IBD

Isolation By Distance

A program for population genetic analyses of isolation by distance

http://www.bio.sdsu.edu/pub/andy/IBD.html

Andrew J. Bohonak
San Diego State University

Citation:
Bohonak, A. J. 2002. IBD (Isolation By Distance): a program for analyses of isolation by distance. *Journal of Heredity* 93: 153-154.

v. 1.52 18 March 2004

Overview

IBD is a fast and simple Macintosh application to analyze isolation by distance in population genetic studies. For diploid, codominant markers (e.g., microsatellites, allozymes) **IBD** will

1) generate estimates of F_{ST} between all pairs of populations using Weir's (1990) estimator.
2) convert these to Slatkin's (1993) measure of similarity $\hat{M}$ and Rousset's (1997) distance $F_{ST} / (1-F_{ST})$.

If the geographic distances between populations are entered, **IBD** will also

3) assess whether the association between genetic similarity (or distance) and geographic distance is statistically significant using a Mantel Test.
4) calculate the slope and intercept of this relationship using Reduced Major Axis (RMA) regression.
5) calculate confidence intervals for the slope and intercept using multiple approaches.
6) repeat 4-5 using log (genetic similarity) and log (geographic distance) jointly and separately.

Alternatively, the genetic distances and geographic distances between all population pairs can be input directly. **IBD** will perform a Mantel test and carry out RMA regression analysis.

A third matrix can also be entered to assess the effects of another variable (e.g., population fragmentation, subspecies affiliation, alternative measures of geographic connectivity) on genetic distance. For clarity, the third matrix is referred to here as an "indicator matrix", although it can hold any values. If an indicator matrix is present, **IBD** will also

7) assess whether the association between genetic similarity (or distance) and the indicator variable is statistically significant using a Mantel Test.
8) calculate the partial correlation coefficients for genetic similarity (Y) as a function of geographic distance (X1) while controlling for the effect of the indicator variable (X2). The significance of this partial correlation coefficient is assessed using a partial Mantel Test.
9) calculate partial correlation coefficient for genetic similarity (Y) as a function of the indicator variable (X2) while controlling for the effect of geographic distance (X1).
10) repeat 7-9 using log (genetic similarity) and log (geographic distance) jointly and separately.

Rationale

The genetic similarity among individuals or populations can be ascertained using a number of statistical techniques. When populations can be defined a priori, one option is to analyze "isolation by distance" by plotting the genetic similarity (or distance) among population pairs as a function of the geographic distance between those pairs (Slatkin 1993, Rousset 1997, Hutchinson and Templeton 1999). These plots have a number of uses. For example:

1. Isolation by distance plots assess whether more distant population pairs are more different genetically
2. They can reveal the importance of specific barriers to gene flow
3. They may help separate the effects of population history from ongoing gene flow
4. The explanatory power of alternative dispersal pathways can be tested. For example, one might assess whether distance along a river or a topographic isocline is more biologically relevant than distance "as the crow flies".
5. Slopes and/or intercepts can be compared between different species, or the same species in different habitats.

Significance in the isolation by distance relationship can be tested statistically using a Mantel test. This test assesses whether the pairwise genetic distance matrix is correlated with the pairwise geographic distance matrix. A null distribution is generated by randomizing rows and columns of one matrix while holding the other constant. Because entire rows (populations) are treated as a single unit, the Mantel test is more appropriate than alternatives that assume that each population *pair* is independent.

Regression techniques must be used to estimate the slope and intercept of the IBD relationship, since the Mantel test provides only an assessment of whether the association is significant. Several options are available. Reduced Major Axis (RMA) regression is more appropriate than standard ordinary least squares (OLS) regression when the independent variable x is measured with error (Sokal and Rohlf 1981). (Assuming, of course, that the goal is to estimate the "true relationship" between the variables.) Error in the independent variable leads to biased estimates of slope. McArdle (1988) suggests as a rule of thumb that RMA should be used when the error rate in x exceeds one-third of the error rate in y. Hellberg (1994) specifically suggested that for isolation by distance analyses, RMA is a more appropriate estimator of slope than OLS .

It is often desirable to assess the partial correlations between genetic patterns, geographic distance and a third variable matrix. Throughout this manual that third variable is generically labeled “indicator variable” for clarity, although it might take many forms. For example, is there an effect of roads on genetic divergence, after the effect of geographic distance is taken into account? The indicator matrix would contain a 0 if the intervening distance between two populations was composed entirely of suitable habitat, or a 1 if a road separates the populations. Does distance along a particular environmental gradient correlate better with gene flow than direct-line distance? In this case, the partial correlation coefficients for alternative distance matrices would be useful. **IBD** analyzes these relationships using partial Mantel tests (described below).

Published Studies Using **IBD**

Burridge, C. P., A. C. Hurt, L. W. Farrington, P. C. Coutin, and C. M. Austin. 2004. Stepping stone gene flow in an estuarine-dwelling sparid from south-east Australia. Journal of Fish Biology **64:** 805-819.

Coyer, J. A., A. F. Peters, W. T. Stam, and J. L. Olsen. 2003. Post-ice age recolonization and differentiation of *Fucus serratus* L. (Phaeophyceae; Fucaceae) populations in Northern Europe. Molecular Ecology **12:** 1817-1829.

Fuselli, S., E. Tarazona-Santos, I. Dupanloup, A. Soto, D. Luiselli, and D. Pettener. 2003. Mitochondrial DNA diversity in south America and the genetic history of Andean highlanders. Molecular Biology and Evolution **20:** 1682-1691.

Hufbauer, R. A., S. M. Bogdanowicz, and R. G. Harrison. 2004. The population genetics of a biological control introduction: mitochondrial DNA and microsatellite variation in native and introduced populations of *Aphidus ervi*, a parisitoid wasp. Molecular Ecology **13:** 337-348.

Jump, A. S., F. I. Woodward, and T. Burke. 2003. *Cirsium* species show disparity in patterns of genetic variation at their range-edge, despite similar patterns of reproduction and isolation. New Phytologist **160:** 359-370.

Marko, P. B. 2004. 'What's larvae got to do with it?' Disparate patterns of post-glacial population structure in two benthic marine gastropods with identical dispersal potential. Molecular Ecology **13:** 597-611.

Vianna, P., R. Schama, and C. A. M. Russo. 2003. Genetic divergence and isolation by distance in the West Atlantic sea anemone *Actinia bermudensis* (McMurrich, 1889). Journal of Experimental Marine Biology and Ecology **297:** 19–30.

System Requirements

IBD for Windows requires Windows 95 or later (including Windows NT). **IBD** will run on Intel x86, Intel Pentium and compatible processors.

For Power Macintosh, **IBD** RAM requirements vary from 3 MB to more than 10 MB, depending on size of the data set. For very large data sets, a G3 processor or faster is suggested. Extensive testing as been conducted using System 9.1-9.2.2 on G3 and G4 processors, although the application is expected to perform well on System 8. The system should perform well on OS X under Classic mode, but this has not been tested extensively.

Installation of **IBD** only requires downloading and expanding the SEA (self expanding archive) available on the **IBD** web site. For Macintosh, if the default RAM requirements are insufficient, they can be changed manually within the Finder (through the Get Info->Memory menu).

Critiques, suggestions, commentary and bug reports are welcomed. Send email to <bohonak@sciences.sdsu.edu> and include the version number and error number (if applicable). Updates will be made available periodically at <http://www.bio.sdsu.edu/pub/andy/IBD.html>.

Parameters and settings

Upon launching the program, six parameters and settings are displayed. Enter the corresponding number to adjust each one.

General settings

0 Perform 1000 randomizations
Can be changed to any number between 50 and 100,000. For moderately sized data sets, at least 10,000 randomizations are recommended.

1 Analyze original genetic distances and log(genetic distance)
Toggle this setting to exclude additional analyses with the genetic distance log-transformed.

2 Analyze original geographic distances and log(geographic distance)
Toggle this setting to exclude additional analyses with the geographic distance log-transformed.

3 Calculate Rousset's distance (Raw Data format only)
Toggle this setting to calculate the genetic distance $F_{ST}/(1–F_{ST})$ as suggested by Rousset (1997).

4 Residuals from RMA regression will not be saved
Toggle this setting to save residuals from the RMA regression. Hutchinson and Templeton (1999) suggest that analysis of the residuals will provide some information about equilibrium vs. nonequilibrium conditions.

5 (Raw data format) Summary stats for each locus will not be saved
Toggle this setting to save all of the summary statistics from raw data analysis to the outfile. (By default, detailed locus by locus results only appear on the screen.)

When NOT log-transforming genetic distance ...

6 Negative genetic distances will be used without transformation
Alternatively, all negative genetic distances can be set to zero for the first set of analyses. For log-transformation, zero genetic distances will be adjusted a second time (see next setting).

When log-transforming (genetic distance) ...

7 Genetic distances of zero will be set to 0.000100
Zero values cannot be log-transformation. Enter a very small number to be substituted for genetic distances/similarities of zero, just prior to log transformation.

8 Negative genetic distances will be set to 0.000100
Negative values cannot be log-transformation. Enter a very small number to be substituted for all negative genetic distances, just prior to log transformation. If setting #4 is toggled to "Negative genetic distances will be set to zero", then this option is irrelevant and will be disabled.

When calculating Rousset's distance = F/[1-F] ...

9 Genetic distances of 1.00 will be set to 0.999900
F = 1 renders the above expression undefined. Should this occur, a number close to 1.0 will need to be substituted.

Parameter limits

Maximum number of populations	100
Maximum number of population pairs	4950
Maximum number of loci	30
Maximum number of alleles per locus	500
Maximum number of individuals per population	1000
Minimum number of randomizations/ bootstraps	50
Maximum number of randomizations/ bootstraps	100,000

Statistical considerations

Obviously, not all points in the isolation by distance relationship are independent. A single population will be involved in multiple pairwise contrasts. The Mantel test is expected to provide an appropriate test of significance for isolation by distance because it appropriately considers the unit of replication to be a population (and not a pairwise contrast). Similarly, to generate confidence limits for the RMA slope or intercept of an isolation by distance plot, bootstrapping over independent population pairs would seem to be the most conservative approach (see 5e under *Screen Output* below). For a small number of populations, jackknifing over populations provides the next best alternative.

Partial Mantel tests

In some cases, analysis of three matrices may be of interest (e.g., a dependent matrix of genetic similarity, an independent matrix of geographic distances, and a second independent "indicator" matrix that codes for particular geographic features). In this case, a partial Mantel test for the three matrices is analogous to a multiple regression on one "Y" variable and two "X" variables. **IBD** conducts partial Mantel tests using methods described in Legendre and Legendre (1998). Significance values in these tests may be somewhat biased (see Raufaste and Rousset 2001, Castellano and Balletto 2002, Rousset 2002), although it seems that the bias may be small in many situations (see Castellano and Balletto 2002 and Fig. 1 in Rousset 2002).

Input File Format

The input file requirements are as follows:

1) The file must be a text file. Tab-delimited text files generated by a spreadsheet application (e.g., Microsoft Excel) work well. Any number of white space characters (space or tabs) can separate fields with in a line, and no fields should be left blank.
2) The file must be in the same folder as the **IBD** application.
3) The file name should not contain any spaces. (Windows users should note that, depending on system settings, a text file may have a hidden ".txt" suffix. If the file name entered by the user cannot be found, **IBD** will attempt to search for alternate files with a ".txt" or ".TXT" suffix.)
4) Three file formats are recognized, as described below. Some error-checking routines are included, to verify that the file structure is correct.

Pairwise distance format

Example file "IBD.3" follows the pairwise distance format.

1) The first line in the input file must declare in capital letters:
   `GENETIC_DISTANCE`
2) Successive lines designate the genetic distance or similarity between all population pairs. All populations must be designated with a unique number. Each line contains:
   `<pop. A> <space or tab> <pop. B> <space or tab> <genetic distance>`
   The population pairs must be entered in a consecutive order, so that all pairs beginning with the first population are entered first, then all remaining pairs that begin with the second population and so on. This allows **IBD** to easily check for missing pairs. For example, a file with four populations would be entered as:
   ```
   1 2 0.758900
   1 3 0.313202
   1 4 -1.026451
   2 3 -0.089314
   2 4 -1.337712
   3 4 -0.633899
   ```
   The populations do not need to be consecutively numbered; for example, population `2` could be renamed `5`. However, each line (consisting of a population *pair*) must be entered in *consecutive, nested order*. If the line beginning "1 4" were moved to the end, **IBD** would generate an error message.
   Negative genetic distances are possible in some cases (e.g., F_{ST} estimated using Weir's methods), and can be handled in several ways by **IBD** (see *Parameters and Settings*).
3) After entering the genetic distances, the next line must declare:
   `GEOGRAPHIC_DISTANCE`
4) Successive lines designate the geographic distance between all population pairs in the exact same order and format as 2) above. Only geographic distances > 0 are permitted.

Matrix distance format

Example files "IBDM1.3" and "IBDM2.3" follow the two matrix distance formats described below. Note that they contain the same data as in file "IBD.3".

1) The first line in the input file must declare in capital letters:
   `GENETIC_GEOGRAPHIC_MATRIX`
   or
   `GEOGRAPHIC_GENETIC_MATRIX`
   The former indicates that genetic distances are above the diagonal, and geographic distances below. The latter indicates that geographic distances are above the diagonal, and genetic distances below.

2) The second line must delimit the number of populations as follows:
   `? POPULATIONS`
   where a number replaces the question mark.

3) The remainder of the file should contain a matrix with POPULATIONS rows and columns. A tab or space character separates each field within a row. Diagonal elements should be set to zero, but will be ignored. For example, if a file with 4 populations began with the declaration GENETIC_GEOGRAPHIC_MATRIX, then the second row of data would be:
   `<geogr.dist. 1->2>  0  <genetic dist. 2->3>  <genetic dist. 2->4>`
   Examples files "IBDM1.3" and "IBDM2.3" are provided in the IBD folder.

4) Negative genetic distances are possible, and can be handled in several ways by **IBD** (see *Parameters and Settings*). Only geographic distances > 0 are permitted.

Raw Data format

IBD also accepts raw genotype data for diploid, codominant markers (e.g., allozymes, microsatellites). The raw data format is similar to that required by TFPGA (Miller 1998). See example files "IBD.1" and "IBD.2".

1) The first line in the input file must delimit the file maxima as follows:
   `? LOCI ? ALLELES ? INDIVIDUALS`
   where the question marks are replaced by the number of loci in the data set, the maximum number of alleles per locus, and the maximum number of individuals per population.

2) Each line that follows lists the genotypes at all loci for a single individual. Each line begins with the population number, followed by the diploid genotype at each locus (e.g., `1,1  2,3  4,6`). (Allele 'names' must be numbers.) A space or tab separates each genotype, and a comma separates the alleles at a locus. A missing genotype at a locus can be designated with "`0,0`". For example, an individual in population 1 might look like:
   `1 1,3 2,2 3,3`
   Loci that are monomorphic across all populations will not affect the calculation of F-statistics or $\hat{M}$. These loci will be used to calculate average heterozygosity, however.

3) **IBD** knows it has reached the end of the genetic data when it reads a population number of "`0`", followed by genotypes of "`0,0`".

4) **IBD** will correctly parse files with loci that are invariable (i.e., only one allele was sampled).
5) (OPTIONAL): the geographic distances between all population pairs can be entered as described above (see 3-4 in *Pairwise distance format*). If geographic distances are not provided, **IBD** will analyze the genetic data, but Mantel tests and RMA regression will not be carried out. See file "IBD.2".

Optional: Indicator variables

A third matrix of "indicator variables" can be entered. The values may be continuous (e.g., an alternative estimate of geographic distance) or numerical but categorical (e.g., 0 if a pair of sampling sites are separated by suitable habitat; 1 if they are separated by inhospitable terrain).

Indicator variables in pairwise format

Example file "IBDI.3" shows the indicator variables 0 and 1 entered in pairwise list format. After the genetic and geographic data are entered, the next line must declare in capital letters: `INDICATORS`. Indicator variables are entered in pairwise list format as described above. Note that the "indicator variables" can be continuous (e.g., an alternative estimate of geographic distance) or categorical (0 if a pair of sampling sites are separated by suitable habitat; 1 if they are separated by inhospitable terrain).

Indicator variables in matrix format

Example file "IBDMI.1" shows the indicator variables 0 and 1 entered in matrix format. After the genetic and geographic data are entered, the next line must declare in capital letters: `INDICATOR_MATRIX`. The lower left portion of the matrix will be read. The diagonal values and the upper right portion of the matrix will be ignored, but must be included.

Screen output

After launching the application, enter the # portion of the input file name. (This file must be in the same folder as the application.) Next, enter the number of randomizations desired for the Mantel test and bootstrapping. **IBD** will provide the following:

For *Raw data format* only:

1) **Allele counts** for each population and locus
2) Locus-specific and overall $\boldsymbol{F_{ST}}$ **for each population pair**, estimated using the methods of Weir (1990).
3) Slatkin's (1993) similarity measure $\hat{M}$ **for each population pair**, estimated as $\hat{M} = \left(\frac{1}{F_{ST}} - 1\right)\Big/4$. If $F_{ST} \leq 0$ or F_{ST} is undefined (all loci fixed for a particular population pair), the user will be prompted to enter $\hat{M}$. This number should be ≥ the largest $\hat{M}$ in the data set. Because this

value will directly affect all calculated statistics, it should be chosen carefully. A preliminary examination of the $\hat{M}$ vs geographic distance plot is useful in this case.

If geographic distances have been entered the following output is available:

4) **Mantel tests**:

For the genetic distance matrix **A** and the geographic distance matrix **B** the test statistic *Z* is calculated as $\sum_{i,j} A_{ij} B_{ij}$. (Matrix **A** will consist of $\hat{M}$ for raw data files, or the user-entered genetic distances for pairwise distance files.) *r* is the correlation between the two matrices, ranging from -1 to +1; *r* is essentially a standardized *Z* (Manly, 1994). Significance is assessed by comparing r_{actual} to a distribution of *r*-scores obtained by randomizing rows/columns of the B matrix. One-tailed p-values for positive or negative matrix correlations are provided.

If an indicator matrix has been entered the following output is available:

5) **Partial Mantel tests**:

For convenience, denote the third "indicator matrix" **C**. Following Legendre and Legendre (1998), the *r* statistics *r*(**AB**), *r*(**AC**) and *r*(**BC**) are calculated as described above. The partial correlation *r*(**AB.C**) denotes the correlation between genetic distance and geographic distance, after controlling for the effects of matrix **C**.

$r(\mathrm{AB.C}) = \dfrac{r(\mathrm{AB}) - r(\mathrm{AC})\,r(\mathrm{BC})}{\sqrt{1 - r^2(\mathrm{AC})}\sqrt{1 - r^2(\mathrm{BC})}}$ and *r*(**AC.B**) is calculated similarly.

Significance of the partial correlations is calculated by comparing the actual statistic to a distribution of *r*-scores based on random permutations of the genetic distance matrix **A**. Potential biases to these significance values have been debated (Raufaste and Rousset 2001, Castellano and Balletto 2002, Rousset 2002).

6) **RMA regression:**

Following Sokal and Rohlf (1981), the RMA slope *b* is calculated as:

$$b_{\mathrm{RMA}} = \pm\sqrt{\frac{SSy}{SSx}} = \pm\sqrt{\frac{\sum y^2 - \left(\sum y\right)^2 / n}{\sum x^2 - \left(\sum x\right)^2 / n}}$$

As with OLS, the intercept $a_{\mathrm{RMA}} = \bar{y} - b_{\mathrm{RMA}}\bar{x}$ and $r^2{}_{\mathrm{RMA}} = SSxy / (SSx)(SSy)$.

Error estimation for *a*, *b* and r^2 is considered using five methods:

a) standard linear model formulas for RMA: $b_{\mathrm{RMA}}[\mathrm{SE}] = \sqrt{\dfrac{MSE}{SSx}}$, $a_{\mathrm{RMA}}[\mathrm{SE}] = \sqrt{MSE\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{SSx}\right)}$

(see Sokal and Rohlf 1981).

b) jackknife over <u>population pairs</u>: one-delete jackknife estimates of *a*, *b* and r^2, and the associated standard errors are calculated following Weir (1990). Confidence intervals are calculated using the *t* distribution and (# pairs – 2) degrees of freedom. *** Note that there is *not* general agreement on the validity of confidence intervals calculated from jackknife estimates of variance (see McArdle 1988,Dixon 1993, Manly 1997).

c) one-delete jackknife over <u>populations</u>

d) bootstrapping over population pairs: confidence intervals are calculated by creating new "pseudoreplicate" data sets, each with the same number of population pairs, by random sampling with replacement. The middle 95% and 99% of the bootstrap pseudoreplicates constitute the confidence intervals.

e) bootstrapping over independent population pairs: random data sets are created by sampling completely independent population pairs. For p populations, each data set will contain $p/2$ population pairs if p is even, otherwise $(p-1)/2$ pairs for odd p. For example, if $p=6$ populations, then one pseudoreplicate might be {(1,4), (3,6), (2,5)}.

7) If the appropriate settings are enabled, genetic and geographic distances will be log-transformed individually, and then jointly (see *Parameters and Settings*). For each of these three additional data sets, **IBD** will repeat 4) and 5). NOTE: Slatkin (1993) suggests log-transformation of both $\hat{M}$ and geographic distance.

Depending on the program settings, log-transformation of genetic distance values <= 0 may be handled in several ways:

- genetic distances of zero will be changed to a small constant prior to transformation (setting #5)
- negative genetic distances will either be ignored or transformed to the same constant (setting #6).

8) (Raw data format only) If the appropriate setting is enabled, the genetic distance $F_{ST}/(1-F_{ST})$ will also be calculated. All of the above analyses will be repeated. NOTE: Rousset (1997) suggests not log-transforming either axis for a one-dimensional stepping stone model, and transforming only the geographic distance for a two-dimensional stepping stone.

Output file

IBD saves a text-only output file following each batch of 1-8 analyses. This file contains the raw data matrices in column format, followed by a summary of the results for each analysis. Because fields in this file are tab-delimited, it should be easy to open and manipulate this file in a spreadsheet application such as Microsoft Excel.

Clicking "Save" when you quit the application will allow you to save a second text file with all of the screen output. (Note that some information from the screen output is not included in the tab-delimited output file. So click "Save" upon quitting to retain all of the results.)

Acknowledgments

I thank Victor Seguritan and George Roderick for constructive commentary and support. **IBD** has been written in C, and executables compiled using CodeWarrior v. 8 for Macintosh. Source code will be made available upon request.

Literature cited

Dixon, P. M. 1993. The bootstrap and the jackknife: describing the precision of ecological indices. Pages 290-318 *in* S. M. Scheiner, and J. Gurevitch, editors. The design and analysis of ecological experiments. Chapman and Hall, New York.

Castellano, S., and E. Balletto. 2002. Is the partial Mantel test inadequate? Evolution **56:** 1871–1873.

Hellberg, M. E. 1994. Relationships between inferred levels of gene flow and geographic distance in a philopatric coral, *Balanophyllia elegans*. Evolution **48:** 1829-1854.

Hutchinson, D. W., and A. R. Templeton. 1999. Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. Evolution **53:** 1898-1914.

Legendre, P., and L. Legendre. 1998. Numerical ecology. 2nd edition. Elsevier, New York.

McArdle, B. H. 1988. The structural relationship: regression in biology. Canadian Journal of Zoology **66:** 2329-2339.

Manly, B. F. J. 1997. Randomization and Monte Carlo methods in biology. 2nd edition. Chapman and Hall, New York.

Miller, M. P. 1998. TFPGA: Tools for population genetic analyses for Windows. Arizona State University.

Raufaste, N., and F. Rousset. 2001. Are partial mantel tests adequate? Evolution **55:** 1703-1705.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics **145:** 1219-1228.

Rousset, F. 2002. Partial Mantel tests: Reply to Castellano and Balletto. Evolution **56:** 1874-1875.

Slatkin, M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47:** 264-279.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry. 2nd edition. Freeman, New York.

Weir, B. S. 1990. Genetic data analysis: methods for discrete population analysis. Sinauer Associates, Sunderland, MA.