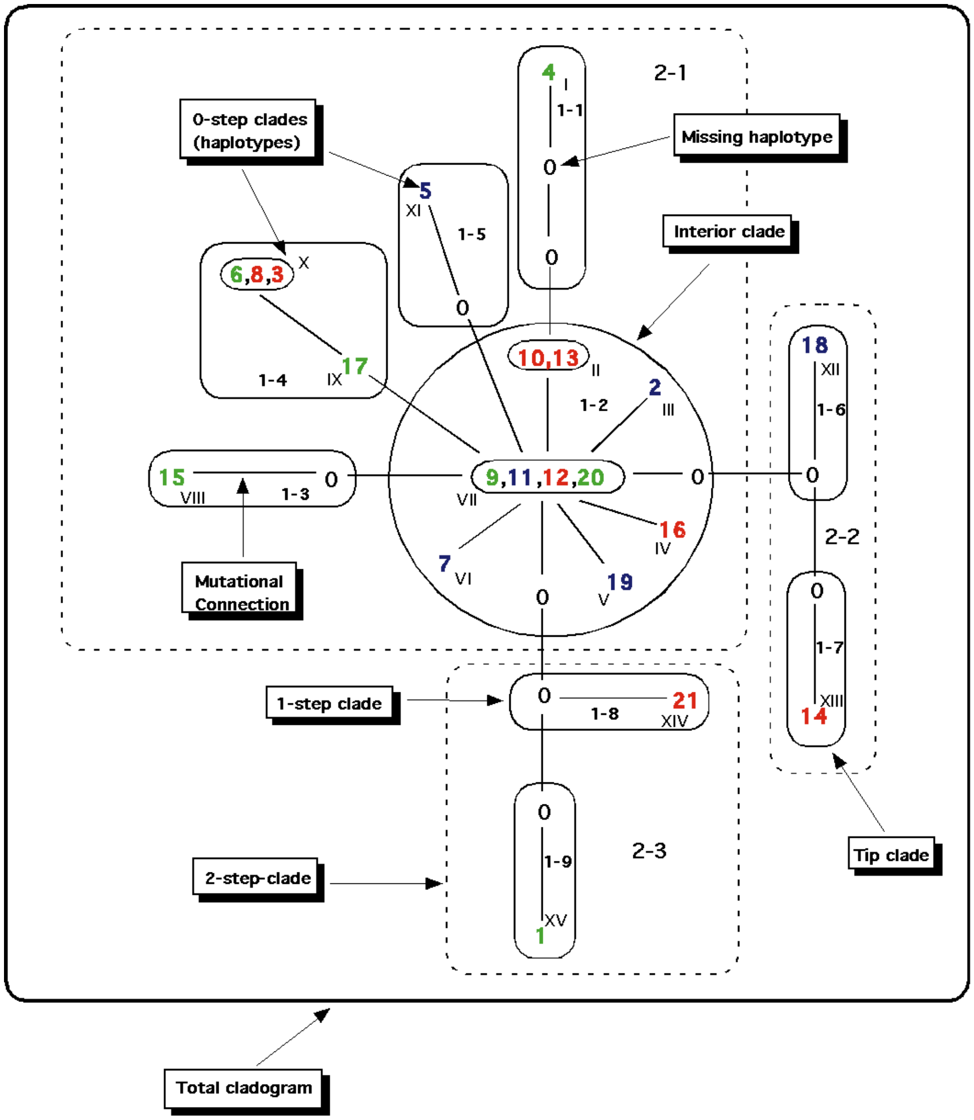
GEODIS 2.0 DOCUMENTATION

Contact: David Posada, Department of Zoology, 574 WIDB, Provo, UT 84602-5255, USA
Fax: (801) 378 7423 e-mail: dp47@email.byu.edu

1. INTRODUCTION

GeoDis is program written in C and Java (two different programs that implement the same calculations) implementing the nested cladistic analysis developed by Templeton et al. (1987). Its input consists in the description of a nested cladogram (Templeton & Sing 1993) estimated from RFLPs or DNA sequences. The theory and applications are described elsewhere (see recommended reading). The first step is the estimation of a cladogram and the defining of a nested structure. The cladogram estimation is described in Templeton et al., (1992) and the nesting rules are described elsewhere and extended in Crandall (1996). We are currently working in the development of software for the cladogram estimation. Meanwhile, you can find some tools to help you building the cladogram at http://bioag.byu.edu/zoology/crandall_lab/programs.htm. Outgroup probabilities (Castelloe & Templeton 1994) can also be included in the analysis. Here is a typical nested cladogram:



This cladogram consists of 21 individuals corresponding to 15 haplotype. The nested cladogram is described below in the input file for GeoDis.

2. INPUT FILE

The first line on the file is the name of the data set being analyzed. After that, the population information is indicated:

2.1 Populations

The description of the populations can be specified by their coordinates and sample size. However, in the case of riparian or coastal species, distances are not adequately measured simply through geographical coordinates, and a matrix of pairwise distances among the different locations better describes the geographical distribution in these one-dimensional habitats

2.1.1 Coordinates (2-dimensions)

2.1.1.1) Degrees, minutes and seconds

Latitude and longitude can be specified with the standard notation degrees, minutes and seconds, followed by the letter N (North) or S (South) in the case of latitude and E (East) or W (West) in the case of longitude. For example:

```
23 45 00 N      34 56 78 E
```

2.1.1.2) Decimal degrees

Latitude and longitude can be also be specified as decimal degrees. In this case latitude is expressed as 0-90 degrees (North {+} and South {-}), while longitude is expressed as 0-180 degrees (East {+} and West {-}).

For each population the format is:

Line 1: the population number and name is specified, for example : 1 Green Mountain

Line 2: the sample size, the latitude and longitude are indicated, for example:

```
7 60 22 01 N 15 20 34 E   or    7  60.35  15.41
```

2.1.2 User-defined population pairwise distances (1-dimension)

This information is specified as a lower triangle matrix without a diagonal (the diagonal would be made by zeroes). The number of populations (i.e. the dimensions of the matrix, is specified above the matrix). The population number, name and size are specified at each line. The distance can be specified in any unit. A matrix for 5 populations would look like:

```
5
1    Pop-1 name    Pop-1 size
2    Pop-2 name    Pop-2 size    distance 2-1
3    Pop-3 name    Pop-3 size    distance 3-1    distance 3-2
4    Pop-4 name    Pop-4 size    distance 4-1    distance 4-2    distance 4-3
5    Pop-4 name    Pop-5 size    distance 5-1    distance 5-2    distance 5-3    distance 5-4
```

2.2 Clades

The next step in the input file is the description of the nested cladogram. Clades without geographical or genetic variation (e.g. 1-8) are not included in the analysis. Clades at one level are subclades at the next one (e.g., clade 1-5 is a subclade in the nested clade 2-1). 0-step clades are haplotypes. The information is specified using the nesting clade as the unit. For each *nesting clade*, the composition of the clades nested within is described. The clades nested within a nesting clade are denominated simply *clades*. Hence the specification of cladogram starts at the 1-step level. For each nesting clade, it follows this format:

Line 1 name of the nesting clade, for example Clade 1-1

Line 2 number of clades nested within this nesting clade.

Line 3 name of the clades nested within this nesting clade. At the nested 1-step level, the clades nested within are haplotypes. We can give a name to these haplotypes, for example I, II, III,…. At higher nested levels (2-step, 3-step, 4-step…Total Cladogram), the name of these clades would we something like Clade 1-2, Clade 2-3,…)

Line 4 for each clade, its topological situation (tip = 1; interior = 0) is specified.

Line 5 number of populations represented in the nesting clade

Line 6 the populations are specified by their numbers

Line 7 In this line starts the observation matrix. The number of rows in this matrix corresponds to the number of clades specified in *line 2*, while the number of columns corresponds to the number of locations specified in *line 5*. For each row, and starting with the first clade (following the order specified in *line* 3), the number of individuals or copies of the clade is specified for each population.

Line (6 + number in line 2) last line of the observation matrix

This structure is repeated for each nesting clade. After the last nesting clade (the total cladogram), in the next line, the word "END" indicates the end of the input file.

2.2.1 Outgroup weights

Outgroup probabilities for each clade can be included in the analysis (see Castelloe and Templeton 1994). If so, they have to be specified for all the clades. The outgroup weights are specified for each clade as an extra line after line 4.

Line 4' For each clade, the corresponding outgroup probability is specified

3. RUNNING GeoDis

To run GeoDis, the input file needs to be specified. If an output file is not specified, the results are echoed to the screen. If the C version is used, the program prompts the user for all the needed information. For the Java version, the appropriate checkboxes need to be specified.

Number of permutations
A minimum number of 1000 permutations is recommend for a 5% level of statistical significance.

4. GeoDis OUPUT

The output of GeoDis saved to a file with the same name as the input file plus the extension .out. The value of the different statistics calculated is indicated for each nesting clade and its nested clades at each level. Two probabilities are indicated, those corresponding to significantly small (P <=) and large values (P >=) of the test statistic. It is highly encouraged to use the reference key in (Templeton et al., 1995) for a consistent interpretation of the output.

5. INPUT FILE EXAMPLES

1) With DMS coordinates and without outgroup weigths

```
Hallucigenia  mtDNA              // Name of the data set
3                                // Number of populations
1 Green Mountain                 // Population number and  name
7  15 41 12 N  60 21 12 E        // Sample size, latitude and longitude
2 Blue Mountain
6  17 16 21 N  61 45 00 E
3 Red Mountain
8  33 01 25 N  66 00 00 E
5                                   // number of clades in the file
Clade 1-2                           // name of the nested clade
6                                   // number of subclades included in the nested clade
II III  IV  V  VI VII               // name of subclades in the nested clade
 0   1   1  1   1   0               // position of each subclade: tip(1) or interior(0)
3                                   // number of populations in the nested clade
1  2  3                             // number of each population represented in the nested clade
0  0  2     // number of individuals in subclade II for each population
0  1  0     // number of individuals in subclade III for each population
0  0  1     // number of individuals in subclade IVfor each population
0  1  0     // number of individuals in subclade V for each population
0  1  0     // number of individuals in subclade VI for each population
2  1  1     // number of individuals in subclade VII for each population
Clade 1-4
2
IX X
 0 1
2
1 3
1 0
1 2
Clade 2-1
5
1-1 1-2 1-3 1-4 1-5
  1   0   1   1   1
3
1 2 3
1 0 0
2 4 4
1 0 0
2 0 2
0 1 0
Clade 2-2
2
1-6 1-7
  0   1
2
2 3
0 1
1 0
Clade 2-3
2
```

```
1-8 1-9
  0   1
2
1 3
0 1
1 0
Total Cladogram
3
2-1 2-2 2-3
  0   1   1
3
1 2 3
6 5 6
0 1 1
1 0 1
END
```

2) With user-defined distances and without outgroup weigths

```
Hallucigenia  mtDNA                // Name of the data set
3                                  // Number of populations
1 Green  7                         // Population number, name, sample size and distance (lower triangle matrix)
2 Blue   6    765
3 Red    8    234  256
5                                  // number of clades in the file
Clade 1-2                          // name of the nested clade
6                                  // number of subclades included in the nested clade
II III  IV  V  VI VII              // name of subclades in the nested clade
 0   1   1  1   1   0              // position of each subclade: tip(1) or interior(0)
3                                  // number of populations in the nested clade
1  2  3                            // number of each population represented in the nested clade
0  0  2                            // number of individuals in subclade II for each population
0  1  0                            // number of individuals in subclade III for each population
0  0  1                            // number of individuals in subclade IVfor each population
0  1  0                            // number of individuals in subclade V for each population
0  1  0                            // number of individuals in subclade VI for each population
2  1  1                            // number of individuals in subclade VII for each population
Clade 1-4
2
IX X
 0 1
2
1 3
1 0
1 2
Clade 2-1
5
1-1 1-2 1-3 1-4 1-5
  1   0   1   1   1
3
1 2 3
1 0 0
2 4 4
1 0 0
2 0 2
0 1 0
Clade 2-2
2
1-6 1-7
  0   1
2
2 3
0 1
1 0
Clade 2-3
2
1-8 1-9
  0   1
2
1 3
```

```
0 1
1 0
Total Cladogram
3
2-1 2-2 2-3
  0   1   1
3
1 2 3
6 5 6
0 1 1
1 0 1
END
```

3) With coordinates (decimal degrees) and with outgroup weigths

```
Hallucigenia  mtDNA                // Name of the data set
3                                  // Number of populations
1 Green Mountain                   // Population number and  name
7  15.41  60.35                    // Sample size, latitude, and longitude
2 Blue Mountain
6  17.67  61.81
3 Red Mountain
8  33.01  65.59
5                                  // number of clades in the file
Clade 1-2                          // name of the nested clade
6                                  // number of subclades included in the nested clade
II III  IV  V  VI VII              // name of subclades in the nested clade
 0   1   1  1   1   0              // position of each subclade: tip(1) or interior(0)
0.80 0.01 0.02 0.10 0.06 0.01      // outgroup probabilities
3                                  // number of populations in the nested clade
1  2  3                            // number of each population represented in the nested clade
0  0  2                            // number of individuals in subclade II for each population
0  1  0                            // number of individuals in subclade III for each population
0  0  1                            // number of individuals in subclade IVfor each population
0  1  0                            // number of individuals in subclade V for each population
0  1  0                            // number of individuals in subclade VI for each population
2  1  1                            // number of individuals in subclade VII for each population
Clade 1-4
2
IX X
 0 1
0.9 0.1
2
1 3
1 0
1 2
Clade 2-1
5
1-1 1-2 1-3 1-4 1-5
  1   0   1   1   1
0.75 0.05 0.05 0.10 0.05
3
1 2 3
1 0 0
2 4 4
1 0 0
2 0 2
0 1 0
Clade 2-2
2
1-6 1-7
  0   1
0.09 0.91
2
2 3
0 1
```

```
1 0
Clade 2-3
2
1-8 1-9
  0   1
0.05 0.95
2
1 3
0 1
1 0
Total Cladogram
3
2-1 2-2 2-3
  0   1   1
0.01 0.01 0.98
3
1 2 3
6 5 6
0 1 1
1 0 1
END
```

4) With user-defined distances and with outgroup weigths

```
Hallucigenia  mtDNA                 // Name of the data set
3                                   // Number of populations
1 Green  7                          // Population number, name, sample size and distance (lower triangle matrix)
2 Blue   6    765
3 Red    8    234  256
5                                   // number of clades in the file
Clade 1-2                           // name of the nested clade
6                                   // number of subclades included in the nested clade
II III  IV  V  VI VII               // name of subclades in the nested clade
 0   1   1  1   1   0               // position of each subclade: tip(1) or interior(0)
0.80 0.01 0.02 0.10 0.06 0.01       // outgroup probabilities
3                                   // number of populations in the nested clade
1  2  3                             // number of each population represented in the nested clade
0  0  2                             // number of individuals in subclade II for each population
0  1  0                             // number of individuals in subclade III for each population
0  0  1                             // number of individuals in subclade IVfor each population
0  1  0                             // number of individuals in subclade V for each population
0  1  0                             // number of individuals in subclade VI for each population
2  1  1                             // number of individuals in subclade VII for each population
Clade 1-4
2
IX X
 0 1
0.9 0.1
2
1 3
1 0
1 2
Clade 2-1
5
1-1 1-2 1-3 1-4 1-5
  1   0   1   1   1
0.75 0.05 0.05 0.10 0.05
3
1 2 3
1 0 0
2 4 4
1 0 0
2 0 2
0 1 0
Clade 2-2
2
1-6 1-7
  0   1
0.09 0.91
2
2 3
0 1
1 0
```

```
Clade 2-3
2
1-8 1-9
  0   1
0.05 0.95
2
1 3
0 1
1 0
Total Cladogram
3
2-1 2-2 2-3
  0   1   1
0.01 0.01 0.98
3
1 2 3
6 5 6
0 1 1
1 0 1
END
```

Recommend reading

The use of this program is pointless without the understanding of the methodology…

Castelloe J, Templeton AR (1994) Root probabilities for intraspecific gene trees under neutral coalescent theory. *Molecular Phylogenetics and Evolution* **3,** 102-113.

Crandall KA (1996) Multiple interespecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences. *Molecular Biology and Evolution* **13,** 115-131.

Georgiadis N, Bischof L, Templeton A *et al.* (1994) Structure and history of African elephant populations: I. Eastern and Southern Africa. *The Journal of Heredity* **85,** 100-104.

Hammer MF, Karafet T, Rasanayagam A *et al.* (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Molecular Biology and Evolution* **15,** 427-441.

Karafet TM, Zegura SL, Posukh O *et al.* (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. *American Journal of Human Genetics* **64,** 817-831.

Templeton AR (1998a) Human Races: A Genetic and Evolutionary Perspective. *American Anthropologist* **100,** 632-650.

Templeton AR (1998b) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology* **7,** 381-397.

Templeton AR (1998c) The role of molecular genetics in speciation studies. In *Molecular Approaches to Ecology and Evolution (*ed. De Salle R, Schierwater B), pp. 131-156. Birkhaüser-Verlag, Basel.

Templeton AR (1998d) Species and speciation: geography, population structure, ecology and gene trees. In *Endless forms: Species and Speciation (*ed. Howard DJ, Berlocher SH), pp. 32-43. Oxford University Press, Oxford.

Templeton AR (1999) Using gene trees to infer species from testable null hypothesis: cohesion species in the *Spalaxhrenbergi* complex. In *Evolutionary Theory and Processes: Modern Perspectives, Papers in Honour of Eviatar Nevo (*ed. Wasser SP), pp. 171-192. Kluwer Academic, Dordrecht.

Templeton AR, Boerwinkle E, Sing CF (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics* **117,** 343-351.

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132,** 619-633.

Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **134,** 659-669.

David Posada
June 99