GenoDive

Version 2.0b23, Manual.

Software for analysis of population genetic data

October 2013

Patrick Meirmans

I.B.E.D. Universiteit van Amsterdam

p.g.meirmans@uva.nl

http://www.patrickmeirmans.com/software/

Contents:

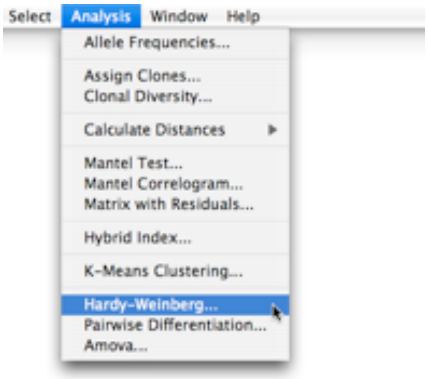
About GenoDive

Discover GenoDive

GenoDive is a user-friendly program to perform population genetics analyses. It features an intuitive and clutter-free user-interface, behind which lie powerful statistical tools, including several analyses that are not available in other programs. GenoDive can handle genetic data as well as distance matrices and ecological data, enabling you to combine data from several different sources into a single analysis. Best of all, GenoDive works on a Mac!

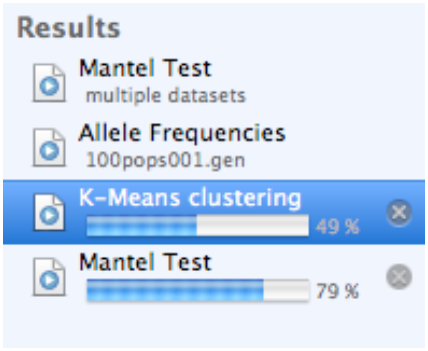
For questions about GenoDive or for reporting bugs, contact me at: p.g.meirmans@uva.nl

Many different types of inferences



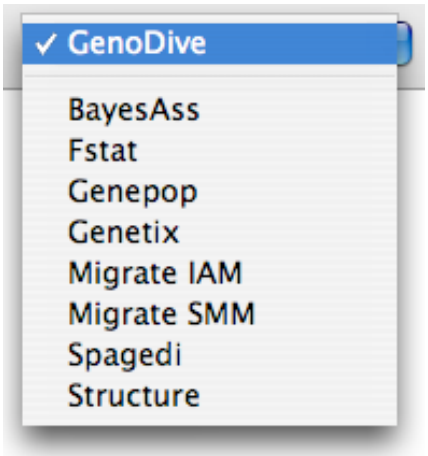
GenoDive features many different types of statistical inferences, some of which are not available in any other population genetics software. Included inferences are among others: a flexible Analysis of Molecular Variance, estimation of standardised coefficients of population differentiation, k-means clustering of populations using a simulated annealing approach, assigning genotypic identity (clones) to individuals, testing for clonal reproduction, testing Hardy-Weinberg equilibrium, calculation of the hybrid index for individuals, and different types of Mantel tests.

Analyse multiple datasets... simultaneously!



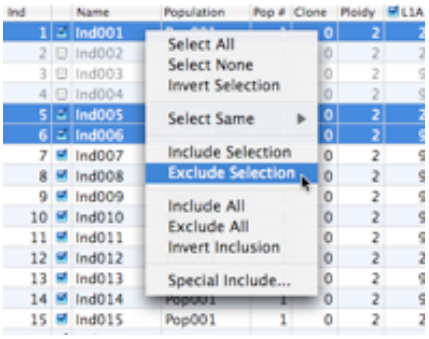
All Macs that are sold nowadays contain multiple processor cores. However, most programs are able to use only a part of the power of these new computers as they employ only a single processor at a time. GenoDive automatically detects the number of processors present in the computer and divides lengthy tasks over them all. Calculations always take place in the background, so you can perform other analyses while waiting for the results of a more lengthy analysis.

Multiple data formats



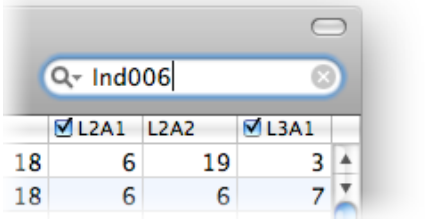
GenoDive can handle three different types of data: markerdata (e.g. from microsatellites), distance matrices (e.g. pairwise genetic distances between populations), and generic other data (e.g. the coordinates of populations), which can be combined in the statistical inferences. Importing data into GenoDive is easy: just open a file and GenoDive will automatically determine what type of data it is and will load the data into memory. GenoDive will even detect whether markerdata is in GenoDive, Fstat, Genepop, Spagedi, Genetix, BayesAss, Structure, or Migrate format.

Easily include and exclude observations



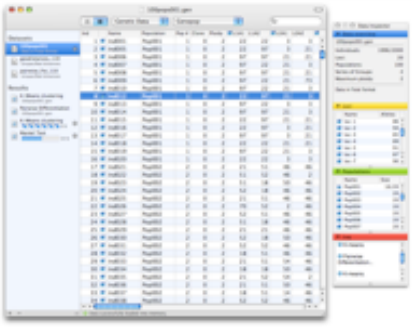
Most real-life datasets are far from perfect, some loci may be more difficult to score than others and there will always be individuals whose DNA-extraction yielded low quality DNA. In most programs this means quitting the program, editing the inputfile and then restarting the program again with the new file. GenoDive makes it easy to include or exclude loci, individuals, clones, ploidy levels, populations, or even whole groups of populations, so the effect of those outliers on the outcome of the analyses can be assessed.

Extensive Search Options



In GenoDive you can search your data in different ways: by allele, by individual name, by population, by clone, or by all those categories at the same time. The search can be very specific: you can for example search all population names that begin with the string "NL" or all alleles that are longer than 150 bp.

A clear view over your data



GenoDive has a clear, uncluttered, interface that makes it easier for you to keep an overview of your data and analyses. You can choose between viewing your data as plain text or viewing it as a matrix. The Inspector panel shows the most important information for the current data: how many individuals are included per population, the number of populations groups, which loci are included, the number of alleles per locus and a log with the analyses that have been performed on this dataset.

Known Issues and Limitations

This version of the program is still incomplete (beta). The fact that the program is still in beta means that the it will still contain bugs, and some analyses are not yet fully implemented. If you use this beta-version, please report any issues, bugs or praise that you may have to the author: Patrick Meirmans, email: p.g.meirmans@uva.nl

Even though I took great care to make the analyses correct, there is still a chance that there are some errors that have escaped my attention. It is therefore important to check the results carefully and see whether they make biological sense.

Most of the remaining issues are in the implementation of the AMOVA, though I hope that the available implementation is suitable for most users. In general: if you have data of a single ploidy level without too many missing values and with at least two populations per group and two individuals per population you are save.

Known bugs and issues:
Some important features are still missing.

- AMOVA does not yet allow for missing values. The current implementation fills in all missing data with randomly drawn alleles based on the overall allele frequencies.
- The permutation tests are disabled for AMOVA's performed on datasets with mixed ploidy levels. As a work-around, you can subsample your data to a single ploidy level (see Data transformation).
- AMOVA does not yet take singletons, so you should not have any populations with only one individual, or any population groups with only a single population.
- Not all Spagedi files are read succesfully, only the ones where alleles at a locus are not separated by non-numeric characters.
- On Mac OS 10.7 (Lion) and 10.8 (Mountain Lion) the program is very slow for datasets with a huge number of loci. This is a problem of Mac OS X and not of GenoDive, as far as I am aware. The solution is to view your data as text instead of as a matrix.

Planned features:

- Linkage Disequilibrium
- Principal Coordinates Analysis
- Redundancy Analysis

Data topics

Format of a GenoDive input file

Though GenoDive can read input files for several different population genetics programs, the preferred file format is the special GenoDive format. Only this format allows you to take advantage of some special features of GenoDive such as multiple series of population groups, polyploid data and assigning individuals to clones. A GenoDive input file consists of a text-only file with the following specifications.

GenoDive format specification:

1. First line: Comments, a single line of comments allowing you to give a short description of the data
2. Second line, five numbers separated by tabs:
   - The total number of individuals
   - The number of populations
   - The number of loci
   - The maximum ploidy levels used
   - The number of digits used to code a single allele
3. Next p lines: population names, on a separate line for every population. Optionally, the population name can be followed by the names of the groups to which this population belongs.
4. Line p+3, column headers, separated by tabs:
   - The total number of individuals
   - A generic name for populations, e.g. "Population"
   - A generic name for clones, e.g. "Clones", ONLY if the data contains a column with clones
   - A generic name for individuals, e.g. "Individuals"
   - The name of every locus.
5. Next lines, Individual data (separated by tabs):
   - The total number of individuals
   - The population number (not the population name) to which the individual belongs
   - The number of the clone to which the individual belongs (optional)
   - The name of the individual. This name should not begin with a number, not contain spaces and should be unique for every individual
   - The genotype of the individual at every locus
6. Genotypes are coded as follows: Every allele should be coded using the number of digits indicated on the first line and should be preceded by zeros if necessary (allele 12 should be coded as 012 if three digits are used). All alleles at a locus must be combined without spaces; the leading zeros are not necessary for the first allele at a locus (an homozygote for allele 12 can be coded 12012 if three digits are used). Missing data can be implemented by omitting the allele or by replacing it by the appropriate amount of zeros. These zeros can be either before, after or in between other alleles for the same locus (12, 012, 000012, and 012000 are all valid ways to code a locus with one allele 12 and one missing allele). If there is only missing data at a locus, there should be at least one zero. Differences in ploidy are coded in the same way as missing data.

GenoDive also supports dominant data, such as AFLP or RAPD. The format generally follows that of a normal data file, but the data should be coded as haploids and with one digit per allele. Since zeroes are reserved for missing data the presence-absence data cannot be specified as 0-1, but should be any other combination of two single-digit numbers, such as 1-2, 8-9, 9-5 or whatever you find most handy.

Example input file:

```
example input file (includes this comment line)
4       2       2       3       2
pop1
pop2
pop     ind     loc1    loc2
1       John    102     1214
1       Paul    202     0
2       George  101     121213
2       Ringo   10304   131414
```

Supported Data Types

GenoDive can handle three different types of data: Genetic marker data (e.g. microsatellite of AFLP data), distance matrices (e.g. genetic or geographical distances), and ecological data (e.g. coordinates or measured ecological variables). All three types of data can be read in several different formats. When opening a file, GenoDive will automatically determine the type of data and its format and will attempt to load the data (see importing data).

Data types and their formats:

Genetic marker data

- GenoDive (see format of a GenoDive input file)
- Fstat (Goudet, 1995)
- Genepop (Raymond & Rousset, 1995)
- Spagedi (Hardy & Vekemans)
- Genetix (Belkhir et al., 1996)
- Migrate (Beerli & Felsenstein, 1999)
- BayesAss (Wilson & Rannala, 2003)
- Convert (Glaubitz, 2004)
- Structure (Pritchard et al., 2000)

Distance matrices

- Square with titles: A square matrix where the first row and the first column contains the names of the observations.
- Square with titles: Only a square matrix, without any titles in the first row and column.
- Lower triangular: Only the part of the matrix below the diagonal, unfolded into a single column.
- Upper triangular: Only the part of the matrix above the diagonal, unfolded into a single column.

Ecological data

- Title row and column: A rectangular matrix with the observations in rows and the variables in columns, where the first column contains the observation names and the first row the variable names.
- Title column: A rectangular matrix, where the first column contains the observation names, but without variable names in the first row.
- Title row: A rectangular matrix, where the first row contains the variable names, but without observation names in the first column.
- No titles: A rectangular matrix, without any names in the first row and first column.

Importing data from other programs

Besides reading data in its own GenoDive-format, GenoDive can work with genetic data formatted for several different other population genetics programs: Fstat, Genepop, Spagedi, Genetix, Migrate, BayesAss+ and Structure. For specifications of these data formats, see the manuals of the respective programs. Working with these files should be as easy as working with files in GenoDive format, though they may not be able to contain exactly the same information.

- Migrate data can be both formatted for an Infinite Allele Model (IAM) and for a Stepwise Mutation Model (SMM).
- The Structure file format has many optional features, only few of which are supported by GenoDive. In order to be read by GenoDive, Structure files should have a first row with locus names, no inter-marker distances, and no phase-information. Genotypic data should be presented with several rows per individuals (so not all data on one row) and should always be preceded by the individual name and population data. No population flags or extra columns are allowed.
- Not all Spagedi files are read succesfully, only the ones where alleles at a locus are not separated by a non-numeric characters. Distance matrices that are included in Spagedi files are discarded.

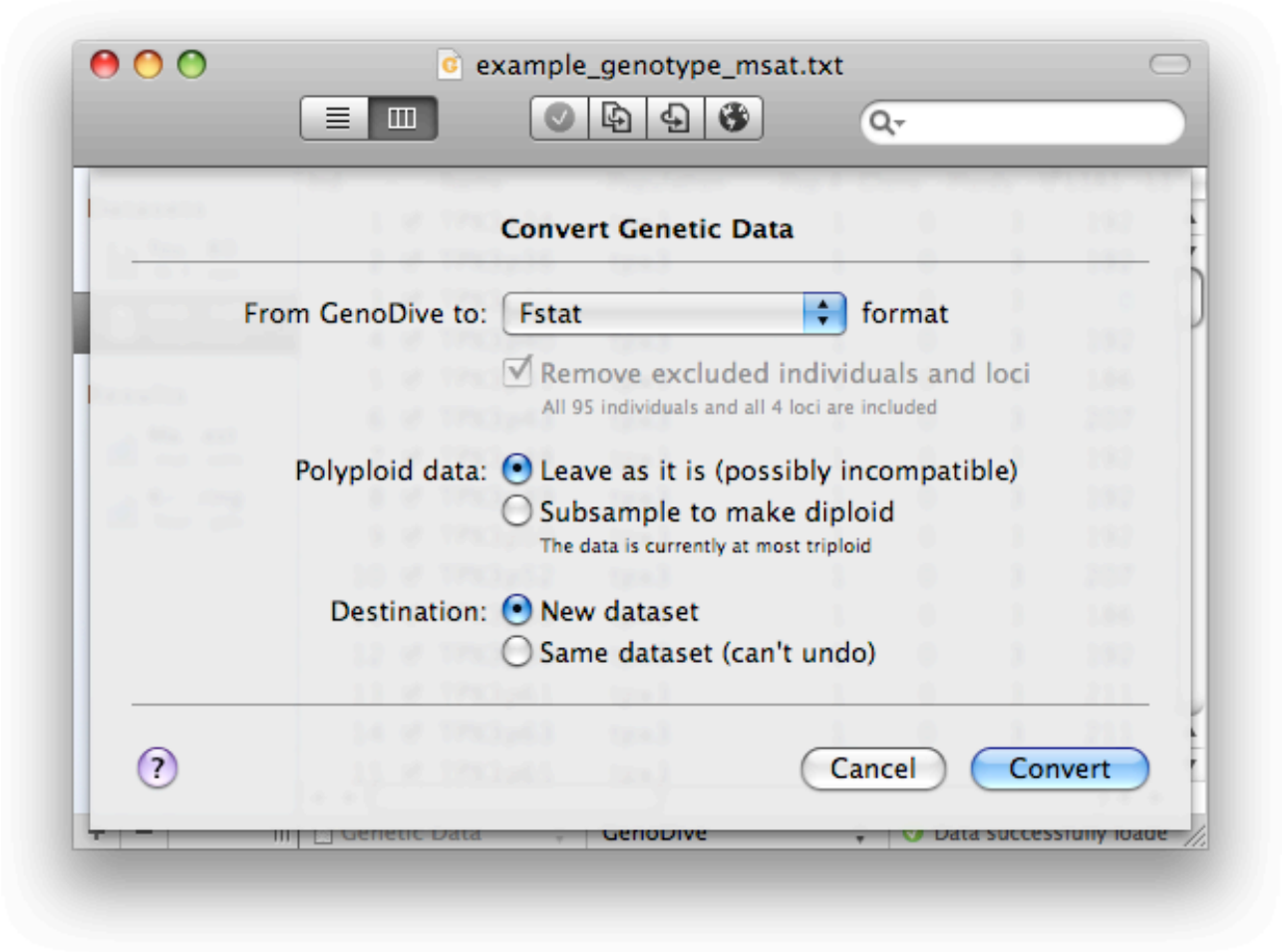
To import a data file:

1. Choose Open... from the File-menu. By default (this can be set in the preferences), GenoDive is set to automatically detect the file format at opening and attempt to load the data
2. If there is an error in the format of your data, GenoDive may not be able to recognise the format, and the data is opened as text. In this case, you should manually set the data format, using the pop-up menu in the toolbar. Then choose Check Format from the Data-menu to try to load the data into memory.

It is also possible to set the preferences so that GenoDive shows an import dialog when opening files. The correct file format can then be selected in the dialog. Alternatively, GenoDive can also be set to always open files directly as text, without attempting to determine the filetype or showing the dialog.

Converting Data

If you have successfully opened a data file, you can easily convert it to any other format that can be applied to that type of data. For example, when you have opened a square distance matrix you can convert it to a lower triangular matrix, or when you have opened a file with marker data coded in GenoDive format, you can convert it to Fstat format.



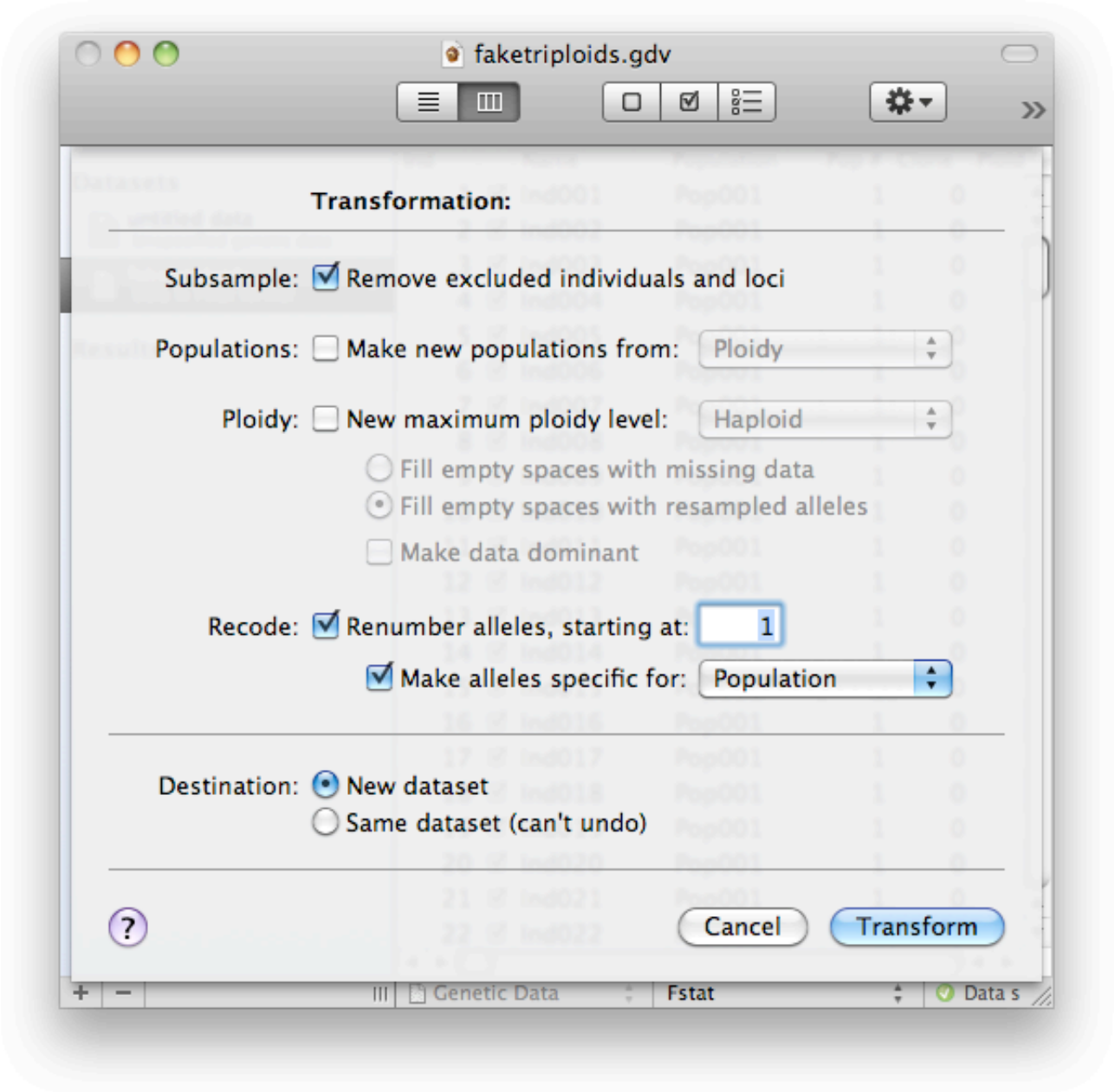
How to convert a file to a different format:

1. Make sure that the selected dataset does not contain results.
2. Choose Convert to Other Format from the Data menu.
3. In the dialog, select the desired format from the pulldown menu
4. If some of the data in the file are excluded, you can choose to subsample the data by switching on the "Remove excluded..." switch. Note that this may lead to a loss of data.
5. If the file contains polyploid marker data, you can choose to randomly subsample the alleles to make the individuals diploid. This will guarantee compatibility with programs that do not support polyploid data. If the file contains haploid marker data, you can choose to duplicate the alleles to create homozygous diploids.
6. Choose a destination window, this can be either a newly created window or the same window. In the latter case, the current data will be overwritten, possibly resulting in a data loss. This operation cannot be undone.
7. Click Convert

To quickly convert data, you can simply choose another format from the popup menu in the status bar after the data has loaded. Polyploid or haploid data will not be resampled and the current data will be overwritten in the new format. This operation cannot be undone, so you will be asked for a conformation.

Transformation of genetic marker data

GenoDive has several different ways to transform genetic marker data, as presented in the transformation dialog. The transformations are applied in the order as they are shown in the dialog, from top to bottom. So in the screenshot below, the data is first subsampled before the alleles are recoded.

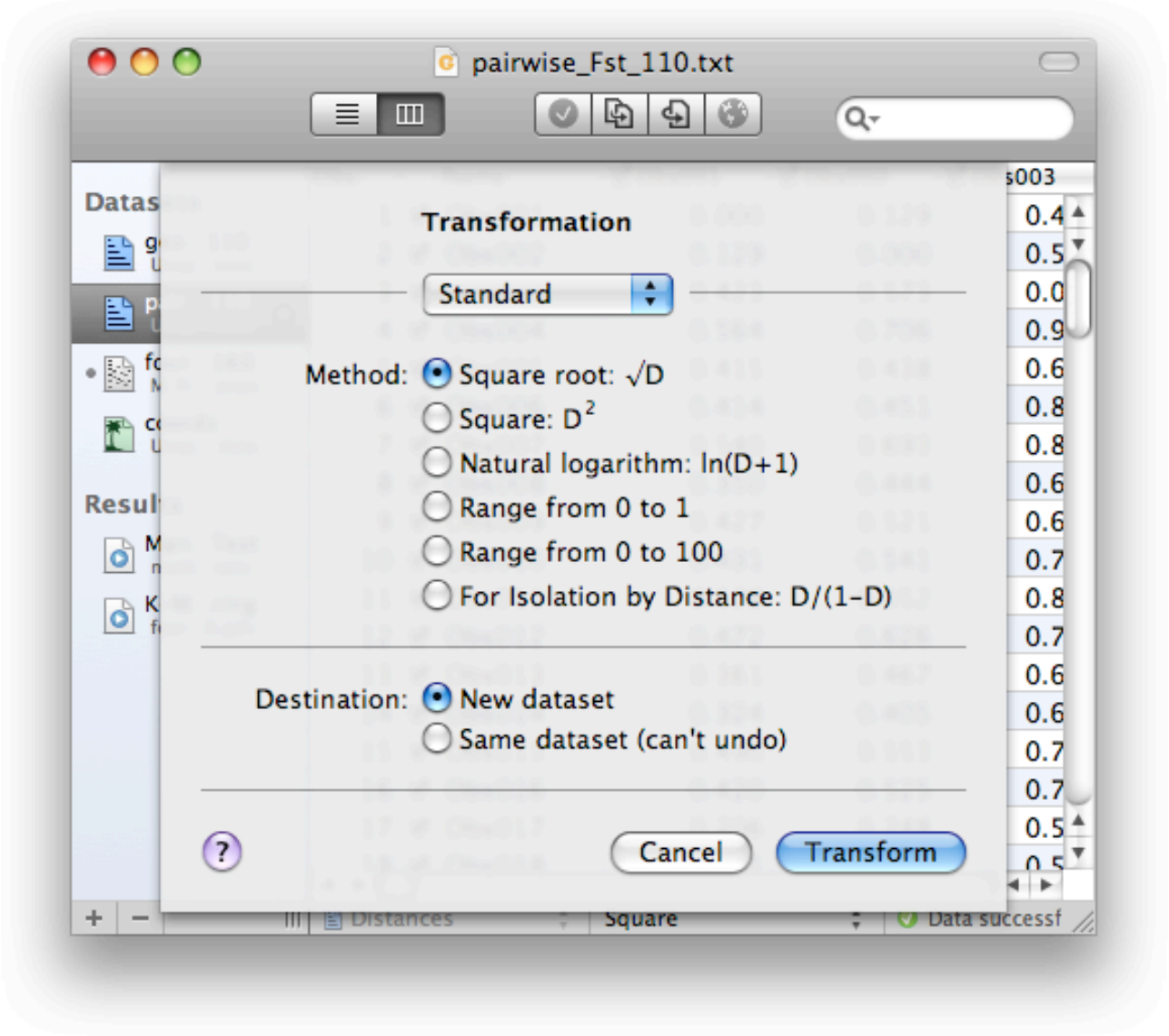


To transform marker data:

1. Make sure that the selected dataset contains genetic marker data.
2. Choose Transformation from the Data menu
3. To change the way individuals are distributed over populations, switch on the "Make new populations from:" button. The new populations can be based on ploidy level (if multiple levels are present), on clones (if clones have been defined), or on population groups (if any are defined).
4. To create a subsample using only the included individuals and loci, switch on the "Remove excluded individuals and loci" button.
5. To change the ploidy of the data, switch on the "New maximum ploidy level" button and choose a new maximum ploidy level from the popup menu. There are two ways in which an increase in ploidy can be implemented: adding missing data, or resampling the present alleles.
6. If the new ploidy is set to haploid, you choose to make the data dominant, by switching on the "Make data dominant" button. In this case the new data will contain as much biallelic loci as there are alleles in the current dataset.
7. To recode the data, switch on the "Renumber alleles..." button, and fill in a starting number in the text field. The lowest allele number found at a locus will be recoded top have this starting number, and higher alleles will be consecutively numbered.
8. You can also choose to make alleles specific to a certain category (population, or group of populations) by switching on the "Make alleles specific for" button and choosing the desired category from the popup menu. This option is useful if you want to estimate a standardised measure of genetic differentiation (Hedrick, 2005; Meirmans, 2006) using another program than GenoDive.
9. Choose a destination window, this can be either a newly created window or the same window. In the latter case, the current data will be overwritten, possibly resulting in a data loss. This operation cannot be undone.
10. Click Transform

Transformation of distances and ecological data

GenoDive has several different ways to transform distances and ecological data, as presented in the transformation dialog. Only a single transformation can be applied at a time. The transformation dialog has two modes: standard and advanced.



To transform data:

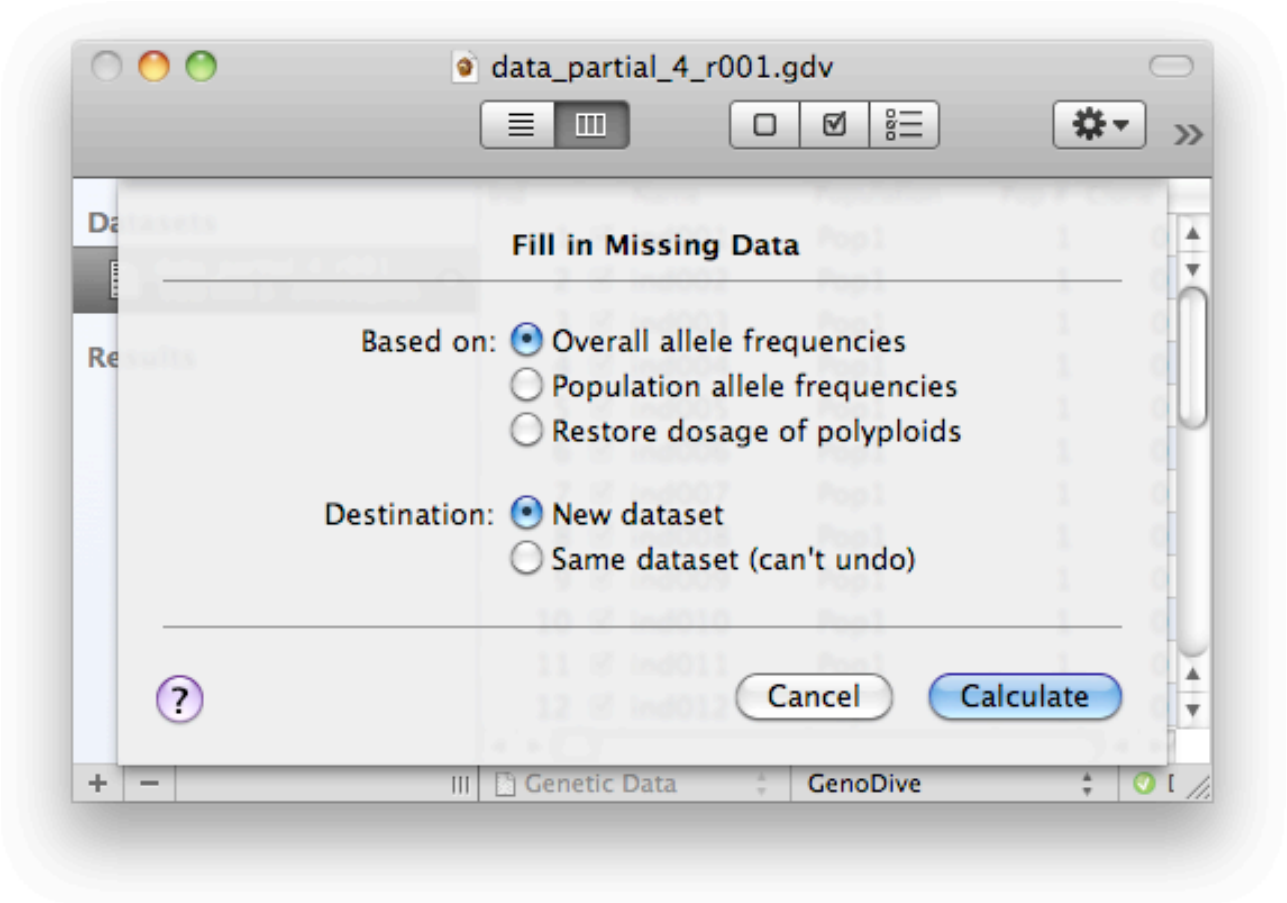
1. Make sure that the selected dataset contains either a distance matrix or ecological data.
2. Choose Transformation from the Data menu.
3. Select either the Standard mode or the Advanced mode from the popup menu.
4. In Standard mode, select the required type of transformation. There is the choice between: Square root, Square, Natural logarithm, Range the data from 0 to 1, Range the data from 0 to 100, and Isolation By Distance. The latter type of transformation is especially meant for converting distance matrices of pairwise Fst-values to Fst/(1-Fst). Under a two-dimensional stepping stone model, a linear relationship is expected between Fst/(1-Fst) and the logarithm of the distance between pairs of populations (Rousset, 1997).
5. In Advanced mode, you will be presented with a list of equation, which all contain the coefficients a and b. Select the required equation and fill in the appropraite values of a and b in the text fields at the bottom of the list.
6. Choose a destination window, this can be either a newly created window or the same window. In the latter case, the current data will be overwritten, possibly resulting in a data loss. This operation cannot be undone.
7. Click Transform

Fill in Missing Data

Some tests require that there are no missing values in a dataset, leading to biased results when there actually are missing values. For example, both a Principal Components Analysis and k-Means clustering tend to cluster observations together if they have missing data at the same loci.

To overcome any bias resulting from missing data, they can be replaced by random values. For allelic data, the most suitable random values are alleles that are randomly picked from the present pool of alleles, where alleles that are present at a high frequency is more likely to be picked than alleles at a low frequency. The baseline frequencies which are used for this can be either the overall allele frequencies or the population frequencies. The former is most suitable for the majority of cases, such as PCA's or K-Means clustering.

For polyploids, there is a special type of missing data that stems from the difficulties of scoring the dosage of alleles in partial heterozygotes. GenoDive also allows filling in such missing data based on the alleles scored for an individual. Alleles are randomly drawn based on the likelihood of the different genotypes given the population allele frequencies. So if a triploid has marker phenotype AB, and allele A is most frequent in the population, it is more likely to be filled to genotype AAB than to ABB. For most analyses, it is best to take the missing data directly by checking the "Correct for unknown dosage of alleles" box. See Correction of allele dosage in polyploids for more information.



One should be careful in applying this method, since a bias may also be introduced by overenthousiastic use. For example, tests for population differentiation will tend to give more false postives if missing values are filled based on the population allele frequencies. On the other hand, filling in missing data based on overall allelel frequencies will tend to make tests more conservative, so you may miss any structure that is present in the dataset.

In general, if you think that missing values may present a problem in your data, you should perform the test both with and without the missing data filled in. If the two tests show widely different results, the missing values infuence the results and you should be very careful in drawing any conclusions from the data. In that case, you should consider using the Special include option to only include individuals without missing data.

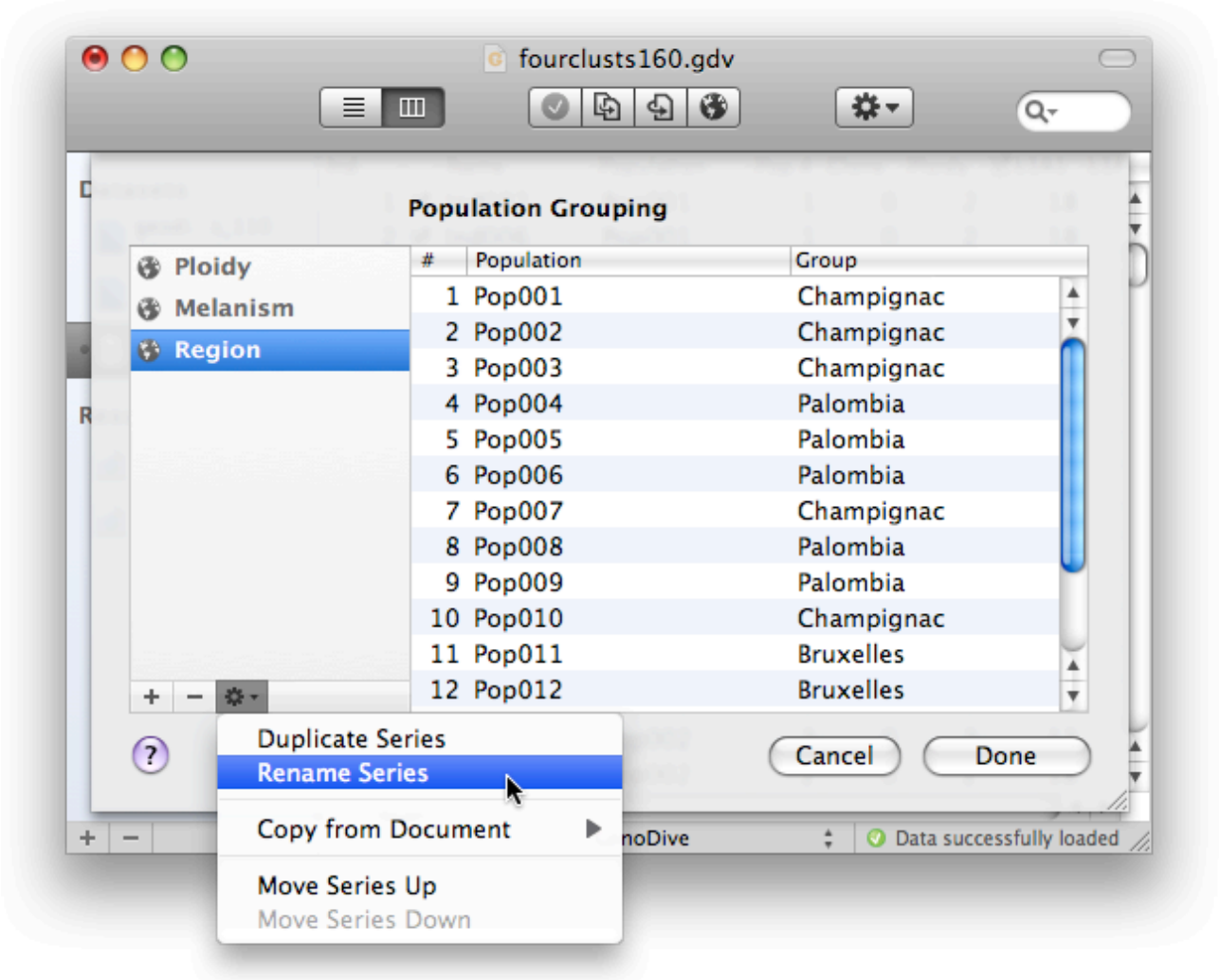
To fill in missing data:

1. Make sure that the selected dataset contains genetic marker data.
2. Choose Fill in Missing Data... from the Data menu.
3. Baseline allele frequencies used for drawing random alleles: overall allele frequencies, population allele frequencies, or restoring the unknown dosage of polyploids.
4. Choose a destination window, this can be either a newly created window or the same window. In the latter case, the current data will be overwritten, possibly resulting in a data loss. This operation cannot be undone.
5. Click Calculate.

Setting Population Groups

GenoDive supports multiple series of groups, this means that you can e.g. base one series of groups based on a North-South division and another one on an East-West division. This allows you to easily compare the effect of different groupings e.g. in an Analysis of Molecular Variance.

If your file is in GenoDive format, the population groups can be specified in the input file. if you did not specify the required groups, or if your file is not in GenoDive format, you will have to set the population groups in the groups-dialog.

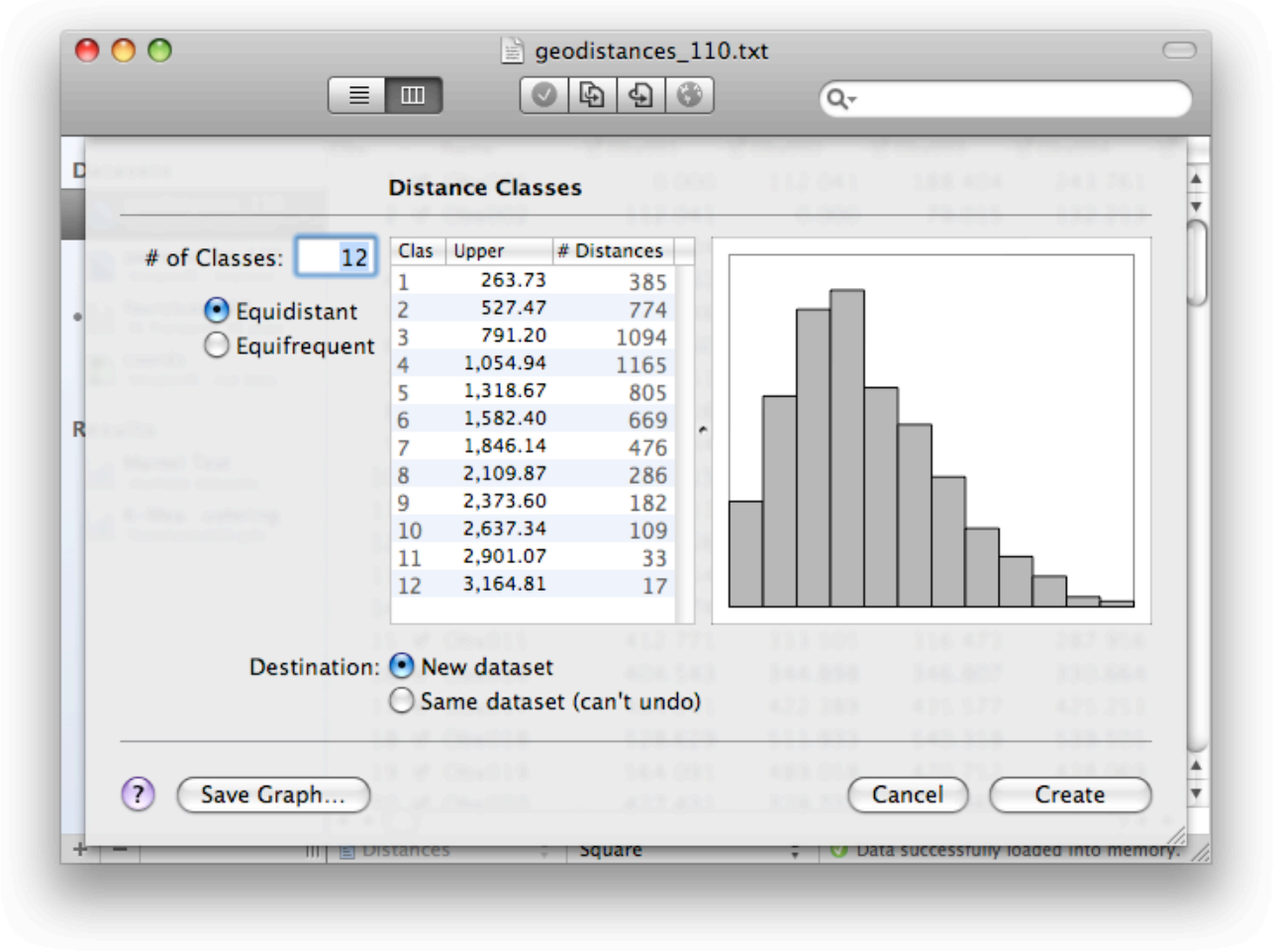


To define population groups:

1. Make sure that the selected dataset contains genetic marker data.
2. Choose Population Grouping from the Data-menu.
3. On the left-hand side of the Population Groupings dialog there is list with all series of population groups that are currently set for the data. In the table on the right there is, for the currently selected series, for every population the name of the group it belong to.
4. To add a series of population groups click the  button, to delete a series click the  button.
5. Some additional commands are presented in the menu that appears when clicking the  button. These commands allows you to duplicate a series of population groups, rename a series of population groups or move a series up or down in the list.
6. To change the group for a population in a series, double click the corresponding cell in the table to edit the name.
7. Click Done

Creating a matrix of distance classes

Tests of spatial autocorrelation, e.g. due to isolation by distance, often use a distance matrix where every distance has been assigned to one of a specified number of distance classes. GenoDive can create such a matrix of distance classes based on a matrix of geographic (or other) distances.



To create a matrix of distance classes:

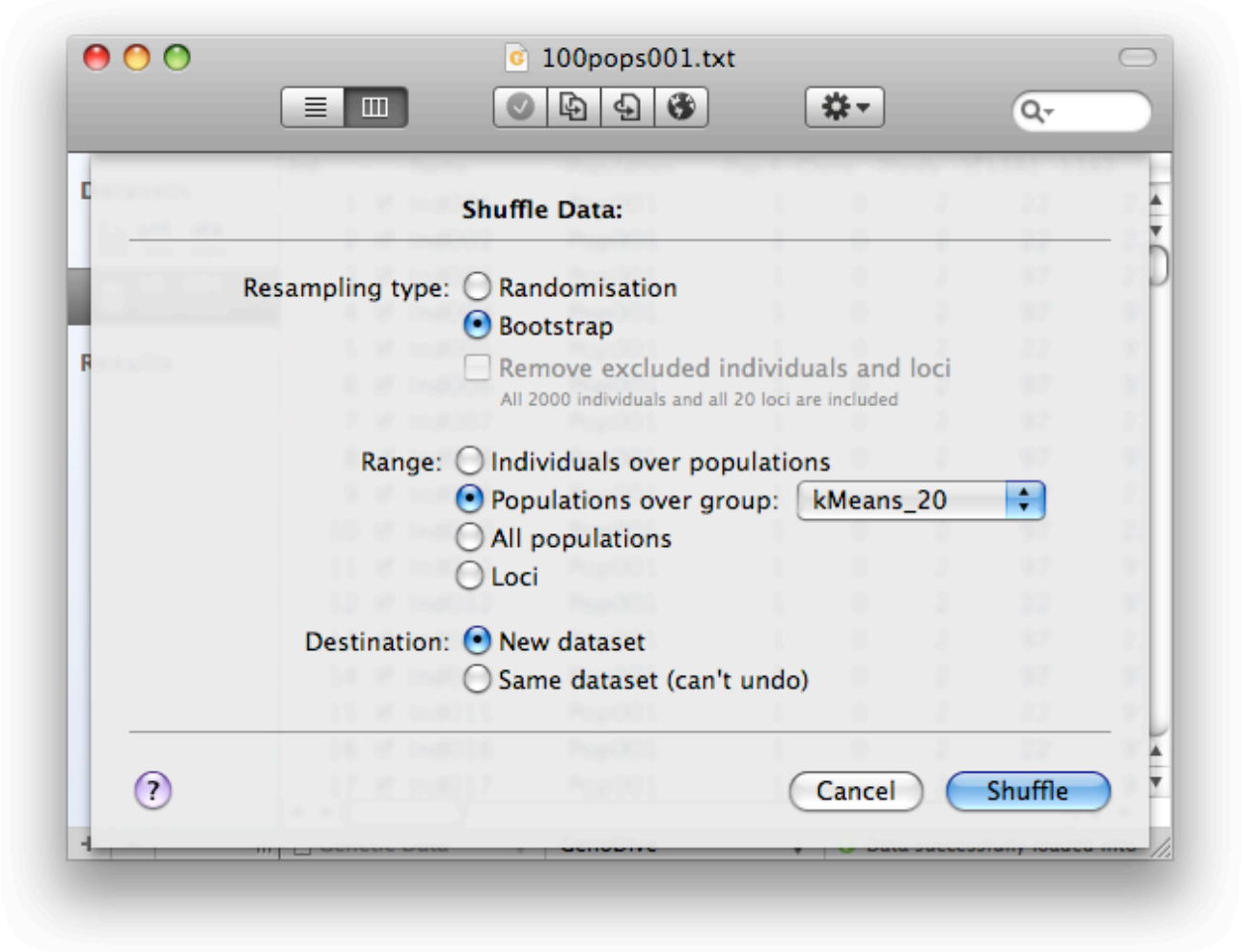
1. Make sure that the selected dataset contains a distance matrix. For information on how to calculate a matrix of geographical distances, see Non-Genetic Distances
2. Choose Distance Classes... from the Data menu.
3. Fill in the number of required distance classes in the text field at the top of the dialog. This field is fileld automatically with the number of distance classes according to Sturge's rule of thumb (see Legendre & Legendre, 1998): 1 + 3.3 * log(number of distances).
4. Choose whether the upper limits of the distance classes are spaced such that all distance classes are equally wide (Equidistant) or such that all distance classes contain approximately the same number of pairwise distances (Equidistant).
5. It is also possible to set the upper limits of the distance classes manually by double clicking the value in the table and entering the desired upper limit.
6. The histogram at the right-hand side of the dialog gives a graphical overview of how the distance classes are distributed, with on the x-axis the upper limit, and on the y-axis the number of pairwise distances.
7. Choose a destination window, this can be either a newly created window or the same window. In the latter case, the current data will be overwritten, possibly resulting in a data loss. This operation cannot be undone.
8. Click Create. The matrix with distance classes will be created at the chosen destination, and overview of the matrix is presented in the Results window.

Shuffle Data

Many standard methods in population genetics use resampled data to test for significance or to calculate the confidence intervals. Indeed, all tests in GenoDive are based on permutations. In some cases in may be handy to be able to randomise data yourself, e.g. in order to see what values of a certain statistic can be found with completely random data, or to calculate bootstrapped confidence intervals for certain genetic distances. GenoDive therefore contains a data shuffling option with which genetic data can be randomised in different ways, based on the two main methods to resample data: randomisation and bootstrapping.

Under randomisation, the data is kept intact, but their order is changed. This is also known as resampling without replacement. For example, when randomising individuals over two populations, their population labels are reordered. Both populations still have the same size as before, but the individuals they contain have been changed. Randomisations are often used for significance testing as they can generate data that conforms to the null-hypothesis.

Bootstrapping is resampling with replacement. Under bootstrapping a new dataset is created by repeatedly picking observations from the dataset. Because of the replacement, the resampled dataset will include some observations repeatedly, while others have not been included at all. Bootstraps are best known for their use for estimating confidence intervals, but they can also be used for significance testing (such as implemented in GenoDive's pairwise test for a difference in clonal diversity).



How to resample your data:

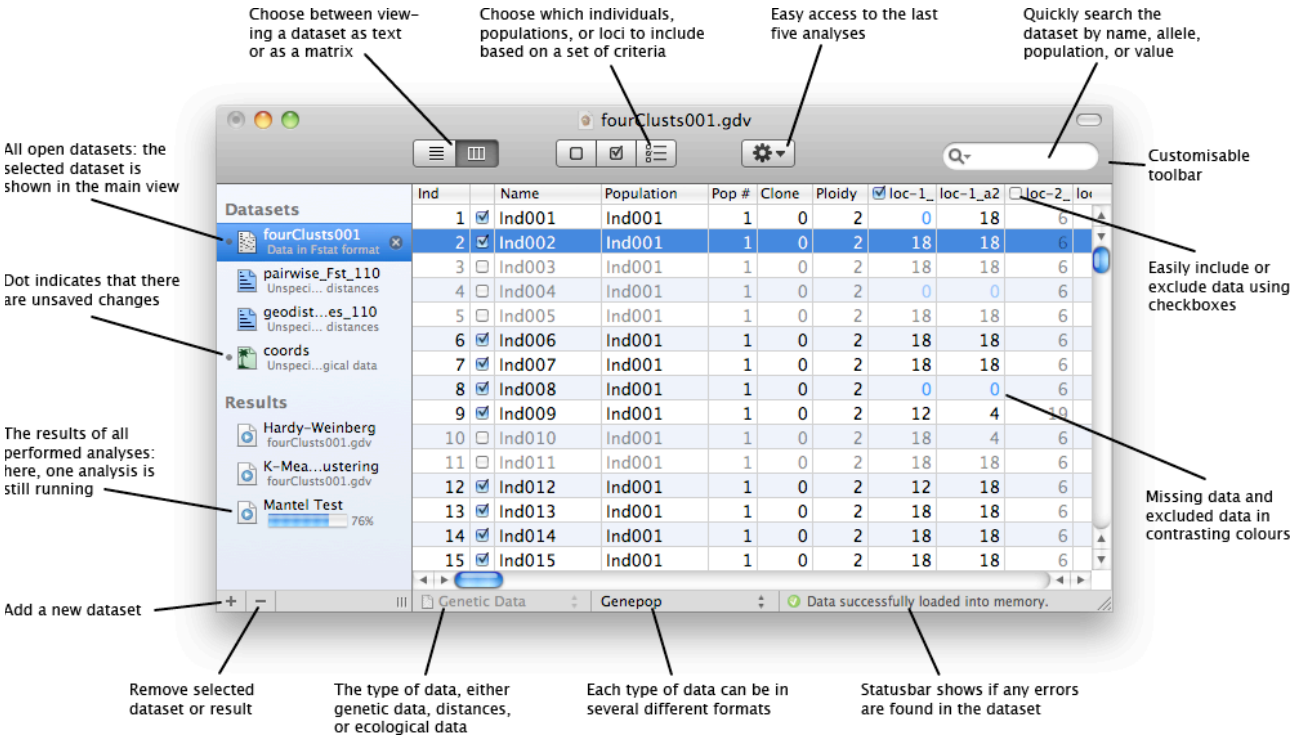
1. Make sure that the selected dataset contains genetic data.
2. Choose Shuffle Data from the Data menu.
3. In the dialog, choose whether you want to use randomisations of bootstraps.
4. If some of the data in the file are excluded, you can choose to subsample the data by switching on the "Remove excluded..." switch. Note that this may lead to a loss of data.
5. Select the range over which you want the resampling to be performed. For randomisations, you can choose to either randomise individuals over populations, or populations over groups. For bootstrapping, you can also choose to bootstrap over all populations, or over loci.
6. Choose a destination window, this can be either a newly created window or the same window. In the latter case, the current data will be overwritten, possibly resulting in a data loss. This operation cannot be undone.
7. Click Convert

Interface topics

Understanding the interface

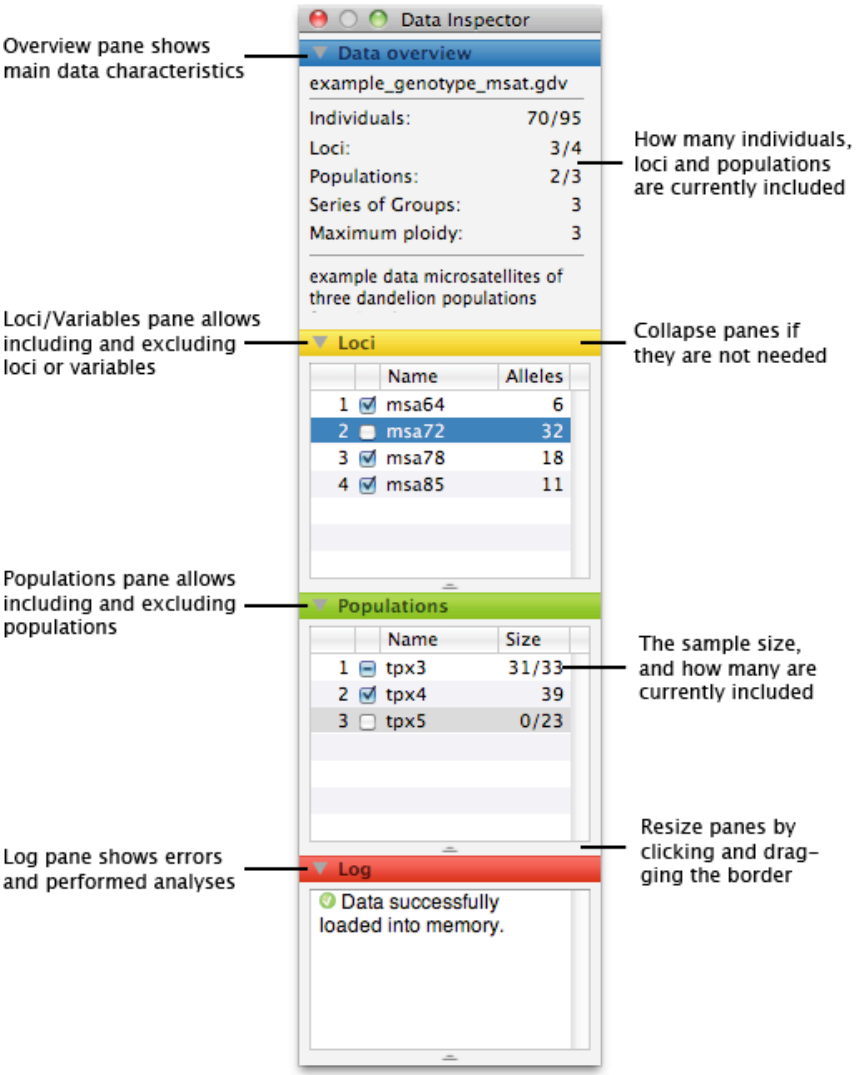
GenoDive features a document-based interface. This means that it is possible to have several datasets open at the same time, making it easy to perform analyses that combine data from multiple documents: e.g. perform an AMOVA using a custom distance matrix, or cluster populations using a distance matrix calculated on ecological data. The document-based interface even allows you to continue working while an analysis works in the background.

GenoDive has a single window with a side-bar in which all currently open datasets and all performed analyses are shown:



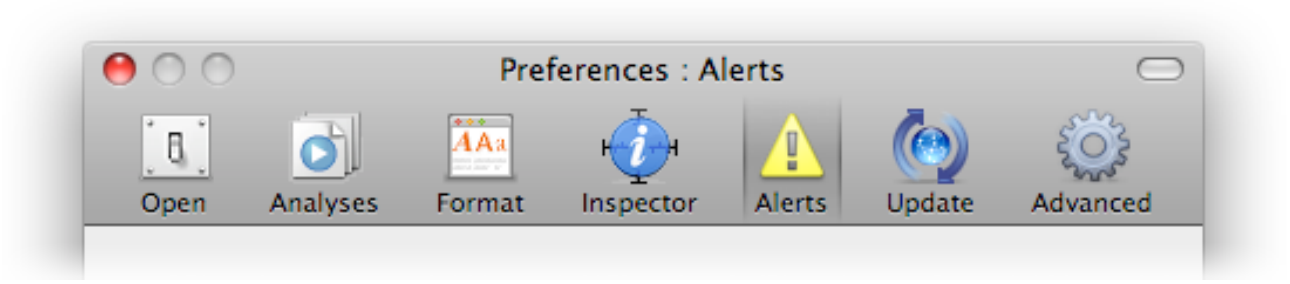
The Inspector panel

The Inspector panel gives you a summary of the data in the selected dataset. As the different types of data that are supported by GenoDive have different characteristics, the Inspector shows different information depending on whether the main window contains genetic data, distances, ecological data or results. Below you find the Inspector window with the information it shows for genetic data.



Setting the preferences

Through the preferences panel it is possible to change a number of basic settings. The preferences window is divided into seven sections:

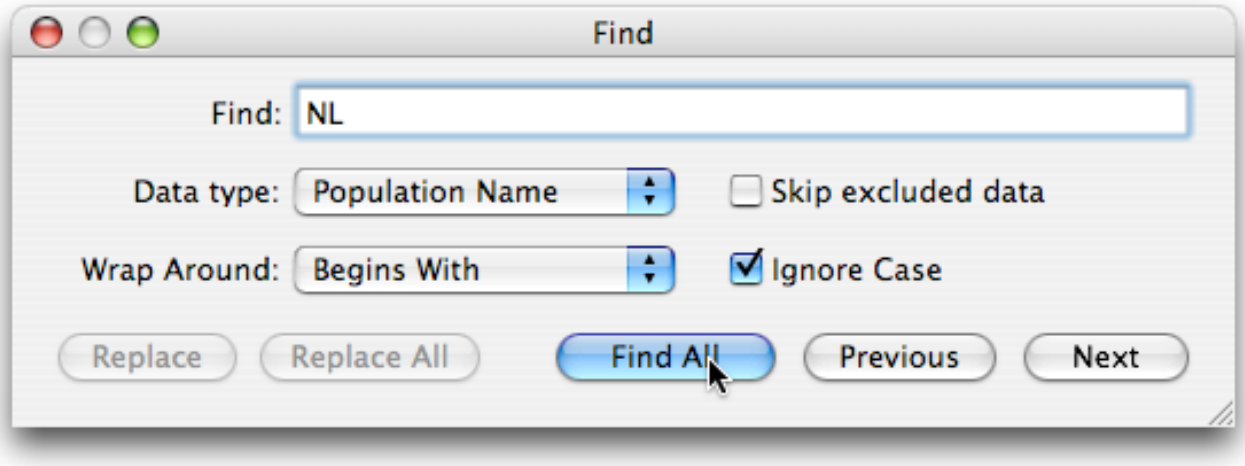


Available preference panes:

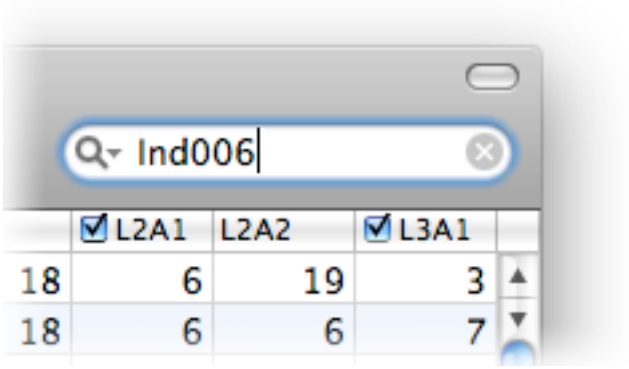
- Open - This allows you to change the behaviour of GenoDive at startup and when files are opened.
- Analyses - This allows you to change the destination of the results and the notifications about the progress of running analyses.
- Format - Here you can set the format of the data as it is displayed both in Text View and Matrix View.
- Inspector - Every self-respecting Mac-application needs some useless eye-candy. In GenoDive it is possible to display the Inspector panel in fourteen different "flavours" (plus a neutral setting).
- Alerts - Here you can change whether you want GenoDive to display many alerts, which is more informative, or not so many, which is less annoying.
- Update - Here you can set whether GenoDive automatically checks for updates on the GenoDive website whenever GenoDive is launched. If an update is available, you will be given the option to download and install it automatically.
- Advanced - Allows you to set threading behaviour and the seed for the random number generator (if you don't understand this, don't change anything here).

Searching

GenoDive has powerful searching tools that help you to quickly find any phrase inside your dataset. It is also possible to select all instances of a certain phrase at once, allowing you to easily include or exclude observations based on certain criteria. For example, if you would like to redo an AMOVA without all the populations from the Netherlands, you could simply search for all population names that begin with "NL" and then choose Exclude Selection from the Select menu.



You can start a search either by using the find-panel that is obtained by choosing Find from the Edit menu, or by using the search box in the toolbar of the document window. When using the search box, the Return key will select the next instance of the search term, while Shift-Return will select all instances of the search term. You can choose the search options from the manu that appears when you click the little triangle next to the magnifying glass in the search box.



It is important to realise that the search options that are available depend on whether you show your data as a matrix or as text.

- When viewing your data as a matrix, you have the options to search for specific data types: individual/observation name, population name, population number, clone, allele or all data. Furthermore, you can specify the "wrap around": contains, begins with, ends with, equals, smaller than, or greater than. You can also set whether you want to skip data that is currently excluded.
- When viewing your data as text, you cannot search for a specific data type, but you can replace occurrences of the search phrase with another phrase.

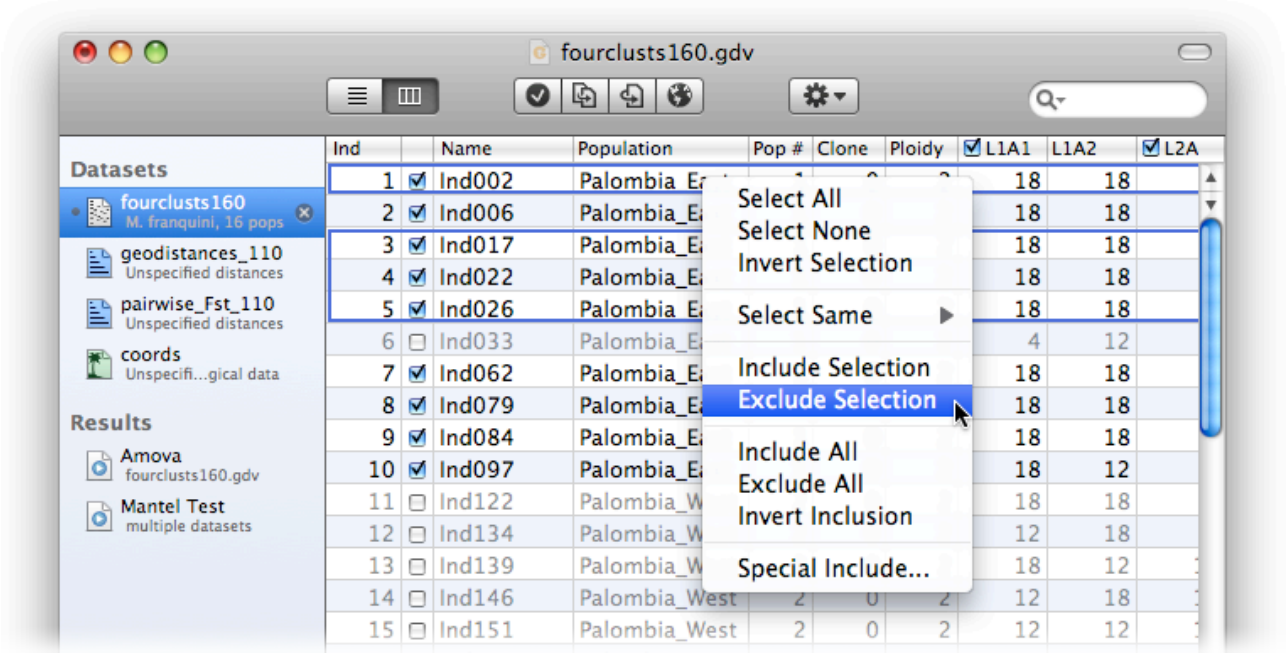
Selecting, including, and excluding data

GenoDive has extensive features that lets you select, include, and exclude data. This is handy if you want to test several different hypotheses with different subsets of your data, or if you want to assess the influence of some outlier loci and individuals on the results of your analysis.

Including and excluding data can be done in several ways: using the Inspector Window and while viewing the data in Matrix View

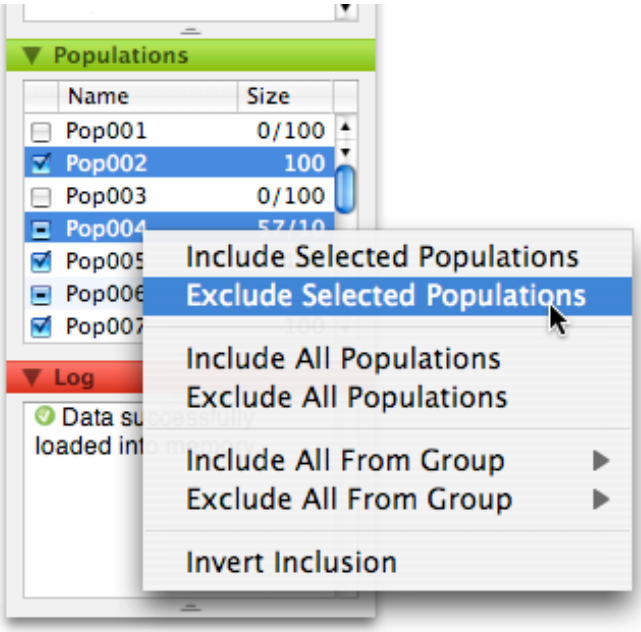
Including and exluding data when viewing the data in Matrix View:

1. If necessary, display the currently selected dataset in Matrix View. Choose View from the Data menu and then As Matrix. Alternatively, you can click select Matrix View from the view selector: 
2. Data can be included or excluded by switching the checkboxes in the second column and in the columnheaders. Data that has been excluded becomes greyed out.
3. To include or exclude several rows at a time, select all rows you want to include and select Include Selection or Exclude Selection from the Select menu, or from the contextual menu that appears when control-clicking the selection.
4. GenoDive allows you to quickly expand the current selection to include all individuals from the same population, clone, ploidy level or population group. To expand the current selection, choose Select Same from the Select menu, and then one of the available categories.
5. It is also possible to make a selection using the Search function. Choose Find from the Edit menu, and then Find..., Type your search term in the Find-panel that appears, set the search options and click Find All. Alternatively, type a search term in the search field in the toolbar and hold down the shift-key while pressing Return. All individuals or observations that contain an occurence of the search term will be selected.



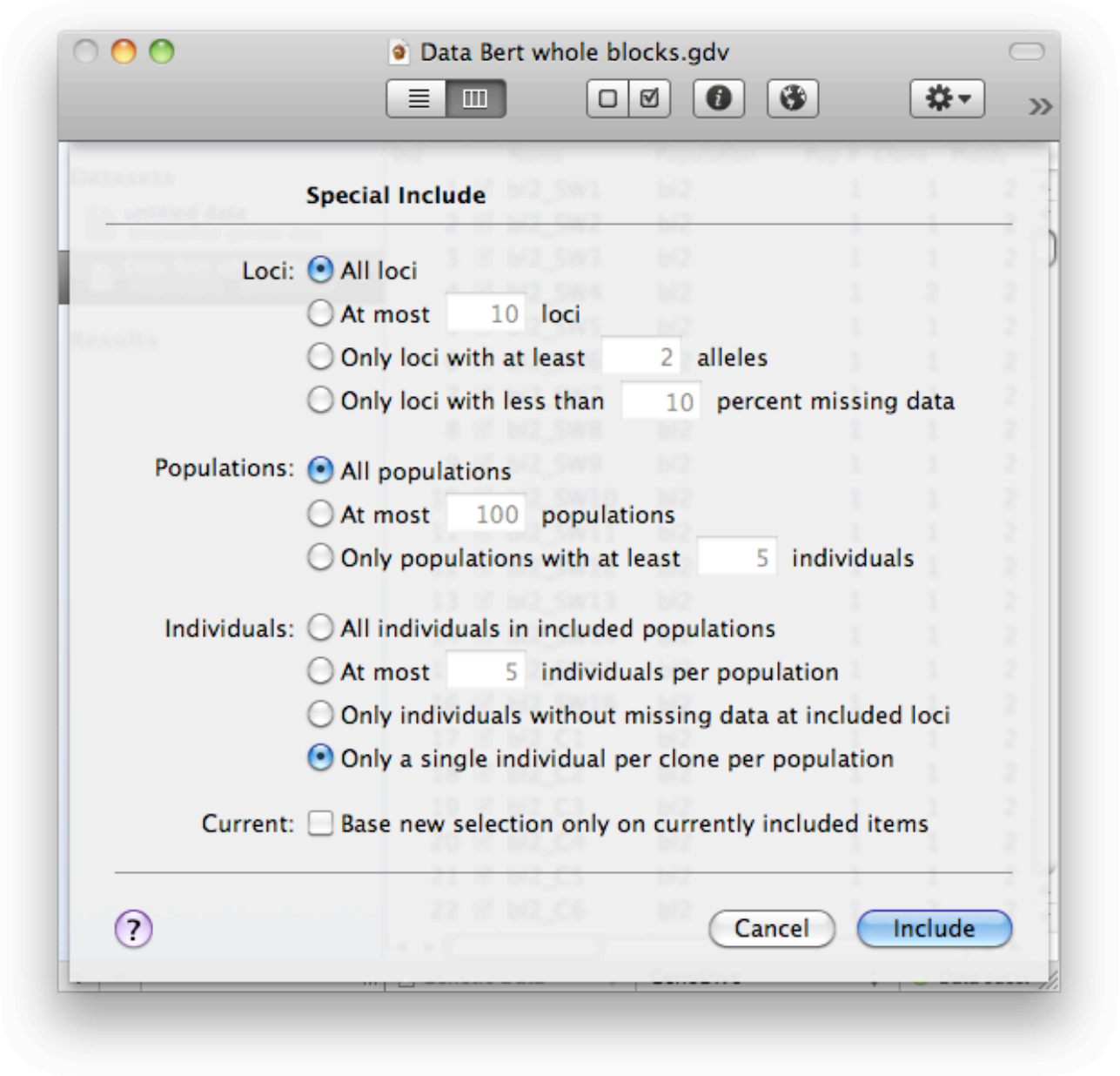
Including and exluding data using the Inspector window:

1. If necessary, make the Inspector window visible by choosing Show Inspector from the Window menu
2. Click the appropriate checkbox in the first column of the table in the Populations, Loci or Variables pane of the Inspector window. In the population pane, a checkmark in the button indicates that all individuals from the population are included, an unchecked button indicates that all individuals have been excluded, and a dash indicates that a only subset of the individuals is included.
3. For more advanced options select multiple rows in a table and Control-click to show a contextual menu with which you can include or exclude multiple items at a time.



Special include options

GenoDive has many different ways to choose which data to include in an analysis. Some special inclusion options are available in the "Special include" dialog.



How to base the included data on certain characteristics:

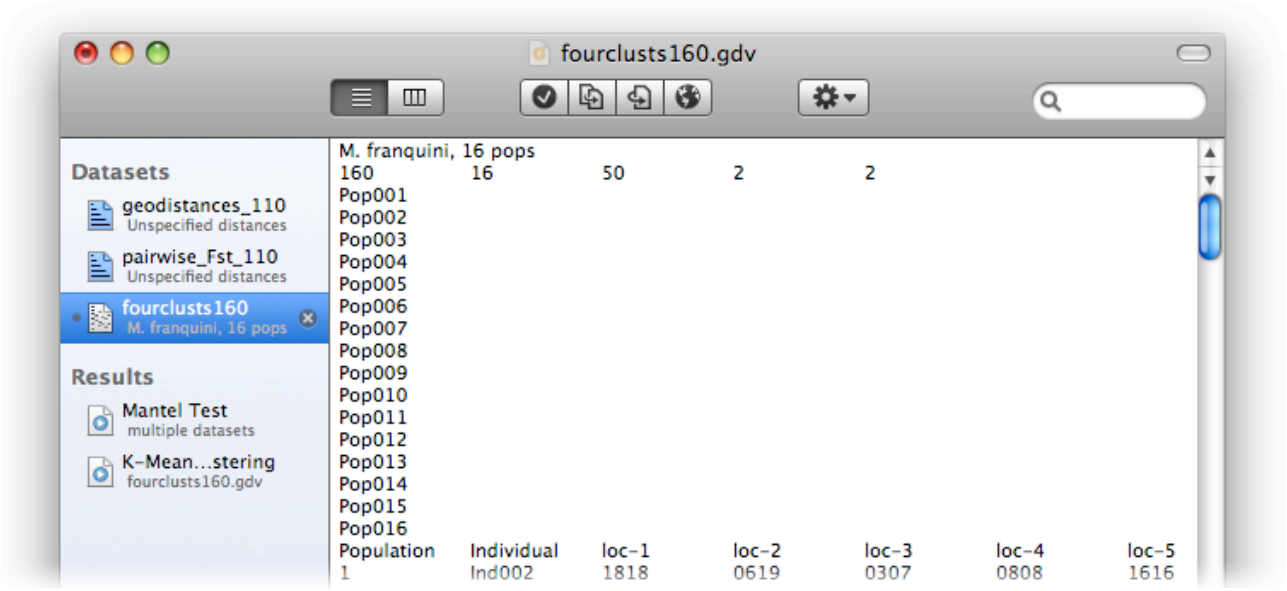
1. Make sure that the selected dataset does not contain results.
2. Choose Special Include... from the Select menu.
3. For marker data you set the inclusion based on locus, population, or individual specific criteria, or a combination of these.
   - For loci, you can choose to include all loci, only a specified number of randomly chosen loci, or only loci with at least a specified number of alleles.
   - For populations, you can choose to include all populations, only a specified number of randomly chosen populations, or only populations with at least a specified number of individuals.
   - For individuals, you can choose to include all individuals within the populations selected above, only a specified number of randomly chosen individuals per population, or only individuals without missing data at the loci selected above (allowing for differences in ploidy). In addition, it is possible to only include a single individual per clone per population. This allow the construction of a "clone-corrected" dataset for organisms with asexual or vegetative reproduction (see also Assign Clones).
4. For distances and ecological data there is only one option:
   - Including only a specified number of randomly chosen observations.
5. With the "Current" option, you can choose whether the new inclusion using the above criteria should be based on the currently included items or, if the switch is off, on the whole dataset.
6. Click Include.

Viewing the data as text and as a matrix

There are two different ways in which you can view your data, either as text or as a matrix. To change between these two modes, choose View from the Data menu and then As Text or As Matrix. Alternatively, you can click the view selector in the toolbar:

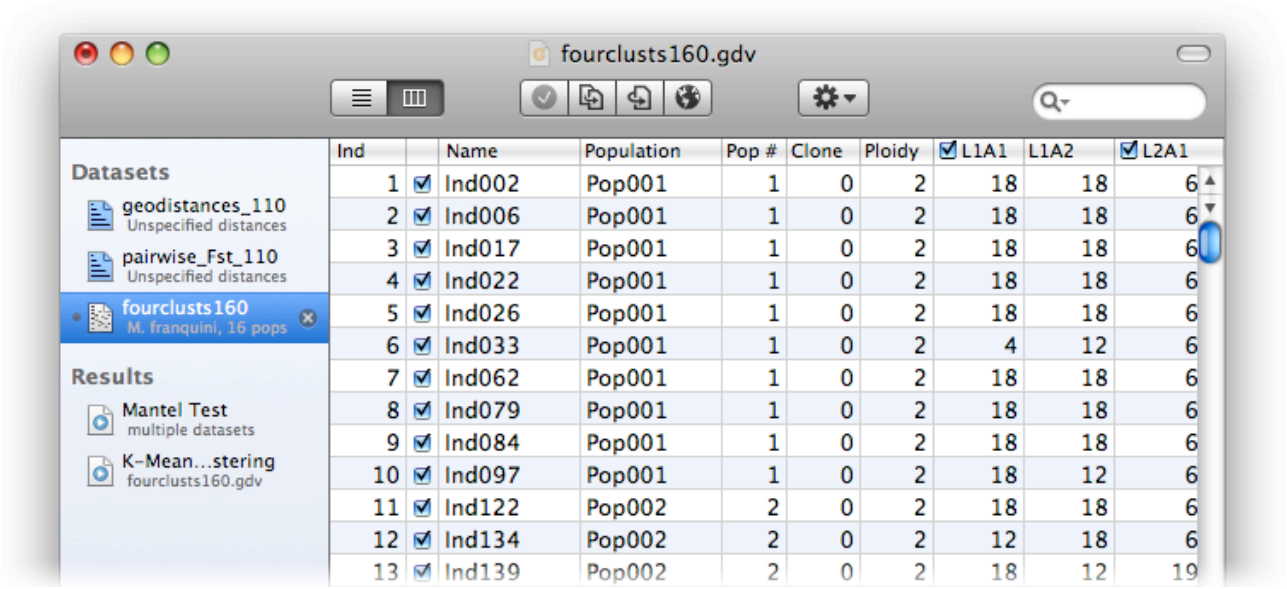
Text View

Text View presents your data in the same format in which it will be saved. This allows you to manually edit the data, for example if there is an error in the format of the data and the data cannot be loaded into memory.



Matrix View

Matrix View gives you a much clearer overview over your data as Text View. Furthermore, Matrix View allows you to easily include or exclude data from the analyses by simply clicking the buttons at the left side of each row or at the headers of the columns. Excluded data is greyed out so that you can immediately see which data are included and which are excluded. It is possible to correct your data by just double clicking a cell and editing the value.



Ind		Name	Population	Pop #	Clone	Ploidy	L1A1	L1A2	L2A1
1	☑	Ind002	Pop001	1	0	2	18	18	6
2	☑	Ind006	Pop001	1	0	2	18	18	6
3	☑	Ind017	Pop001	1	0	2	18	18	6
4	☑	Ind022	Pop001	1	0	2	18	18	6
5	☑	Ind026	Pop001	1	0	2	18	18	6
6	☑	Ind033	Pop001	1	0	2	4	12	6
7	☑	Ind062	Pop001	1	0	2	18	18	6
8	☑	Ind079	Pop001	1	0	2	18	18	6
9	☑	Ind084	Pop001	1	0	2	18	18	6
10	☑	Ind097	Pop001	1	0	2	18	12	6
11	☑	Ind122	Pop002	2	0	2	18	18	6
12	☑	Ind134	Pop002	2	0	2	12	18	6
13	☑	Ind139	Pop002	2	0	2	18	12	19

Note that if you have changed your data in Matrix View, the data in Text View will be overwritten to match the updated content. This may cause your data to appear slightly different in format.

Multithreaded analyses

GenoDive is fully multithreaded. Threads provide a way for a program to split itself into several different simultaneously running tasks. One advantage of multi-threading is that a program can have a lengthy calculation on one thread while another thread is used to keep the user-interface responsive. Another advantage is that the program can have multiple calculations running at the same time: for example, you can quickly perform a calculation of pairwise genetic distances on one document while a lengthy k-means clustering is running for another. The Dock-icon of GenoDive shows progressbars for all running processes, so that you can keep track of the status of your analyses.

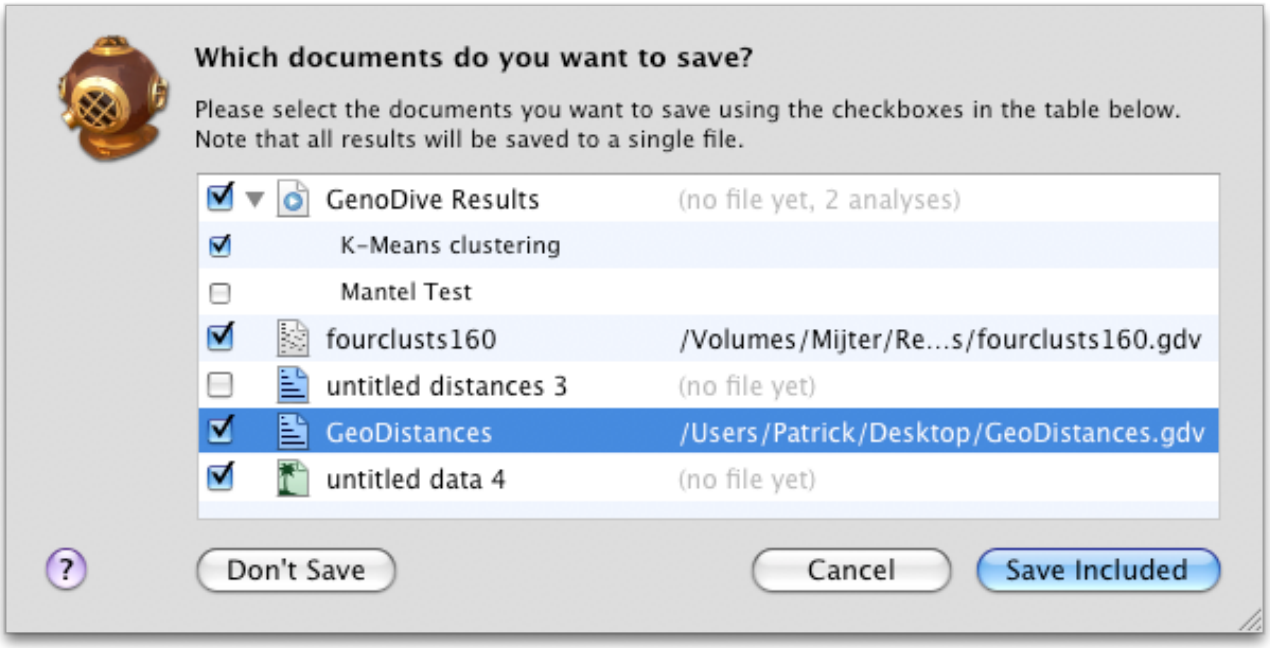
All Macs sold nowadays have processors with multiple cores. Most models have a single dual-core processor, but the top-of-the-line Mac Pro actually has two processors with four cores each. If you have such a computer with multiple processor cores, multithreading provides an additional benefit, as GenoDive automatically distributes length calculations over all available processor cores. So if you have a dual-core iMac (such as mine) and you are performing a test with 1000 permutations, each core will do 500 so that the analysis is finished in half the time it would have taken otherwise; if you are lucky and have an 8-core Mac Pro, each core will do 125 permutations.

In the "Advanced" section of the Preferences-panel it is possible to specify whether you want GenoDive to automatically detect the number of cores, or manually set the maximum number of threads that is used for a single analysis. The latter option is handy if you want to keep some processing power available for other programs that you may be running (even though Mac OS X already does a very good job of distributing the tasks over the available processors).

Saving multiple datasets or results

The GenoDive interface allows you to have many datasets and results open at the same time. When you quit GenoDive, the program will present a sheet with a list of all open datasets and results that have unsaved changes. You can select which datsets and results you want to save by switching the checkboxes in the table on or off.

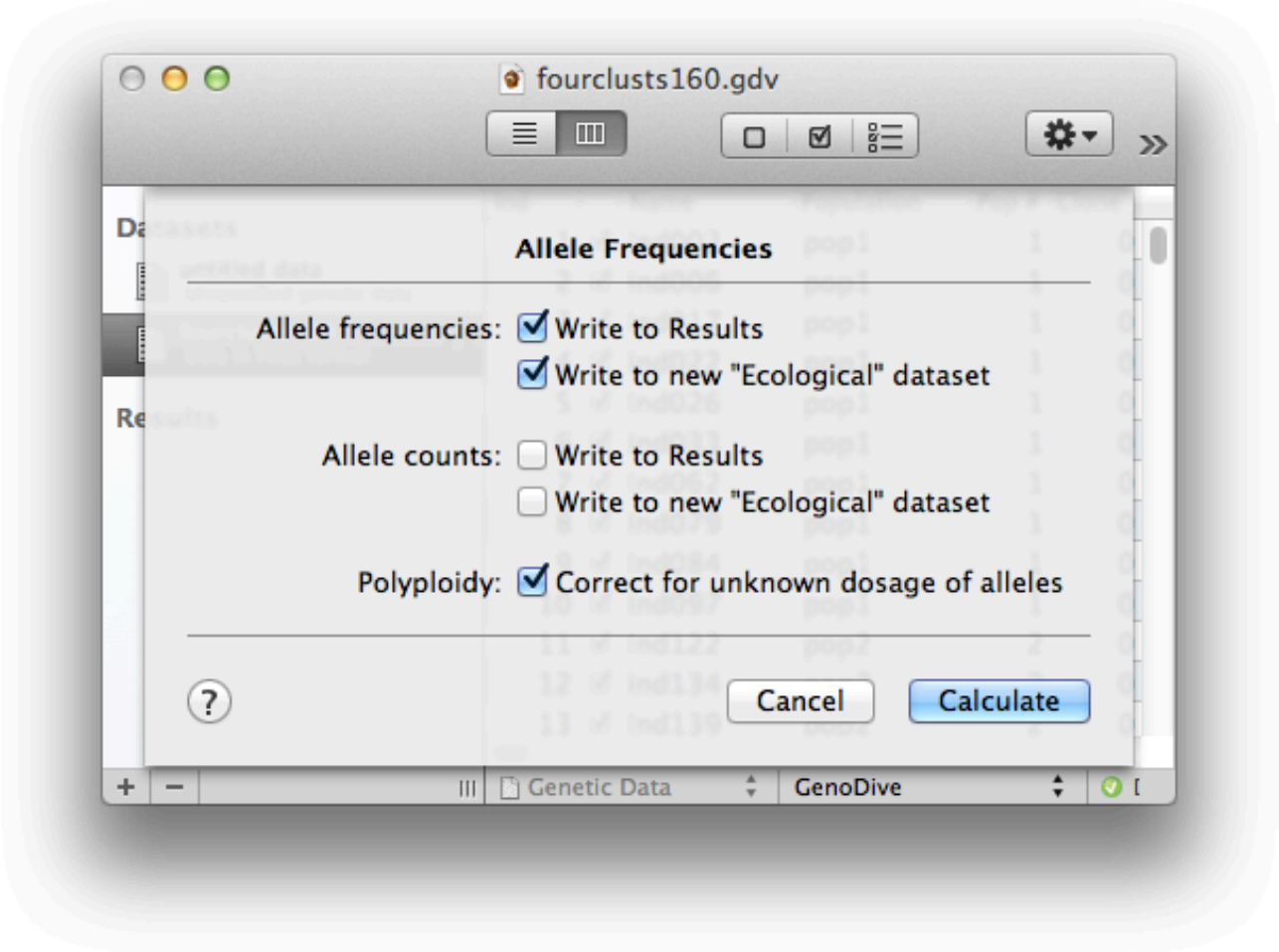
GenoDive handles results differently than datasets: all results will be saved to a single file, while each dataset is save to its own file. To determine which results exactly will be saved to the file, you can click the disclosure button next to the "GenoDive Results" item and then use the checkboxes to include or exclude the results of certain analyses.



Available analyses

Allele frequencies

GenoDive can calculate both allele frequencies and allele counts. The allele count is the number of times an allele is found in a population, the allele frequency is the allele count divided by the sum of all allele counts. For allele frequencies, missing data is not included; hence the sum of the frequencies of all non-missing alleles equals one.



To calculate allele frequencies:

1. Make sure the selected dataset contains genetic marker data
2. Choose Allele Frequencies... from the Analysis menu
3. Select the desired output options for the allele frequencies and the allele counts. The output can be written to the Results window, but also to new ecological datafiles. The latter option is useful if you want to calculate non-genetic distances between populations, based on the allele frequencies.
4. For polyploid data it is possible to correct the allele frequencies for the unknown dosage of alleles.
5. Click Calculate.

Analysis of Molecular Variance

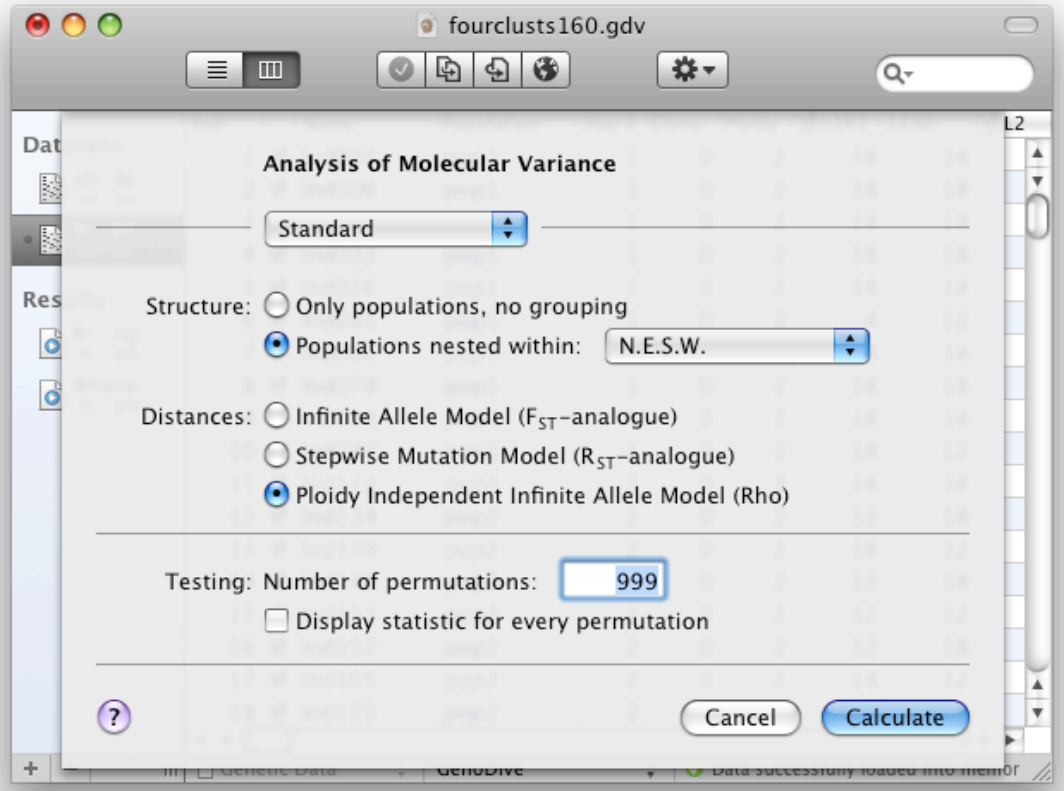
Analysis of Molecular Variance (AMOVA, Excoffier, 1992, Michalakis & Excoffier, 1996) is a method for analysing population structure based on an analysis of variance which is performed on a matrix of squared Euclidean distances. The main strength of AMOVA is its flexibility, as it can analyse several different types of population structure with different numbers of hierarchical levels.

Depending on how the squared Euclidean distances are calculated, different F_{st}-analogues can be calculated. When an Infinite Allele Model is assumed, the F-statistics correspond to those defined by Weir & Cockerham (1984). When a Stepwise Mutation Model is assumed, the calculated statistics correspond to Slatkin's (1995) R_{st}. In addition to these two well-known statistics, GenoDive can calculate rho, which is an F_{st}-analogue that is independent of the ploidy level and breeding system (Ronfort et al. 1998).

Next to calculating the usual F_{st}, F_{sc} and F_{ct} values, GenoDive also calculates standardized versions of these parameters (Hedrick, 2005; Meirmans, 2006).

Warning about this Beta version:
Some important features are still missing.

- AMOVA does not yet allow for missing values. The current implementation fills in all missing data with randomly drawn alleles based on the overall allele frequencies (see Fill in Missing Data). This means that when your data have a lot of missing values, the estimates of differentiation may have a slight downwards bias. Furthermore, the results can differ slightly between different runs of the AMOVA.
- AMOVA does not yet take singletons, so you should not have any populations with only one individual, or any groups of populations with only a single individual. In principle, this also counts for the within-individual level (haploid data), but that is avoided by converting any fully haploid datasets to diploid homozygotes. This means that if you have haploid data, you should ignore the values of F_{is} (which are 1.0 for all loci).



How to calculate an Amova:

1. Make sure that the selected dataset contains genetic data
2. Select Amova... from the Analysis menu.
3. Choose Standard or Advanced from the topmost pop-up menu.
4. Select the population structure that you would like to test.
5. In Standard view, there are two possible population structures: either only individuals nested within populations, or individuals nested within populations, nested within groups of populations. A within-individual level (F_{is} / F_{it}) is always included.
6. In Advanced view, it is possible to specify the population structure with much more flexibility. Using the four popup menus, it is possible to specify up to four hierarchical levels not only including individuals, populations and population groups, but also ploidy levels and clones.
7. Choose the mutation model that is assumed in the analysis, either an Infinite Allele Model, a Stepwise Mutation Model, or a ploidy independent Infinite Allele Model, respectively resulting in F_{st}, R_{st}, or Rho estimates. Under Advanced view it is also possible to use a custom distance matrix from an open file for the calculations; this matrix should contain squared Euclidean distances.
8. Type in the number of permutations to test the significance.
9. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations as it will make the Result file overly large.
10. Click Calculate.

Assign Clones

GenoDive has an algorithm for assigning genotypic identity to individuals, using data from most types of genetic markers. Identification of genotypes ("clones") is especially important in studies of parthenogenetically reproducing organisms, or organisms with vegetative reproduction. In these cases different individuals (sometimes referred to as "ramets") can have the identical multilocus genotypes (referred to as "genets"). The technique can also be used in studies with noninvasive sampling, to check whether some individuals have been sampled twice. Assigning genotypes is a simple task, but for large datasets it is tedious to do it by hand. This functionality was previously included in a separate program called GenoType (Meirmans & Van Tienderen, 2005).

The algorithm works by first calculating a matrix of genetic distances, and then choosing a threshold distance. If the distance between a pair of individuals is below that threshold, they are deemed to belong to the same clone.

In the same analysis as the assignment of clones, it is possible to test for clonal population structure in the data. This test works by randomising the alleles over individuals and comparing the clonal diversity for the randomised dataset with that of the original dataset.

How to assign individuals to clonal lineages:

1. Make sure the selected dataset contains genetic data.
2. Choose Assign Clones from the Analysis menu
3. You will be shown a series a panels that will guide you through the process in three or four steps. See the dedicated help-pages for more information about each step.
4. First step: calculate a matrix of genetic distances.
5. Second step: choose a threshold distance below which individuals are considered clonemates.
6. Third step: test for clonal population structure. This step is not available if an external distance matrix is used.
7. Fourth step: output options.
8. Click Calculate to perform the genotype assignment and, if applicable, the test for clonal structure.

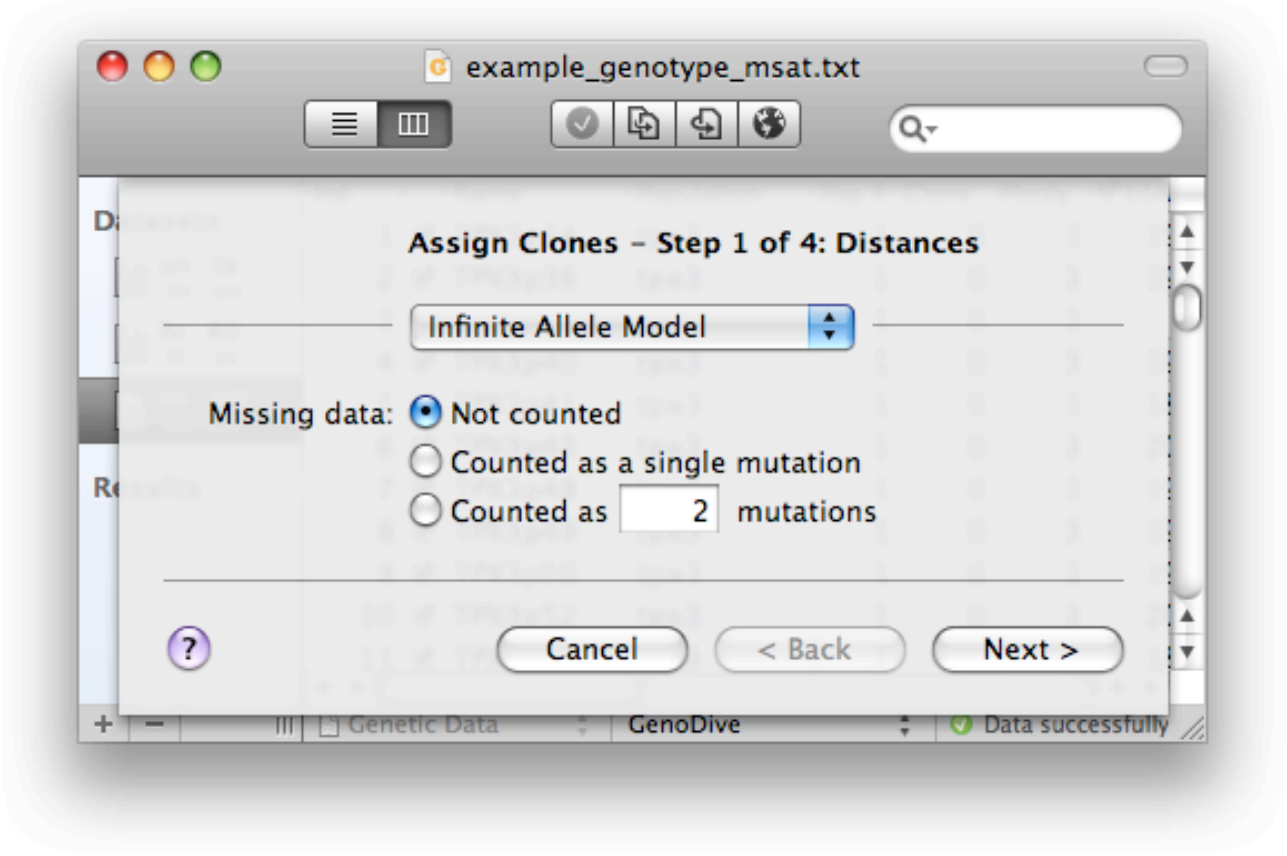
Assign Clones - Step 1: Distances

The clone assignment algorithm works by first calculating a matrix of genetic distances between individuals, using a distance measure that assumes asexual reproduction. Two different mutation models can be used: an Infinite Allele Models and a Stepwise Mutation Model. Alternatively, you can choose to use a distance matrix from a separate file.

Infinite Allele Model

Under this distance index, an infinite allele model (IAM) is assumed meaning that it takes only one mutation step to get from a certain allelic state to any other. This mutation model provides a good approximation for almost all molecular data besides microsatellites, such as allozymes, RAPD's, and AFLP's. The distance measure simply consists of the number of mutations that are needed to transform the genotype of one individual into the genotype of the other, summed over all loci. For haploid data without missing values, this distance measure is equivalent to the Manhattan distance.

The program can handle missing data in three different ways: they are either discarded (first option), equated to one mutation (second option) or equated to a user-specified number of mutations (third option).



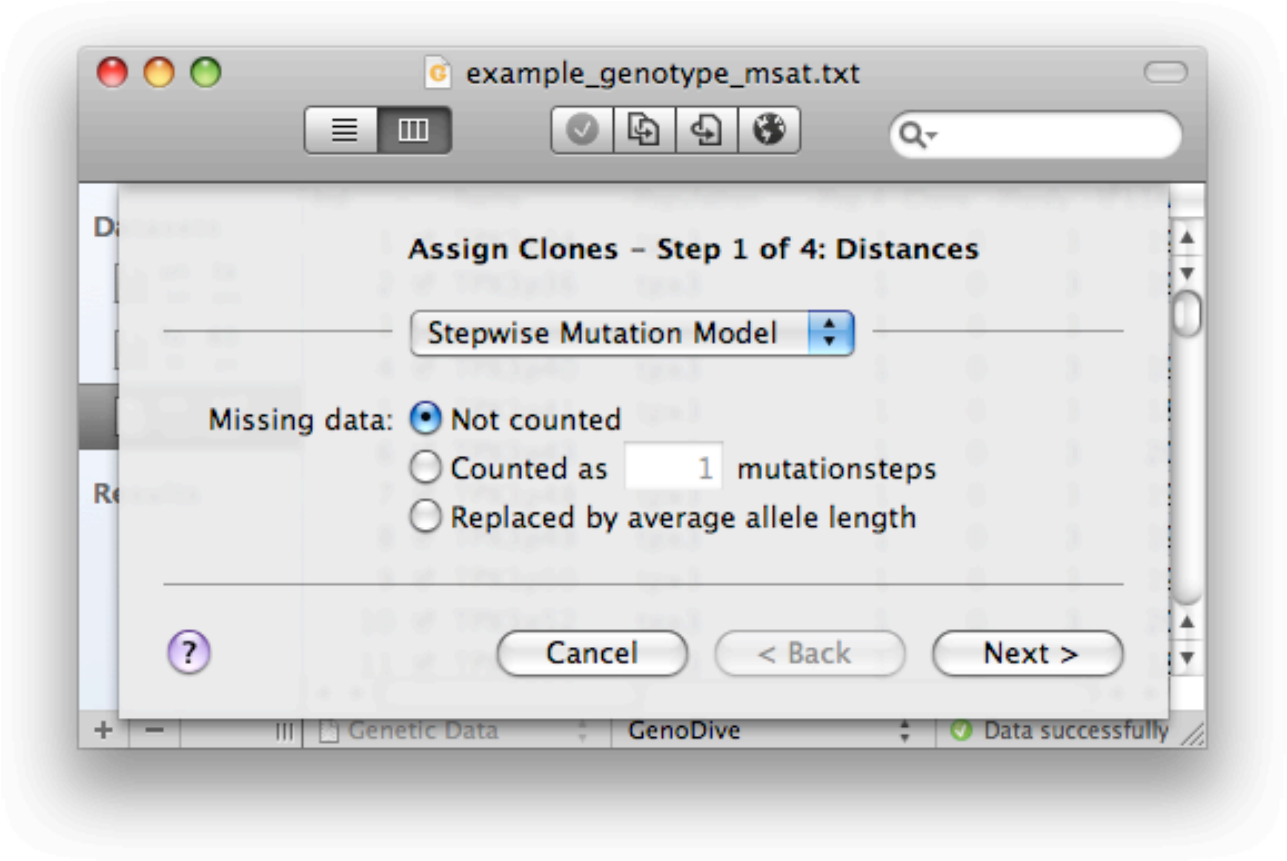
Stepwise Mutation Model

This distance measure is specifically meant for microsatellite data. A stepwise mutation model (SMM) is assumed, meaning that alleles that differ only a few repeats in length are thought to be of more recent common ancestry than alleles that differ a lot of repeats in length. For proper calculation, this index requires that the alleles are given as the number of repeats rather than the length of the fragment. However, also fragment length data can be given, which is handy for microsatellite loci containing imperfect repeats, though this stretches the idea of the SMM a bit.

A straightforward distance measure under the SMM can be calculated by simply calculating the smallest number of mutation steps that is needed to transform the genotype of one individual into

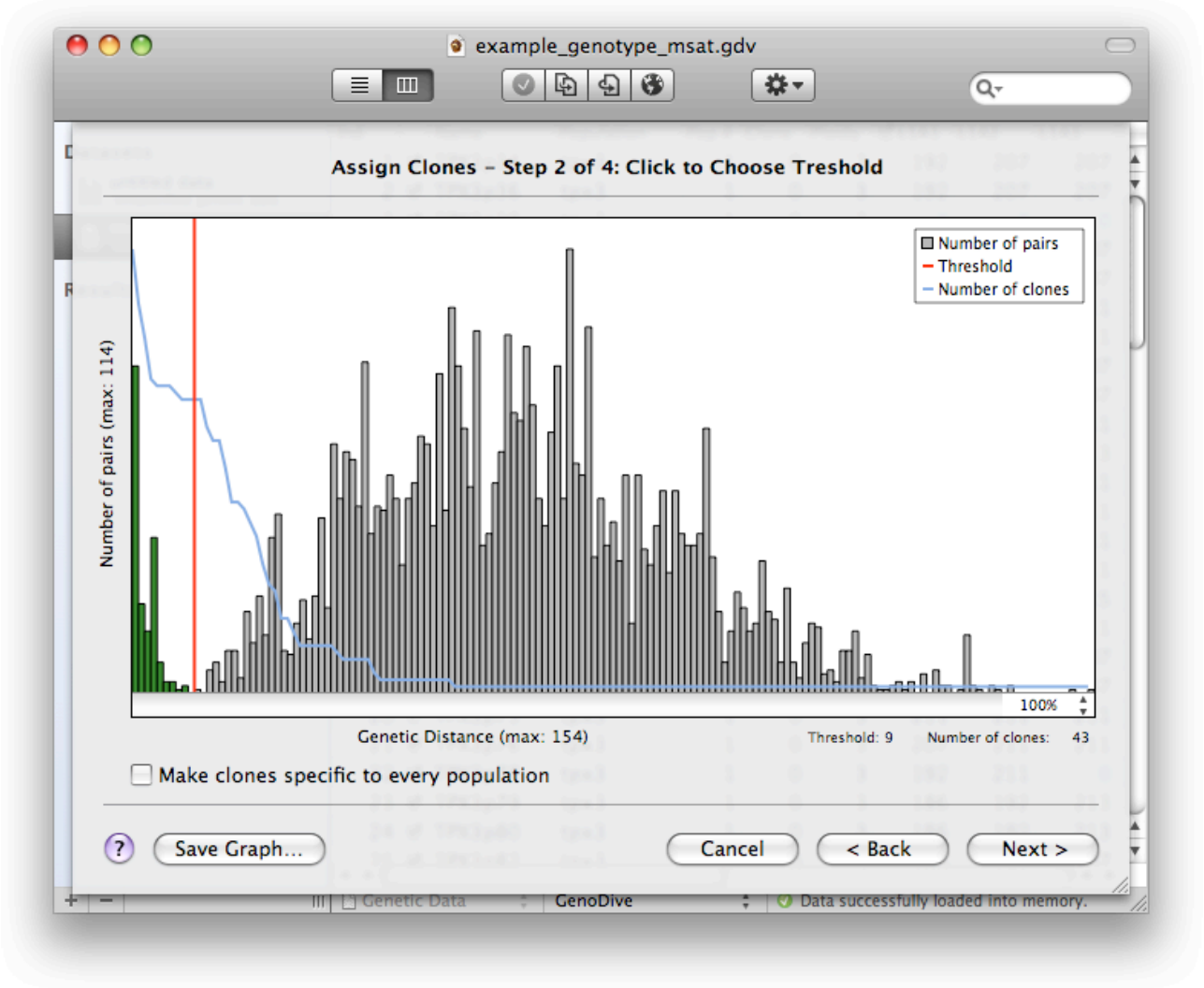
the genotype of the other, summed over all loci. Like the simple IAM-distance, the IAM distance is equivalent to the Manhattan distance for haploid data without missing values.

The program can handle missing data in three different ways: they are either discarded (first option), equated to user-defined number of mutation steps (second option), or substituted by the average allele length for a locus (third option).



Assign Clones - Step 2: Threshold

Individuals are assigned to clones using a distance "threshold". This threshold indicates the maximum distance that is allowed between two individuals to still be clonemates with the "same" multilocus genotype. Scoring errors and mutations may cause individuals from the same clonal lineage (clonemates) to have a pairwise distance larger than zero. To set a limit to this you can draw a threshold for the amount of scoring error or mutation you allow (Rogstad et al, 2002, Meirmans & Van Tienderen, 2005). Choosing this threshold too low inflates the estimates of clonal diversity, choosing this value too high deflates the diversity estimates, so choosing a right threshold is important. Douhovnikoff & Dodd (2003) recently proposed a method to objectively choose a threshold value, based on the means and standard deviations of the two peaks in a bimodal histogram. This method is however not implemented in GenoDive as it is difficult to implement on data from natural populations. Douhovnikoff and Dodd however assumed the determination of the threshold from a dataset of known clones and siblings, and afterwards used this value for natural populations. As most studies on clonal diversity will probably be carried out without the possibility of such prior testing, I did not include this method in the program, but rather propose to test hypotheses concerning clonal diversity using several different threshold-values, to see the effect of the scoring errors and mutations on the used statistics.



The threshold dialog shows a histogram of the frequency distribution of the distances, with the genetic distance on the x-axis and the number of pairs on the y-axis. The current threshold is indicated by the vertical red line, the bars before the threshold are shown in a contrasting green colour to distinguish them from the bars after the threshold line. You can change the threshold by clicking at the desired position in the graph. If there are too many bars in the graph, you can

change the zoom using the popup menu in the lower right corner of the graph; the zoom will only work horizontally. The graph also shows the number of clones for every threshold as a blue line. The number of clones that can be distinguished using the current threshold is also show under the graph.

One problem that may occur when assigning clones is that individuals with clearly different genotypes are assigned to the same clone. The problem is a theoretical one rather than a bug in the program, and is a corollary of working with a non-zero threshold, but it also occurs with a treshold of zero when there are missing data. The problem is best explained with a little example:

```
ind1 0102 0101 0202
ind2 0102 0000 0202
ind3 0102 0202 0202
```

These are three individuals with three loci, where individual two has missing data at the second locus. If we now calculate the pairwise distance between these individuals, we get a matrix that looks like this:

```
dist ind1 ind2 ind3
ind1 0.00
ind2 0.00 0.00
ind3 2.00 0.00 0.00
```

If we now do the genotyping with threshold 0, we see that ind1 and ind2 are the same genotype. Furthermore, ind2 and ind3 are also the same genotype. From this it follows logically that ind1 and ind3 are also the same genotype, even though they differ from eachother! If you have a large dataset or quite some missing data, this problems gets more and more expanded, until there may only be a few genotypes left.

One workaround is to do the genotyping separately for every population, for this switch on the button labeled "Make clones specific to each population". However, this may give problems when certain genotypes are present in multiple populations.

Assign Clones - Step 3: Test for Clonal Structure

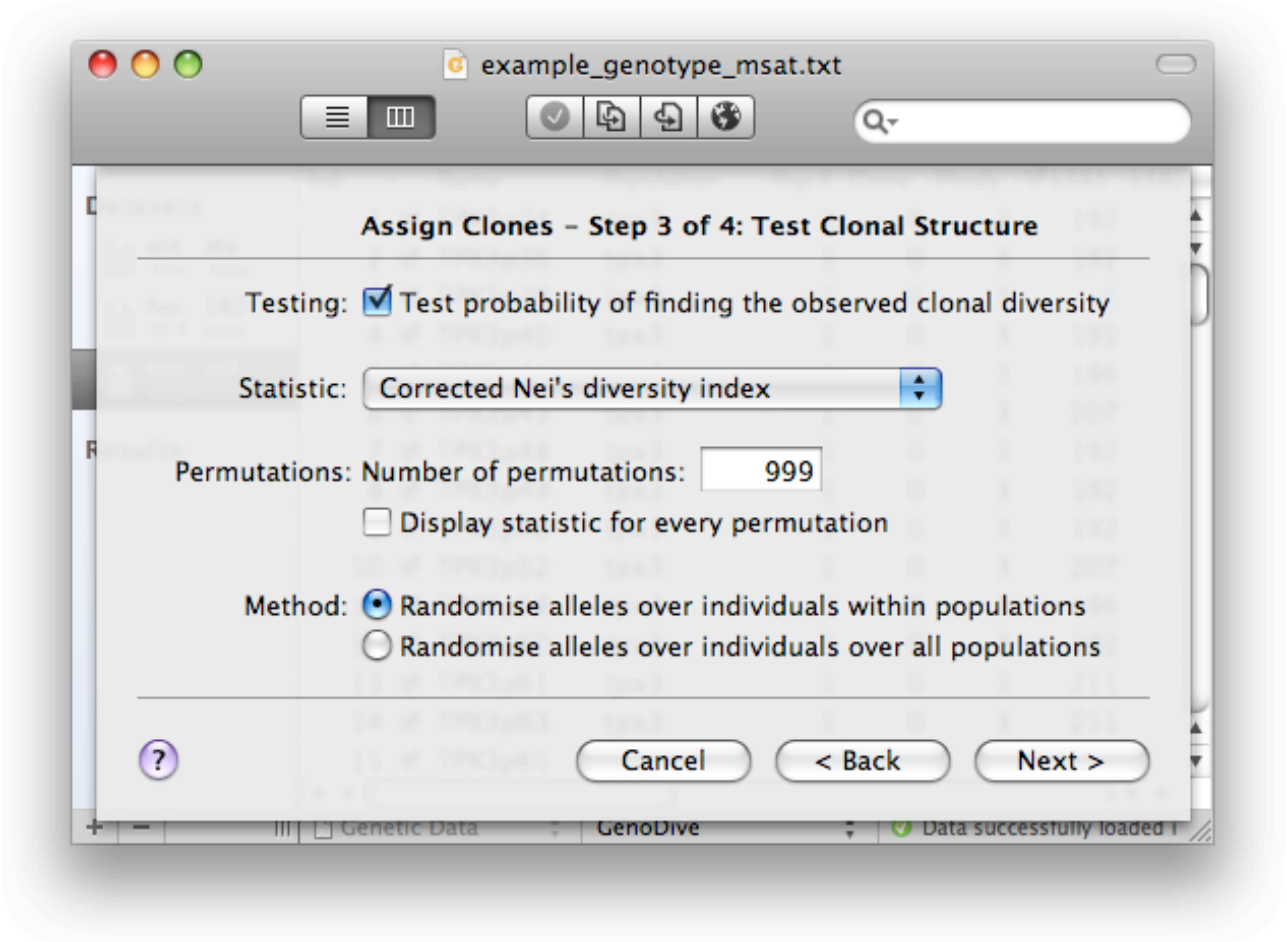
Finding duplicated multilocus genotypes does not necessarily imply asexual reproduction, as identical genotypes can also be produced by random mating, especially when the amount of genetic variation is low or few loci are used. It is therefore necessary to test whether the sample shows a clonal population structure. This test is not available if an external distance matrix is used.

Diversity based test

The clone assignment implemented in GenoDive includes such a test for clonal population structure, based on the concept of clonal diversity. Under asexual reproduction, the clonal diversity is lower than under sexual reproduction. Therefore, we can test the null-hypothesis that the observed clonal diversity is due to sexual reproduction by randomising alleles over individuals and comparing the observed clonal diversity with that of the randomised dataset (Gomez & Carvalho, 2000).

This test allows for the inclusion of the threshold concept used for the clone assignment. For every permuted dataset, the distance matrix is recalculated, clones are assigned using the same threshold as for the original dataset, and a diversity index is calculated using the assigned clones. In this way, the test allows for mutations and genotyping errors within clones. As the test depends on the calculation of distances, it cannot be performed when a custom distance matrix is selected in Step 1 of the clone assignment.

Note that there are several processes other than asexual reproduction that can lead to identical genotypes, such as inbreeding (especially self-fertilisation) and undetected population structure (Halkett et al., 2005).

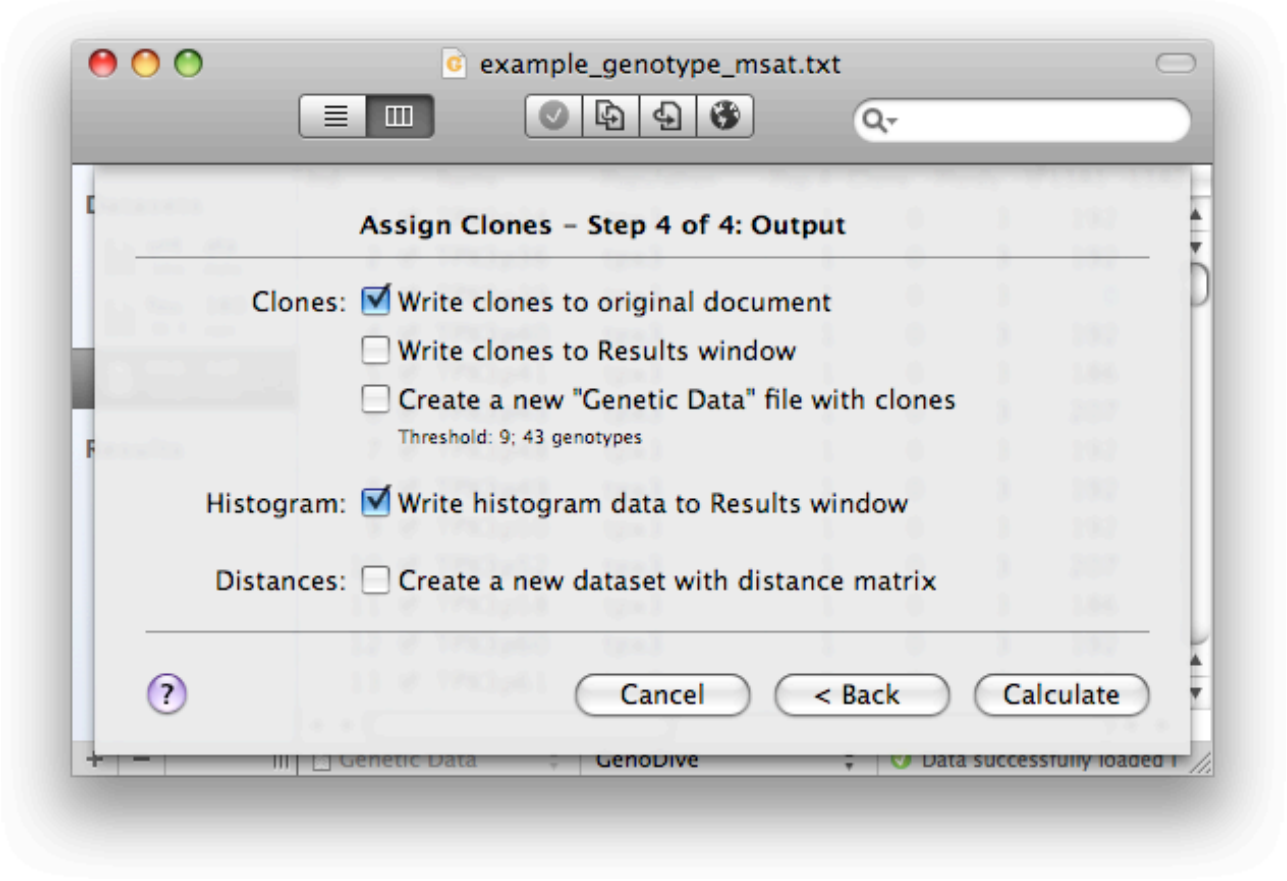


Available options for testing clonal population structure:

1. To perform the test for clonal population structure switch on the button labeled “Test probability of finding the observed clonal diversity”
2. Choose a test-statistic. Seven different diversity indices are available for use a a test statistic, the one with the best statistical behaviour is probably Nei's corrected diversity index. For more information on the available indices, see clonal diversity
3. Enter the number of permutations. Note that the test will take very long for large datasets. An estimate for the required time is given, this estimate is a good approximation for large datasets, but upwardly biased for very small datasets.
4. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations as it will make the Result file overly large.
5. Choose the method of randomisation, either restricting the randomisations within populations or randomising the alleles over all individuals over all populations. The former method should be used if you selected the "Make clones specific to every population" option in Step 2 of the clone assignment.
6. The test will be performed only after you have completed Step 4 of the clone assignment.

Assign Clones - Last Step: Output Options

The last step of the clone assignment are determining the output options. It is possible to save the assigned clones in several different ways, as well as save the histogram data and the used distance matrix. There are no additional output options for the test for clonal population structure, these are all set in Step 3.

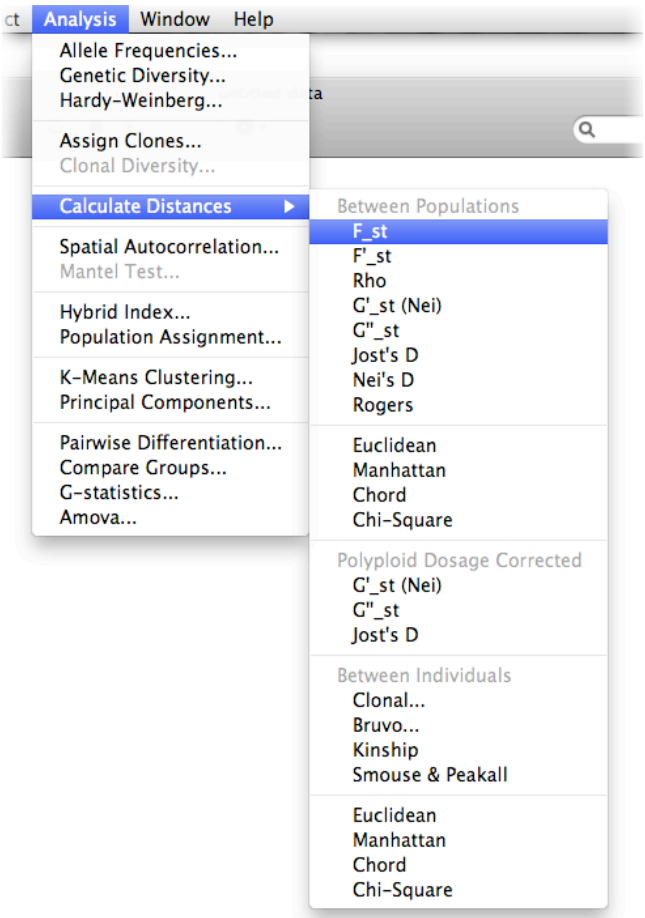


Available options:

1. "Write clones to original document". This option will add the clone information to the genetic data in the current document. The clones will be shown in the "clone" column when the data is viewed as a matrix. The clone information will only be saved together with the data if the document is in GenoDive format.
2. "Write clones to Results window". Per individual, the name and clone will be written to the Results window.
3. "Create a new Genetic Data file with clones. A new file will be created with only a single haploid locus, containing the clone information. This allows you to save the clone information in another format than the GenoDive format and perform some additional analyses in which the clones are treated as allelic data.
4. "Write histogram data to Results window". For every possible distance, the number of pairs of individuals is given, so that you can recreate the histogram in a specialised graphing or spreadsheet program such as Microsoft Excel. In addition, for every distance the number of clones is given if that distance would be used as a threshold.
5. "Create a new window with distance matrix". The used distance matrix is output to a new window.

Calculate Distances

GenoDive can calculate multiple indices of pairwise distances or similarities. The actual indices that can be calculated depend on the type of data of the selected dataset, as some indices cannot be calculated for ecological data, while others cannot be calculated for genetic data. For genetic data, distances can be calculated either between populations or between individuals. See Genetic Distances and Non-Genetic Distances for more information about the available indices for the two different types of data.



To calculate a matrix of pairwise distances:

1. Make sure the selected dataset contains either genetic marker data or ecological data
2. For ecological data it is important to first include only those variables that you want to use for the calculation of distances, otherwise all variables wilol be used (see: including, and excluding data).
3. Choose Calculate Distances from the Analysis menu; the available distance indices for the current data type are shown in the submenu.
4. Select the desired distance index.
5. Some indices, marked by the ... after the name, require some additional input, see Genetic Distances and Non-Genetic Distances.
6. When the calculations are finished, which usually does not take very long, a new document will appear containing the distances.

Genetic Distances

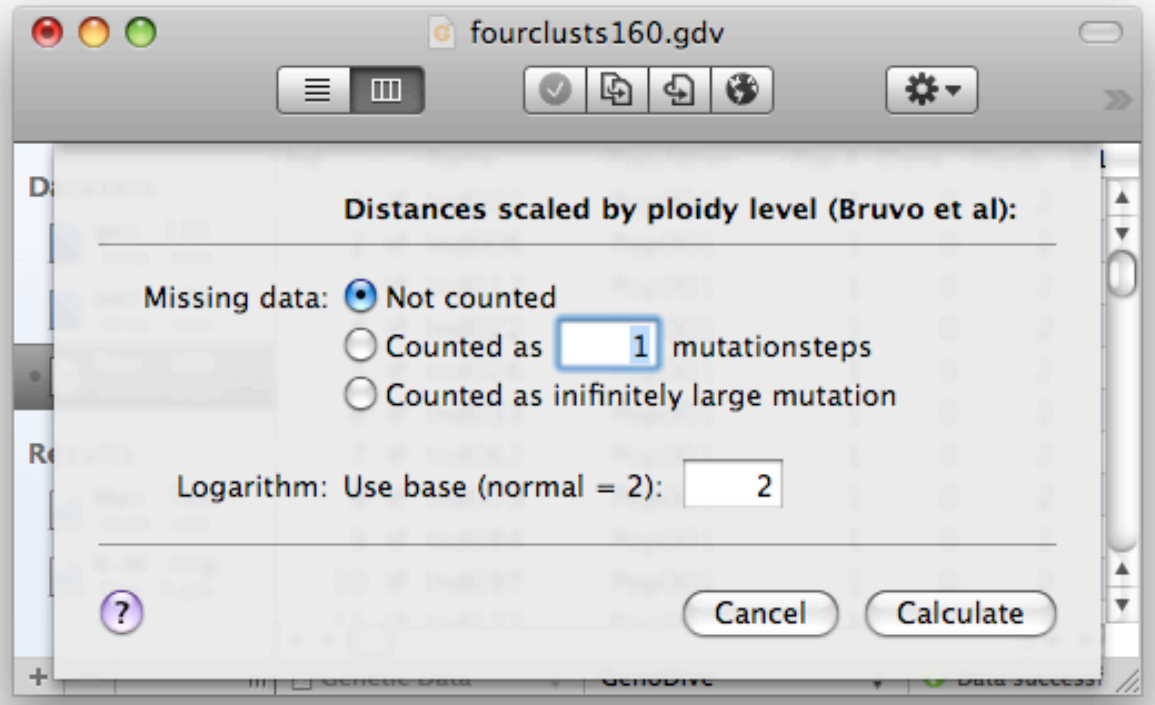
GenoDive can calculate several indices of genetic distance, both between pairs of individuals and between pairs of populations. In addition, there are some distance indices that are not strictly genetic as they are not based on a genetic model, but that may be useful anyway. For a general overview of distance and similarity indices, see Legendre & Legendre (1998)

Distances between populations:

- F_{st} – This distance is the F_{st} (Phi_{st}) statistic resulting from an Analysis of Molecular Variance performed between each pair of populations (AMOVA, Excoffier, 1992, Michalakis & Excoffier, 1996). This distance is the same as the one used for the Pairwise Differentiation test.
- F'_{st} (standardised) – A standardised measure of population differentiation, estimated using an AMOVA (Meirmans, 2006). This is F_{st} relative to the maximum value possible given the observed amount of within-population diversity. This distance is suited for comparison between studies using different types of markers or between organisms with different effective population sizes (Hedrick, 2005).
- Rho – An F_{st} analogue that is independent of the ploidy level and, for polyploids, the rate of double reduction (Ronfort et al., 1998).
- G'_{st} (Nei) – The classic G_{st} calculated by comparing the heterozygosity within and between populations, with a correction for a bias that stems from sampling a limited number of populations (Nei, 1987)
- G''_{st} – A standardised measure of population differentiation, based on Nei's G_{st}. This is Hedrick's (2005) standardised measure of differentiation, but with a correction for a bias that stems from sampling a limited number of populations (Meirmans & Hedrick, 2011).
- Jost's D – Index of population differentiation that is independent of the amount of within-population diversity (H_s) (Jost, 2008).
- Nei's D – This is the standard genetic distance from Nei (1978), which includes a bias-correction for samll sample sizes. Note that this distance gives "inf" when the two populations do not share any alleles, in which case makes this distance becomes unsuitable for further testing (e.g. using a Mantel test).
- Rogers – This is the genetic distance from Rogers (1972). Note that this distance index does not have a correction for bias due to small sample sizes.

- Euclidean – The standard Euclidean distance between populations, calculated using the within-population allele frequencies.
- Manhattan – The standard Manhattan distance between populations, calculated using the within-population allele frequencies.
- Chord – The standard Chord distance between populations, calculated using the within-population allele frequencies. This one is quite often used in genetics following the work of Cavalli-Sforza & Bodmer (1971).
- Chi-square – The standard Chi-square distance between populations, calculated using the within-population allele frequencies.

For polyploid data it is possible to calculate G'_{st} (Nei), G''_{st}, and Jost's D using allele frequencies that are corrected for the unknown dosage of alleles.



Distances between individuals:

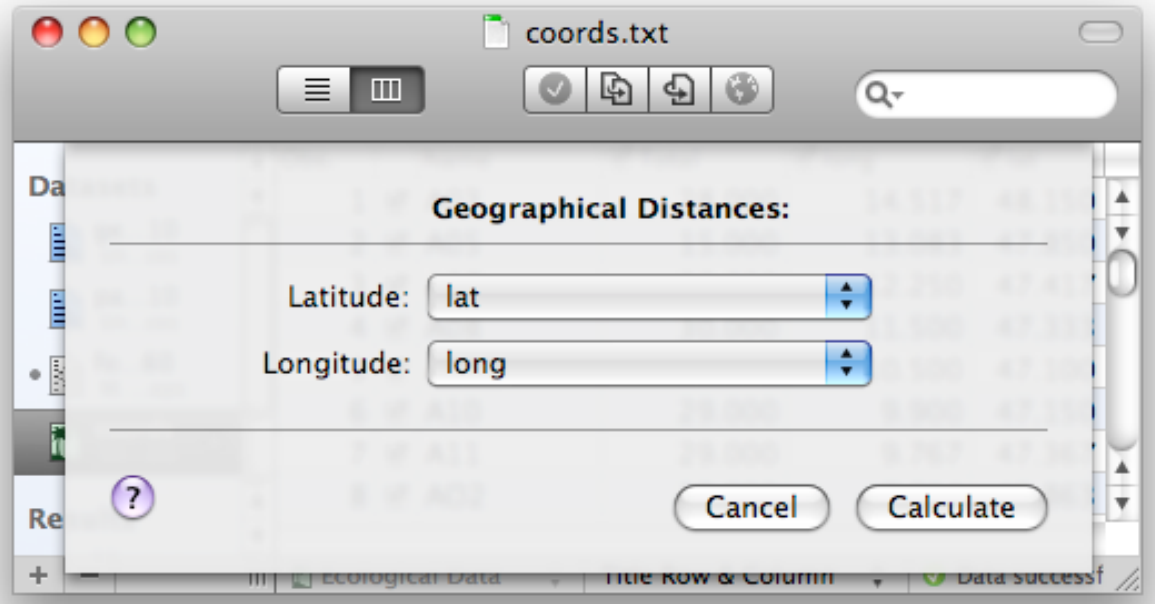
- Clonal – This option calculates the number of mutationsteps that are needed to transfer the genotype of the first individual into the genotype of the second individual (Meirmans & Van Tienderen, 2004). Strictly taken, this assumes clonal reproduction but the distance may also be useful in other cases. Selecting this option from the distances menu will show a dialog in which you can choose the mutation model that is used and set some additional parameters. These options are the same as the ones used for the distance calculations for assigning clones. See Assign Clones - Step 1: Distances
- Bruvo – This option calculates a distance based on the two-phase mutation model for microsatellites (Bruvo et al, 2004). This distance is useful for datasets with a mixture of different ploidy levels (though all other distance indices can also be calculated for different ploidy levels). Selecting this option from the distances menu will show a dialog in which you can set some additional parameters (see screenshot above).
- Kinship – The kinship coefficient of Loiselle et al (1995), based on the relative probability of identity by descent of the alleles within the two compared individuals. Note that this index also uses the allele frequencies within the whole dataset, making the distances between pairs of individuals dependent on all other individuals in the dataset. When the data appears to be dominant (i.e. it is haploid and has only two allelic states per locus), the kinship coefficient of Hardy (2003) is calculated instead, assuming an inbreeding coefficient of 0.
- Smouse & Peakall – This is a squared Euclidean distance based on the number of times a certain allele is found in the two individuals (Smouse & Peakall, 1999). This distance can among others be used for tests of spatial autocorrelation and AMOVAs. GenoDive also uses this distance for AMOVA-based K-Means clustering of individuals. For datasets without any missing data or ploidy variation, this distance is equal to twice the square of the standard Euclidean distance below.
- Euclidean – The standard Euclidean distance between individuals, calculated using the within-individual allele frequencies.

- Manhattan – The standard Manhattan distance between individuals, calculated using the within-individual allele frequencies.
- Chord – The standard Chord distance between individuals, calculated using the within-individual allele frequencies.
- Chi-square – The standard Chi-square distance between individuals, calculated using the within-individual allele frequencies.

Non-Genetic Distances

GenoDive contains several difference distance and similarity measures, which are especially meant for calculation on ecological variables. They can however also be calculated using haploid genetic data. If you want to calculate them using diploid or polyploid data, you will have to transform these to dominant haploid data first.

The included distance measures are: Geographical distances, Euclidean distance, Manhattan distance, Chord distance, and Chi-Square distance. The included similarity measures are Simple matching similarity, Jaccard similarity, Dice similarity, and Steinhaus similarity. For a detailed overview of these and many more measures, see Legendre & Legendre (1998). The Geographical distances are calculated using the coordinates of the observations in latitude and longitude, and are calculated along the surface of the earth, with correction for the flattening of the earth at the poles. To calculate geographical distances, the coordinates should be given in decimal degrees.

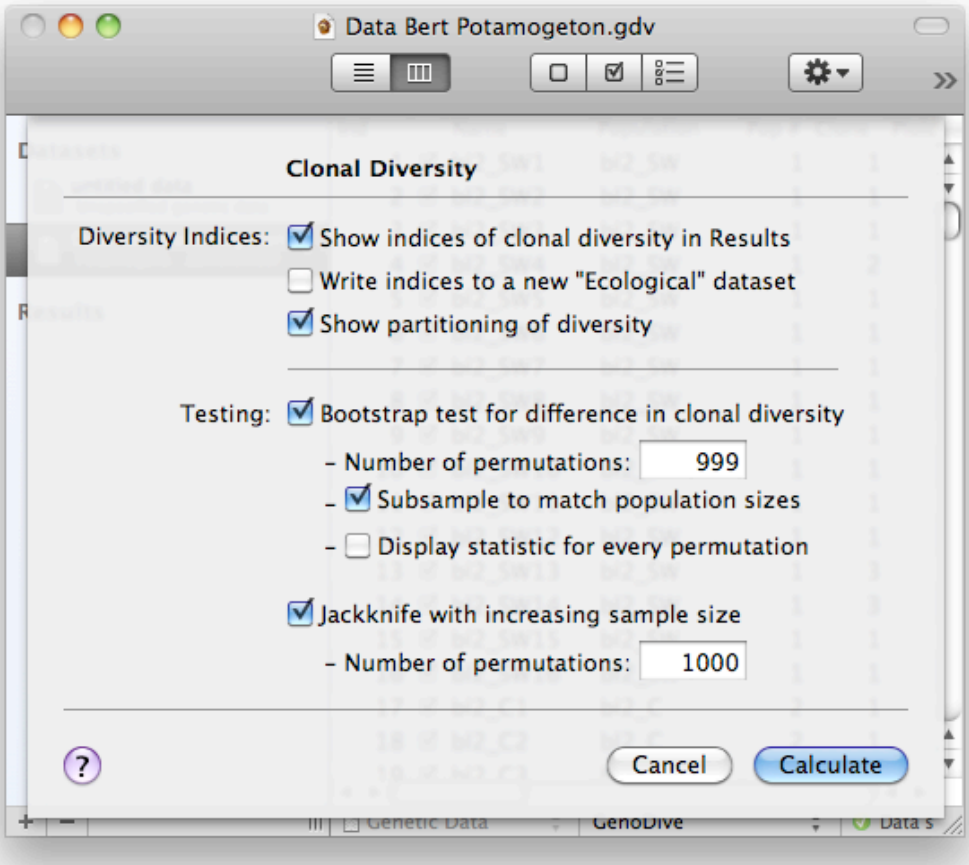


To calculate non-genetic distances:

1. Make sure that the selected dataset contains either ecological variables or haploid genetic data.
2. For ecological data, it is important to include only those variables that you want your distances be based on (see: including, and excluding data).
3. Choose Calculate Distances from the Analysis menu and select your preferred distance index.
4. If you choose Geographical Distances... you will see a dialog (screenshot) in which you have to choose which variable contains the latidude and which variable contains the longitude. Click Calculate.
5. When the calculations are finished, which usually does not take very long, a new document will appear containing the distances.

Estimating Clonal Diversity

Calculation of diversity indices is an important tool in studies on clonal organisms. Studying differences in clonal diversity between populations or species can give major insights in the biological backgrounds of sexual versus asexual reproduction. GenoDive can calculate seven commonly used indices of clonal diversity. For more information, see: background information on clonal diversity.



How to estimate indices of clonal diversity:

1. Make sure that the selected dataset contains either genetic data with clone information or haploid genetic data. To add clone information to a dataset, see Assign Clones
2. Select which diversity indices you want to calculate.
3. If you want to create a new "ecological" datafile with the diversity indices as variables and the populations as observation, switch on the button labeled Write indices to a new "Ecological" datafile.
4. Switch on the Bootstrap test button if you would like to perform a pairwise bootstrap test for difference in clonal diversity between populations. Enter the number of bootstrap permutations and select whether you want to subsample the largest population from each pair to match the size of the smallest population. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations as it may make the Result file extremely large.
5. Switch on the Jackknife button if you would like to perform a jackknife analysis with increasing sample size. Enter the number of jackknife permutations
6. Click Calculate.

Clonal Diversity (background)

Calculation of diversity indices is an important tool in studies on clonal organisms. Studying differences in clonal diversity between populations or species can give major insights in the biological backgrounds of sexual versus asexual reproduction. GenoDive can calculate seven commonly used indices of clonal diversity (abbreviations refer to those used in the results):

Included indices:

- The number of genotypes (num). Simply the number of genotypes found in a population.
- The effective number of genotypes (eff). This is equivalent to the "effective number of alleles" that is sometimes used in allozyme studies. This index may be slightly more insightful than the number of genotypes for comparing diversity between populations, though care should be taken as this index is biased for small sample sizes.
- Nei's (1987) genetic diversity (div) corrected for sample size. Among ecologists, this index is better known as Simpson's diversity index. This is the only diversity index calculated by GenoDive that is truly independent of sample size. The index is also widely used in population genetics under the name of "expected heterozygosity".
- The evenness (eve = eff / num). Basically, every diversity index has its own evenness, which is simply calculated as the estimated value of the index devided by the maximum value possible for the used sample size. For a diversity index that has a maximum of one, the corresponding evenness therefore equals the index itself. The evenness is an indicator for how evenly the genotypes are divided over the population, hence the name. An evenness value of 1 indicates that all genotypes have equal frequencies. GenoDive calculates the evenness of the effective number of genotypes, which is the most widely used one. However, as the effective number of genotypes itself it has an estimation bias related to the sample size, also the related evenness has an estimation bias.
- Shannon index (shu). This index, also known as the Shannon-Wiener index or as the Shannon-Weaver index, is the most widely used diversity index in ecology. However, the index has a huge estimation bias and therefore is not always useful in genetics, unless all sample sizes in a study are more or less equal. The corresponing evenness can easily be calculated by dividing the estimate by log(s). For calculating the Shannon index, some people use a natural logarithm instead of 10log that is used here, so be careful when comparing your results with other studies unless you know what method they used.
- Shannon index (shc) corrected for sample size (Chao & Shen, 2003). This is a recently published version of the Shannon index that uses a non-parametric bias correction. For the correction, the number of singletons (types that are only sampled once) is used to estimate the number of unsampled types. Though this removes the bias rather well for sample sizes > ~50, it still has a bias for smaller sample sizes. The used method of correcting the bias is not possible in some cases: e.g. when all the individuals in a populations have different genotypes. In that case, you will see "nan", instead of a number.
- Nei's uncorrected genetic diversity (diu). This is index number 3 (div), without the n/(n-1) correction term.

Partitioning of diversity

Next to these population specific indices, for Nei's index and the two Shannon indices estimates of the total diversity, the average diversity per population and the among-populations diversity are given. However, there is a difference in calculation of these estimates between the different indices. For Nei's index, I used the formulas from Nei (1987), and the fraction of among-populations diversity is in this case equivalent to Gst. For the two Shannon indices, I found no ways of properly calculating the total and average within population diversity. The total diversity is therefore simply calculated by pooling all individuals and the average within-population diversity is calculated through averaging over populations. This method is not free of bias: The Shannon indices depend on the sample size and the total sample size is always bigger than the average of the population sample sizes. Therefore, the estimate of the total diversity will usually be higher than the estimate of the within-population diversity, even when there should be no among-populations diversity (e.g. when all the samples come from the same population). In other words: the estimate of the fraction of among-population diversity gets inflated. This bias is, however, less for the corrected Shannon index, as it is less dependent on the sample size.

For convenience, the program shows the among-population components of the Shannon indices also under "Gst" in the output, though these are technically no real Gst-values (and certainly not estimates of Fst!!). These "Gst" estimates are not corrected for the number of populations, and therefore not well suited for datasets containing only a few populations. Therefore, also the corrected versions "G'st" are calculated (see Nei 1987). Note that all the "Gst" and "G'st" values are based on genotype frequencies and not on allele frequencies!

The "Gst" and "G'st" estimates for Nei's index are usually lower than those for the Shannon indices. This is not only because of the above-mentioned bias, but also (mainly) because of the different (statistical) behaviour of the indices. This difference in behaviour is the reason that most ecologists prefer the Shannon index, even though it has an estimation bias; they think that the Shannon better represents the diversity seen in nature. It must be said, however, that in Ecology the estimation bias is generally not as important as it is in Genetics.

Bootstrap test

GenoDive includes a bootstrap test to test whether pairs of populations differ in their clonal diversity. This type of test can be especially useful if some populations are expected to have a higher diversity than others; for example due to the presence of sexuals in these populations or because of an expected geographical trend in clonal diversity. The test uses a bootstrapping approach (resampling with replacement); the individuals are resampled from the populations and the diversity indices are compared after every replicate (Manly, 1991). The test is performed for all diversity indices (but not pairwise population comparisons!) simultaneously, which makes the p-values for the different diversity indices dependent on each other. However, I assume that most users will focus on only one index and ignore the others. The bootstrap test has a bias when the sample sizes of the compared populations differ a lot and of course the test also has a bias for diversity indices with an estimation bias. These biases can be overcome by subsampling the samples to be of equal size before bootstrapping.

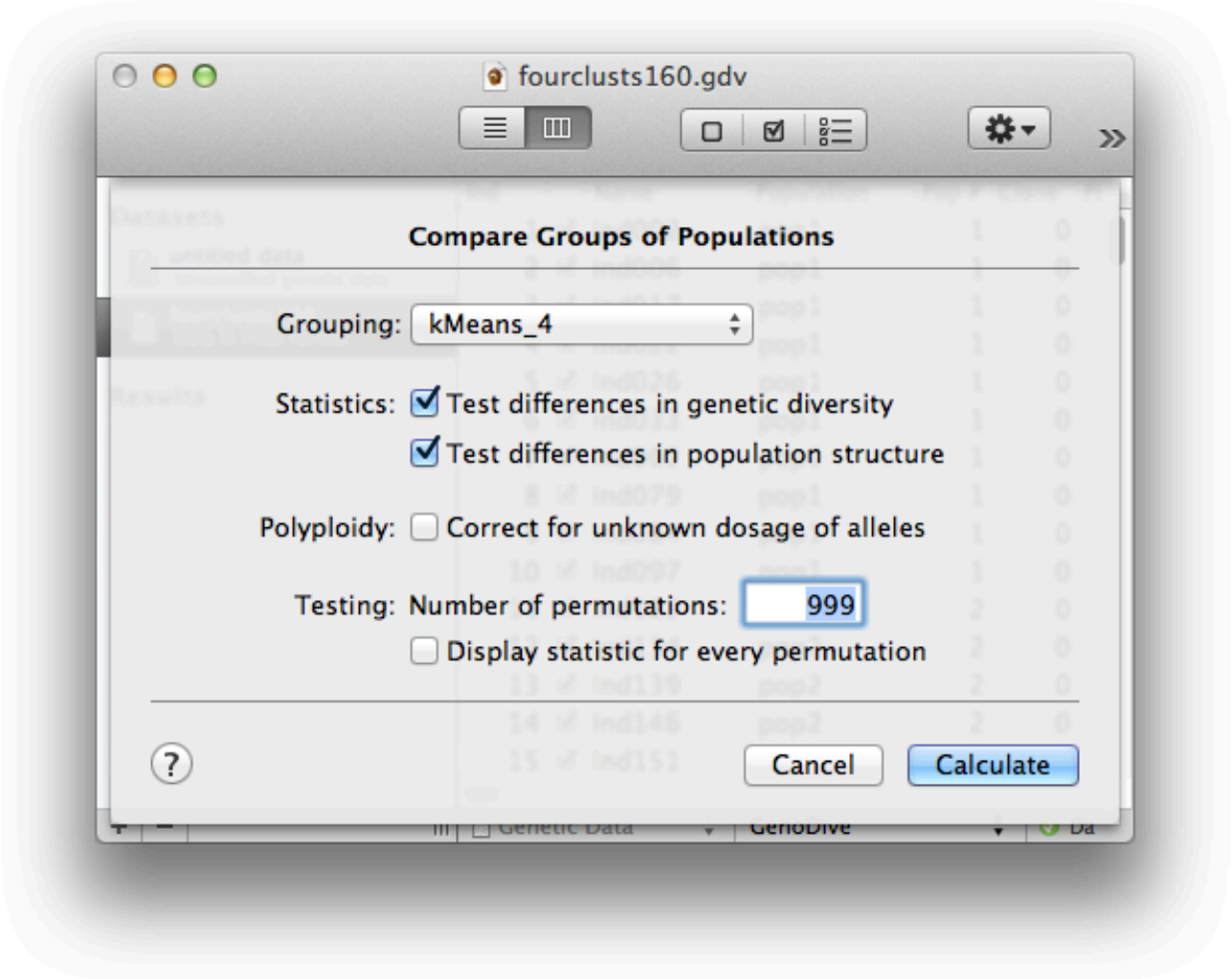
Jackknife analysis

All of the above diversity indices, except the Nei's corrected diversity index, have an estimation bias for small sample sizes, even the "corrected" Shannon index. If you want to use any of these biased indices, a jackknife analysis can be used to check whether sample sizes are big enough to be able to estimate those indices without bias, by taking increasingly large subsamples of your data, starting at 2. If the trend in the value of the index for the different subsample sizes has leveled off when it reaches the actual sample size, the sample size was adequate. If the trend has not leveled off, you should have sampled more individuals, or you should calculate an unbiased diversity index.

Compare Groups

GenoDive can tests whether groups of populations differ in their values of certain summary statistics. For example, one may be interested whether populations from a newly colonised part of the range have a lower genetic diversity than the populations from the original range. For this, GenoDive calculates the average of the summary statistics within every group and then uses a permutation test to test for differences among the groups. As a test statistic, the OSx-statistic is used (Goudet, 1995): the sum of the squared differences in the test statistic over all pairwise combinations of groups. Permutations take place by randomising the populations over the groups.

The test can be performed for several of the summary statistics from the G-statistics analysis: H_o, H_s, G_{is}, G_{st}, G'_{st}, G'_{st}, G''_{st}, D_est. Note however, that the test is very sensitive for the statistics of genetic diversity (H_o, H_s). Stochastic variation in the demographic processes within the groups can easily lead to significant differences in H_o and H_s, even when there are no differences in population size, bottlenecks or mutation rate among the groups. Therefore, it is best to only use this test when you want to test an explicit hypothesis for which you can state a priori predictions about the sign and size of the difference.



To tests for differences among groups:

1. Make sure the selected dataset contains genetic marker data
2. Choose Compare Groups... from the Analysis menu
3. Select which test statistics should be used. It is possible to test for differences in diversity statistics (H_s, H_o, G_{is}) and statistics for population differentiation (G_{st}, G''_{st}, D).
4. Type in the number of permutations to test the significance.
5. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations as it will make the Result file overly large.
6. For polyploid data it is possible to correct the allele frequencies for the unknown dosage of alleles. Note that H_o and G_{is} will not be based on the corrected frequencies and will be biased.
7. Click Calculate.

Correction of allele dosage in polyploids

The analysis of polyploid genetic data is challenging since it is often not possible to obtain the dosage of alleles from the marker phenotypes. For example, in a tetraploid an individual that has marker phenotype AB, so two marker bands, at a locus may have genotype AAAB, AABB, or ABBB. Unsurprising, this missing data leads to a bias in the calculation of the allele frequencies, and hence to a bias in the summary statistics that are calculated based on those frequencies.

In several analyses, GenoDive allows for a correction of the unknown dosage of alleles, which allows estimation of these statistics without or with less bias. This correction is possible for every ploidy level and can also be calculated for populations with a mixture of different ploidy levels. The method assumes random mating within populations; in mixed populations even among individuals with different ploidy levels.

This correction uses a maximum likelihood method based on random mating within populations, which is heavily modified from De Silva et al. (2005). For every incomplete marker phenotype (e.g. AB) it is possible to calculate the likelihood of observing that phenotype given some set of allele frequencies, by combining the likelihoods of the underlying genotypes (AAAB, AABB, or ABBB). The algorithm starts by calculating the likelihood of all observed phenotypes for a population, given the uncorrected (biased) allele frequencies. The allele frequencies are then slightly changed and the likelihood is recalculated. If the fit improves, the changed allele frequencies are accepted and the old ones discarded. This is continued until convergence is reached. This algorithm is run separately for every locus and every population.

Analyses where dosage correction is available:

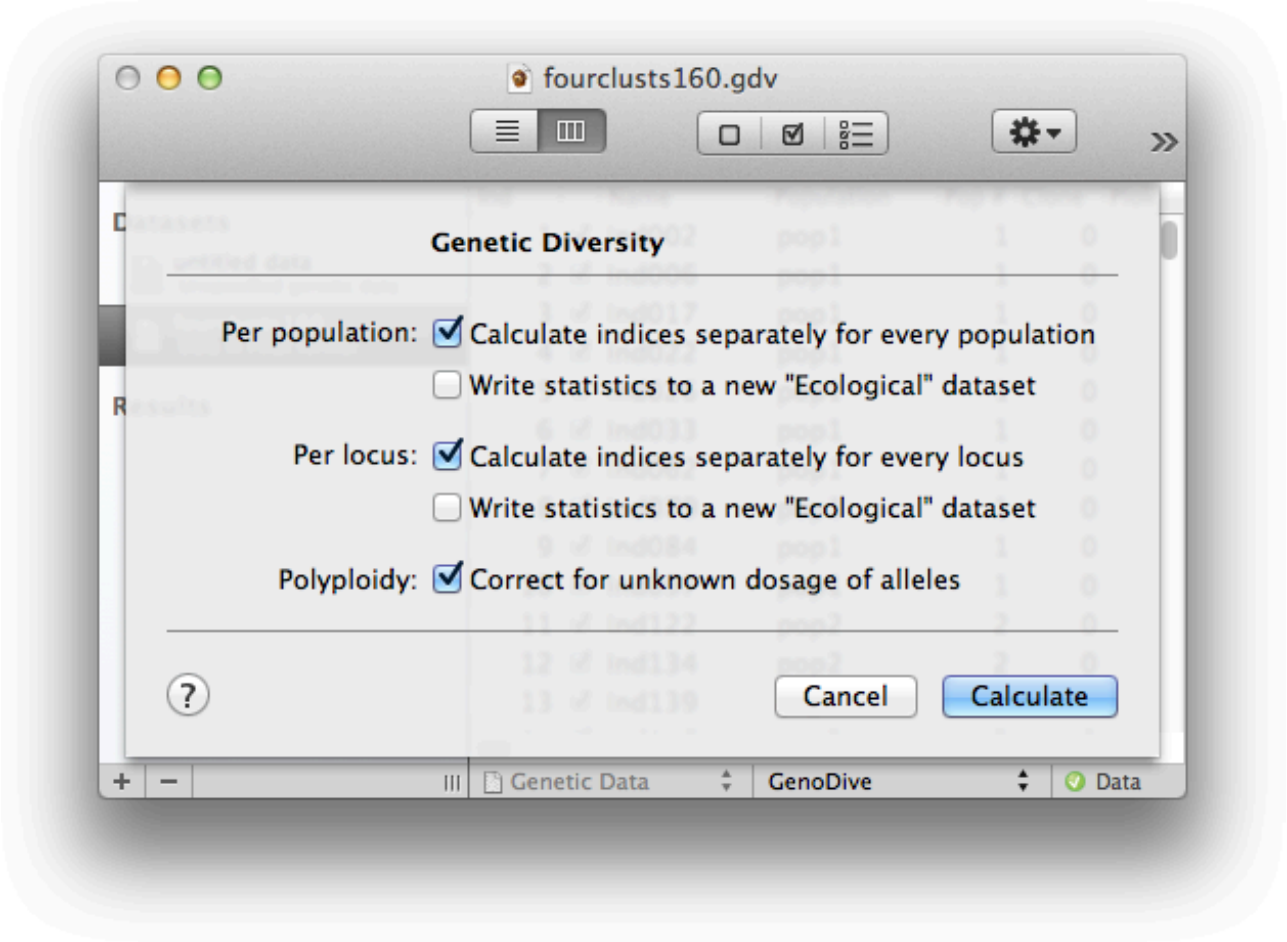
- Allele frequencies
- Genetic distance
- K-means clustering
- Principal Components Analysis
- Compare Groups
- G-statistics
- Filling missing data

Genetic Diversity

GenoDive can calculate various different indices of genetic diversity (see Nei, 1987). In addition to calculating these statistics over all loci and all populations, GenoDive can calculate indices of genetic diversity separately for every population or every locus. This can useful to compare different types of populations to see whether they differ in diversity, for example to see whether there is a correlation between population size and diversity.

Heterozygosity-based Statistics:

- Number of alleles (Num) - The number of alleles observed.
- Effective number of alleles (Eff_num) - The number of alleles in a population, weighted for their frequencies. Care should be taken when comparing this among populations and studies as it is very strongly dependend on the sample size.
- Observed Heterozygosity (H_o) - The observed frequency of heterozygotes within subpopulations, ranging from 0 (all individuals are homozygous) to 1 (all individuals are heterozygous). For polyploids, here the "gametic heterozygosity" (Moody et al., 1993) is given, the chance that two random alleles drawn from the individual are the same.
- Heterozygosity Within Populations (H_s) - The expected frequency of heterozygotes within subpopulations, assuming Hardy-Weinberg equilibrium. This statistic is also known as the "gene diversioty". This is a sample estimate, which includes a correction for sampling bias stemming from sampling a limited number of individuals per population (Nei, 1987).
- Total Heterozygosity (H_t) - The expected frequency of heterozygotes over all populations, assuming Hardy-Weinberg equilibrium. This is a sample estimate, which includes a correction for sampling bias stemming from sampling a limited number of individuals per population (Nei, 1987).
- Corrected Total Heterozygosity (H_t) - The total heterozygosity, with a further correction for a bias that stems from sampling a limited number of populations (Nei, 1987).



To calculate Genetic Diversity:

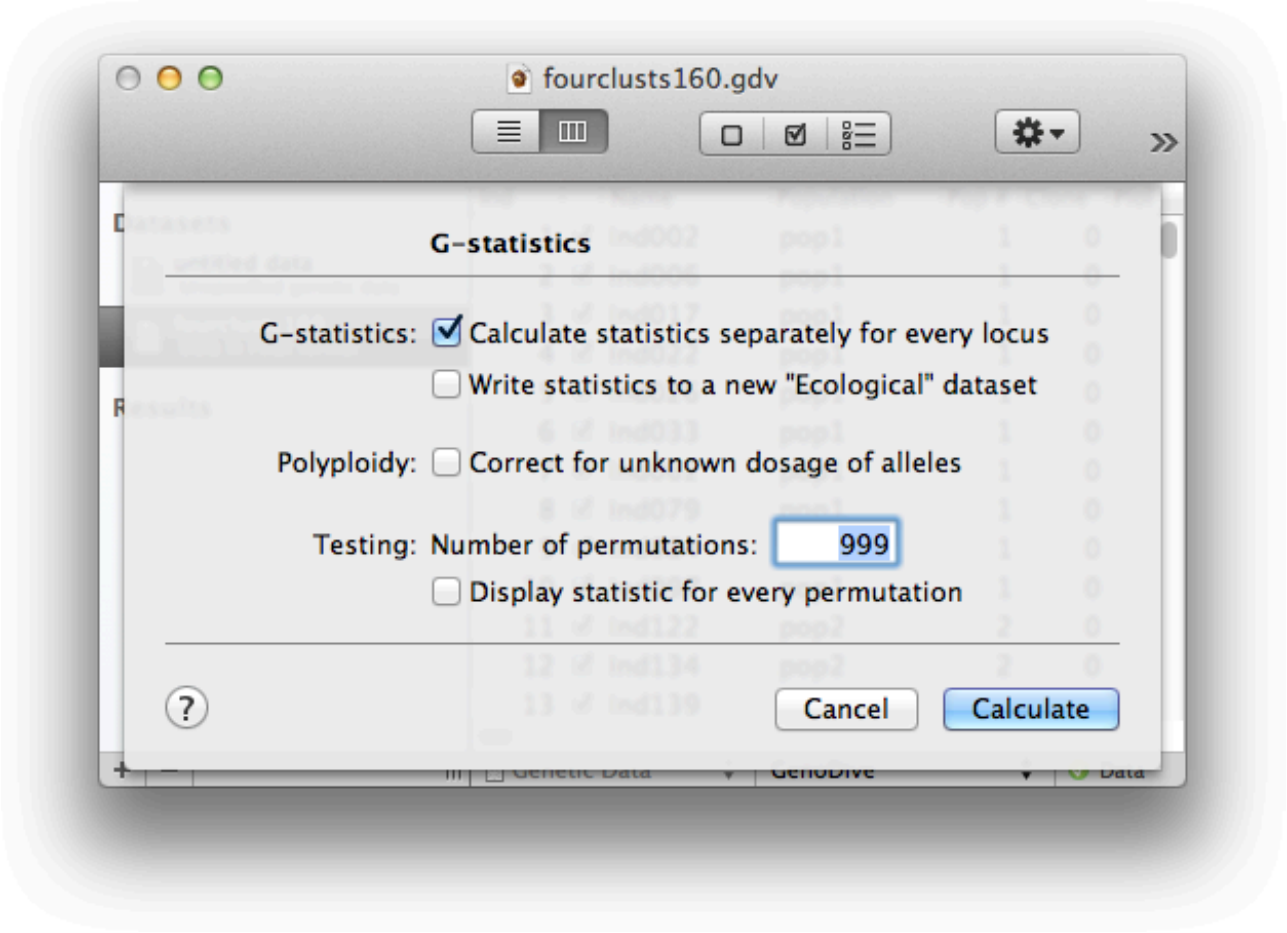
1. Make sure the selected dataset contains genetic marker data
2. Choose Genetic Diversity... from the Analysis menu
3. Select the desired output options. The multilocus values over all populations are always written to the Results, but you can also choose to have the values for all loci or for all populations calculated. The per locus and per population values can also be written to new "Ecological" datafiles.
4. For polyploid data it is possible to correct the allele frequencies for the unknown dosage of alleles. Note that H_o and G_{is} will not be based on the corrected frequencies and will be biased.
5. Click Calculate.

G-statistics

GenoDive can calculate various different diversity and differentiation statistics, which go back to Nei's original work on G_{st} and related statistics (Nei, 1987). G_{st} and related statistics can be used to quantify the degree of differentiation among populations. G_{is} is an analogues of F_{is}, which describes the degree of deviation from Hardy-Weinberg equilibrium. These statistics are calculated by comparing different relating the genetic diversity within populations (H_s) to the diversity within individuals (H_o) or the overall genetic diversity (H_t). Standard errors are calculated for all statistics by jacknifing over loci, when more than 5 loci are available.

Heterozygosity-based Statistics:

- Number of alleles (Num) - The number of alleles observed.
- Effective number of alleles (Eff_num) - The number of alleles in a population, weighted for their frequencies. Care should be taken when comparing this among populations and studies as it is very strongly dependend on the sample size.
- Observed Heterozygosity (H_o) - The observed frequency of heterozygotes within subpopulations, ranging from 0 (all individuals are homozygous) to 1 (all individuals are heterozygous). For polyploids, here the "gametic heterozygosity" (Moody et al., 1993) is given, the chance that two random alleles drawn from the individual are the same.
- Heterozygosity Within Populations (H_s) - The expected frequency of heterozygotes within subpopulations, assuming Hardy-Weinberg equilibrium. This statistic is also known as the "gene diversioty". This is a sample estimate, which includes a correction for sampling bias stemming from sampling a limited number of individuals per population (Nei, 1987).
- Total Heterozygosity (H_t) - The expected frequency of heterozygotes over all populations, assuming Hardy-Weinberg equilibrium. This is a sample estimate, which includes a correction for sampling bias stemming from sampling a limited number of individuals per population (Nei, 1987).
- Corrected Total Heterozygosity (H_t) - The total heterozygosity, with a further correction for a bias that stems from sampling a limited number of populations (Nei, 1987).
- Inbreeding Coefficient (G_{is}) - Relates the observed heterozygosity within subpopulations (H_o) to the expected heterozygosity (H_s), ranging from -1 to 1. Often used to detect departure from Hardy-Weinberg equilibrium; analogues to F_{is}.
- Fixation Index (G_{st}) - Measures the amount of fixation of alleles within subpopulations, relative to the total population (Nei, 1987). Widely used as a measure of genetic differentiation; analogues to F_{st}.
- Corrected Fixation Index (G'_{st}) - Fixation index, with a correction for a bias that stems from sampling a limited number of populations (Nei, 1987).
- Standardised Fixation Index (G'_{st}) - Fixation index, standardised relative to the maximum value it can reach given the amount of within-population diversity (H_s) (Hedrick, 2005).
- Corrected StandardisedFixation Index (G''_{st}) - Standardised fixation index, with a correction for a bias that stems from sampling a limited number of populations (Meirmans & Hedrick, 2011).
- Population Differentiation (D_est) - Index of population differentiation that is independent of the amount of within-population diversity (H_s) (Jost, 2008).

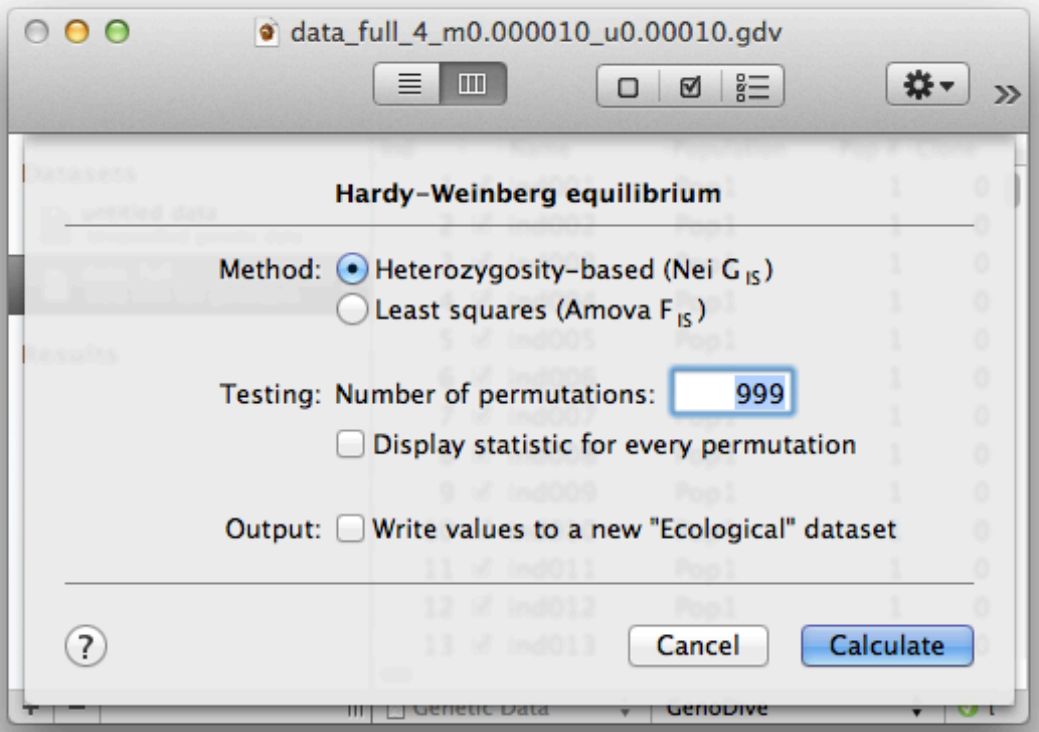


To calculate G-statistics:

1. Make sure the selected dataset contains genetic marker data
2. Choose G-statistics... from the Analysis menu
3. Select the desired output options for the calculated G-statistics. The multilocus values are always written to the Results, but you can also choose to have the values for all loci calculated. The per locus values can also be written to a new "Ecological" datafile.
4. For polyploid data it is possible to correct the allele frequencies for the unknown dosage of alleles. Note that H_o and G_{is} will not be based on the corrected frequencies and will be biased.
5. Type in the number of permutations to test the significance.
6. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations as it will make the Result file overly large.
7. Click Calculate.

Hardy-Weinberg equilibrium

This option performs a test to see whether populations are in Hardy-Weinberg equilibrium; i.e. to see whether the genotype frequencies do not deviate from the frequencies that are expected under random mating. As estimator of F_{is}, GenoDive can use the either the heterozygosity-based G_{is} statistic (Nei 1987), or the F_{is} statistic from an Amova (Excoffier, 1992, Michalakis & Excoffier, 1996) to test for Hardy-Weinberg equilibrium. The latter estimator is equivalent to using Weir & Cockerham's (1984) small f statistics. The two statistics generally give comparable results, though they differ slightly in the way that values are averaged over populations.



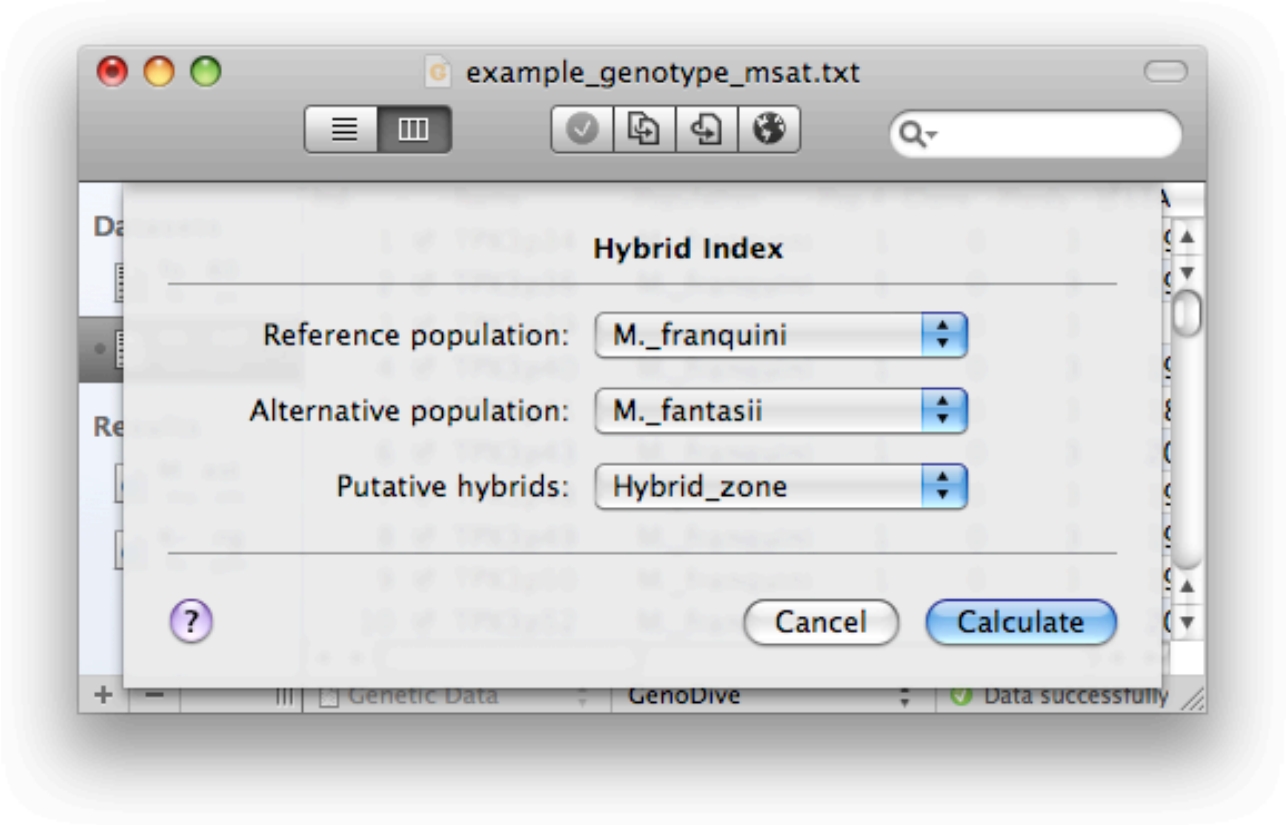
To perform a test for Hardy-Weinberg equilibrium:

1. Make sure the selected dataset contains genetic marker data
2. Choose Hardy-Weinberg... from the Analysis menu
3. Select whether you want to use a heterozygosity based estimator (G_{is}) or a least-squares based estimator (F_{is})
4. Type in the number of permutations to test the significance.
5. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations and for datasets with many populations or loci, as it will make the Result file overly large.
6. Select whether you want to have the results written to a new ecological datafile. This option is useful if you want to perform further tests on the inbreeding coefficient, e.g. calculate a correlation with population size. Note that the results are always written to the Results window.
7. Click Calculate.

Hybrid Index

A hybrid index is a quantitative estimate of the genetic contribution of two parental species or populations to an individual of unknown provenance. GenoDive uses the method of Buerkle (2005) to calculate a maximum likelihood estimate of such a hybrid index. The method of Buerkle is extended to include polyploid individuals, up to octaploids. The analysis requires three datasets, which should be coded as populations in a genetic data file. Two populations should contain the genotypes for the two parental genepools, referred to as the Reference population and the Alternative population; usually these are two species. The third population should contain the genotypes for the putatively hybrid individuals.

The analysis returns the maximum likelihood estimate of the hybrid index, the likelihood value, and the upper and lower limits of the 95% confidence interval.



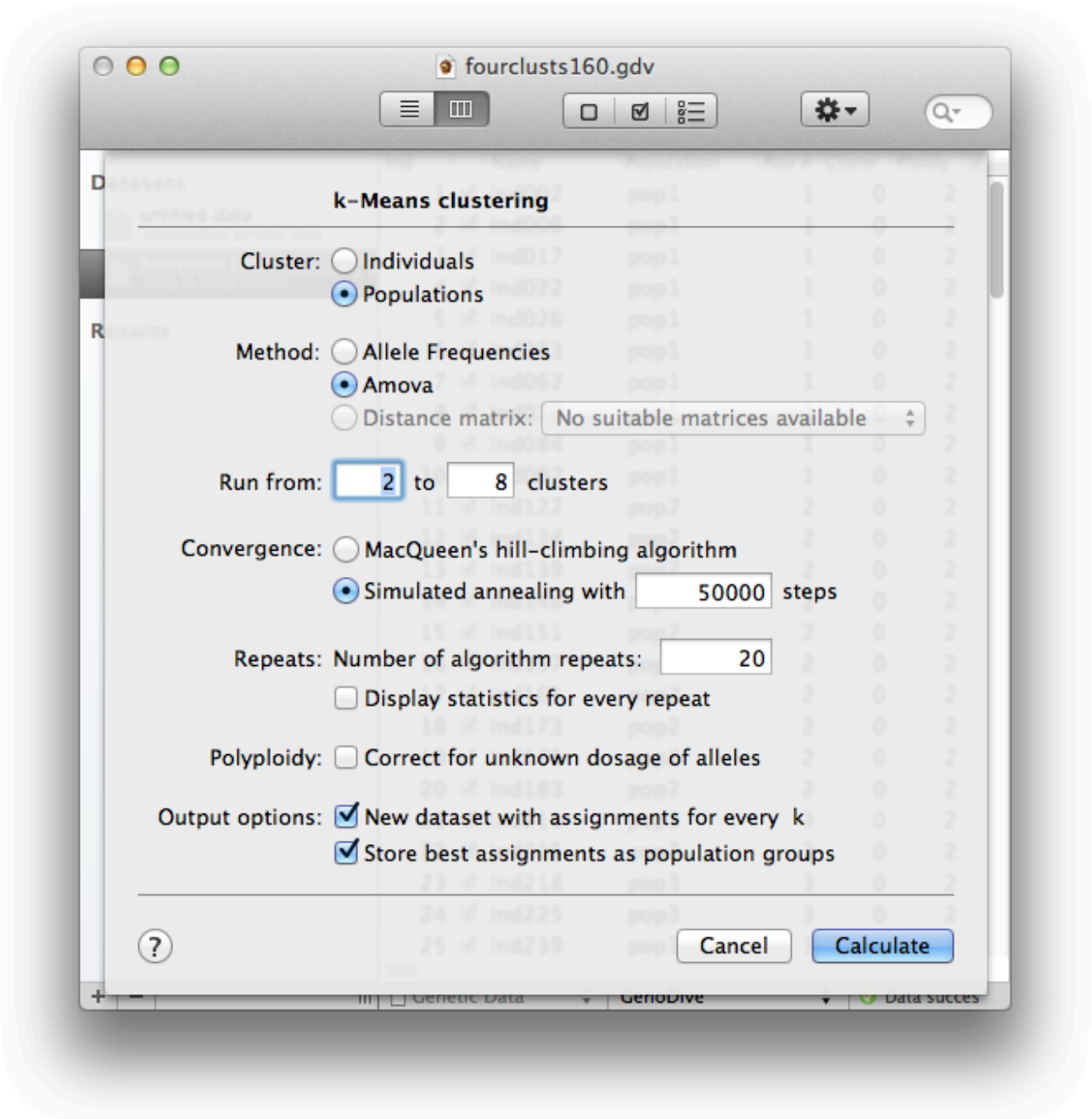
To calculate the hybrid index:

1. Make sure the selected dataset contains genetic data with at least three populations.
2. Choose Hybrid Index... from the Analysis menu
3. Using the three popup menus, choose the correct populations for the Reference and Alternative populations and for the Putative hybrids.
4. Click Calculate.

K-Means clustering

K-Means clustering divides a number of objects into a a priori assigned number (k) of groups in such a way that the within-group diversity is minimised and the amon-group diversity is maximised. The implementation in GenoDive can perform the clustering either on genetic marker data or on ecological data. For genetic data it is possible to either cluster individuals into presumed populations, or to cluster populations into population groups (Meirmans, 2012a). For more information, see: background information on K-Means clustering.

When there are a lot of missing values in a dataset, K-Means clustering can be biased as individuals with missing data tend to be grouped together. Therefore, it is advised to first replace the missing data with random values (see Fill in Missing Data) before performing the clustering.



To cluster populations or individuals using K-Means:

1. Make sure that the selected dataset contains either genetic marker data or ecological data.
2. Choose K-Means Clustering... from the Analysis menu.
3. Select whether you want to cluster individuals or populations. This choice is only available for genetic marker data. In general, clustering is only possible if there are more than four objects included in the dataset.
4. Choose the method: for genetic data either allele frequencies, AMOVA, or an existing distance matrix; for ecological data either Euclidean distances or an existing distance matrix.
5. Fill in the range of values of k for which you want to run the analysis. The minimum possible value is 1, and the maximum is half the number of objects to be clustered.
6. Choose the convergence method, either the classic hill-climbing algorithm or simulated annealing with a specified number of steps.
7. Specify the number of times that the algorithm should be repeated, suitable numbers are 500 for the hill-climbing algorithm and 20 for simulated annealing. If you would like to inspect the results for every random start, switch on the "Display all statistics" button.
8. For polyploid data it is possible to correct the allele frequencies for the unknown dosage of alleles.
9. Choose the output options. For a better overview of the clusterings, you can choose to create a new document with the best results for every value of k. There is an additional output option for genetic data, the effect of which depends on the objects being clustered. If you are clustering individuals, you can choose to write the clusters as clones in the datafile. If you are clustering populations, you can choose to add the best clustering as a series of population groups.
10. Click Calculate.

K-Means clustering (background)

K-Means clustering divides a number of objects into a a priori assigned number (k) of groups in such a way that the among-groups Sum of Squares is maximised. The implementation in GenoDive can perform the clustering either on genetic marker data or on ecological data.

Method outline

The method uses a pairwise matrix of distances between all observations. Given a certain clustering into k groups, for every group the within-group Sum of Squares is calculated by taking the sum of the squared within-group distances. When the distances are Euclidean, this is equivalent to calculating the sum of the squared distances from the points to the group's centroid. The Error Sum of Squares is then found by summing over groups. The amount of variance explained by the grouping is then calculated by dividing the Error Sum of Squares by the Total Sum of Squares.

Because the method uses a matrix of distances, it is very flexible as it can be used with many different types of distance indices. For genetic data, a very straightforward method is to use it with a matrix of Euclidean distances based on within-population or within-individual allele frequencies. On the other hand, the method of calculating the Sums of Squares is very similar to that used by an Analysis of Molecular Variance (Excoffier et al., 1992), so it is also possible to use the Sums of Squares from an AMOVA to perform the clustering (Meirmans, 2012a). A third option is to use an existing distance matrix. A suitable distance matrix contains as much observations as there are individuals or populations (depending on what is being clustered) in the used dataset. The clustering method works best with Euclidean distances, but other distance measures are also possible, though in those cases the presented Anova table does not represent a true Anova.

Using hill-climbing

This is the classical method to perform K-Means clustering first developed by MacQueen (1967), which is implemented in many statistical programs. The analysis starts by assigning every observation at random to one of the k groups and then calculates the Error Sum of Squares. A new clustering is then made by removing one by one each observation and placing it into the group to which centroid it is closest. This process is repeated and every iteration the new clustering will have a smaller Error Sum of Squares, until at some point convergence is reached. The problem with this method is that it can only climb uphill, so it is very likely to get stuck at a local optimum. Therefore the whole procedure is repeated a number of times (say 100) and the best solution is taken from among those repeats. For small datasets, this is a very fast and useful method. However, for very complex problems such as a dataset of many populations genotyped at many highly variable loci, the size of the solution space is huge and may contain many millions of local optima. The chance that the overall minimum is found is then very small, even when a large number of random starts is used.

Using Simulated Annealing

Simulated annealing uses a Monte Carlo Markov Chain (MCMC) that prevents the clustering from getting stuck in local optima. The algorithm has many similarities with the hill-climbing approach, but has some important differences. It again starts by randomly assigning observations to the k clusters. From that clustering, the MCMC is then started which lasts a certain number of user-defined steps. In every step of the MCMC, a random observation is selected and placed in a different, randomly picked, cluster. If the clustering is better than before, i.e. the Error Sum of Squares decreases, the new situation is accepted. However, there is also a chance that the new situation is accepted if the clustering is worse than it was before. This chance depends on the difference in the Error Sum of Squares and on what is called the "temperature" of the chain. This temperature gradually decreases during the run. Initially, the temperature is high and almost all new clusterings are accepted. In the end, when the temperature reaches zero, only the changes that improve the clustering are accepted. The real power of the method lies in the middle, where most of the climbing is uphill, but every now and then a valley is crossed.

One problem with this method is choosing a suitable starting temperature, if this is either too high or too low, the middle part gets very short and the clustering is suboptimal. As a workaround, GenoDive implements an adaptive starting temperature (Meirmans, 2012a). A very high value is used as initial starting temperature, the as long as the acceptance rate is too high, the starting temperature is lowered iteratively until a suitable value is reached. Despite the improved performance relative to the hill-climbing algorithm, multiple repeats may be necessary for the simulated annealing in order to find the overall optimum.

Which method to use?

The hill-climbing algorithm is usually much faster than the simulated annealing and works rather well for small datasets. However, for datasets with a large number of individuals or populations (depending on what is being clustered), the simulated annealing approach usually returns better results, provided that a suitably large number of steps is used (the default setting of 50.000 is a minimum) and the method is repeated several times. In general, I would use both and see which gives the best results. Also try several runs to see whether these give wildly different results.

Optimal value of k

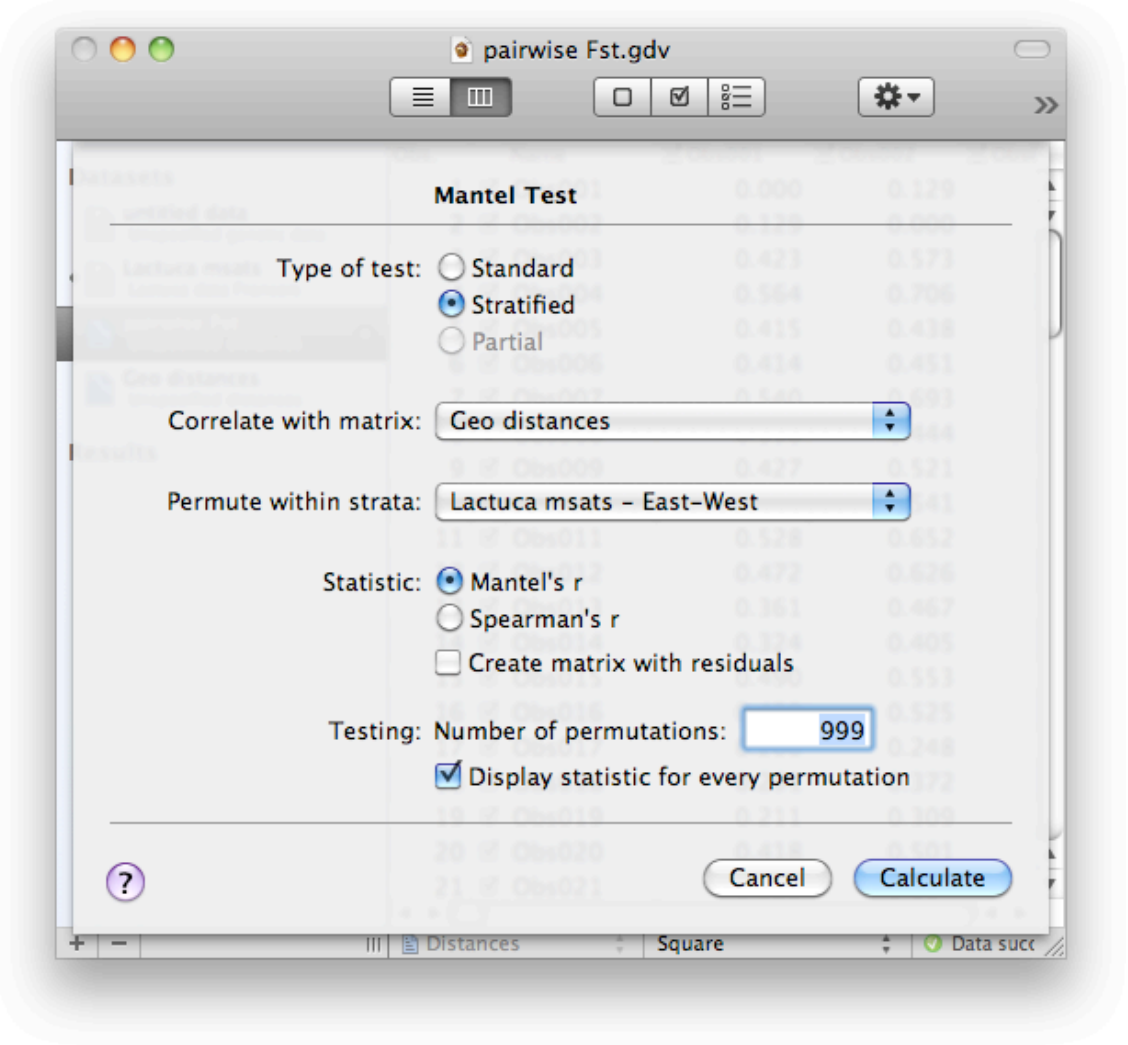
Most often, the number of clusters will not be known a priori. In that case, it is possible to set a range of values for k and perform the clustering for all values. However, this introduces the problem of determining the best clustering among all values of k. The percentage of explained variance is unsuitable for this task as it tends to increase with increasing k. GenoDive provides two different statistics that can determine the number of clusters. The first is the Calinski-Harabasz (1974) pseudo-F-statistic; the optimal clustering is the one with the highest value for the pseudo-f statistic. The other is the Bayesian Information Criterion (BIC, calculated using sum of squares rather than likelihoods); here, the optimal clustering is the one with the lowest value of BIC (Schwarz, 1978).

Simulations have revealed that both work well for clustering populations and individuals, especially when there is random mating within populations (Meirmans, 2012). However, pseudo-F works slightly better when there is non-random mating and for clustering individuals. On the other hand, BIC can be calculated for a single cluster (k=1), whereas pseudo-F can only be calculated for two or more clusters ($k \geq 2$). Therefore, BIC has the benefit that it can be used to determine whether there actually is any population structure at all.

Mantel tests

A Mantel test (Mantel, 1967) is used to test the association between two or three resemblance matrices. Mantel tests are very flexible and can be used in a large number of situations that involve multivariate data. They can e.g. also be used as a generic tool for hypothesis testing, when one of the matrices is a model matrix. Significance testing is performed through permutations, by jointly randomising the rows and columns in one matrix, while leaving the other matrix intact.

Optionally, a distance matrix containing the residuals of a linear regression between the two distance matrices can be calculated. The method assumes that the distances are normally distributed and that the relationship between the two matrices are indeed linear. If you suspect a non-linear relationships bwteen matrices, you can apply a transformation. This method may be used to obtain a distance matrix corrected for the influence of another matrix, e.g. to test for a pattern for population differentiation, corrected for the influence of isolation by distance.



GenoDive offers three types of Mantel tests:

Standard Mantel test

A standard Mantel test tests for the association between two matrices (Mantel, 1967). This test is often used to test whether there is a pattern of Isolation by Distance. In that case the association is tested between a matrix of genetic distances and a matrix of geographical distances. To perform a Partial Mantel test, at least two similar-sized resemblance matrices have to be open.

Stratified Mantel test

A stratified Mantel test, where the permutation scheme is changed to permute observations within specified clusters (Meirmans, 2012b). This test is handy when one wants to test for isolation by distance when there is also hierarchical population structure in the data; in that case sampled populations should be permuted within these clusters. To perform a Stratified Mantel test, at least two similar-sized resemblance matrices have to be open plus one document containing the strata. This document can either contain genetic marker data or ecological data, but not a distance matrix. In the first case, the populations or clones can be used as strata, when the number of individuals corresponds to the size of the matrix. Otherwise the population groups (if specified) can be used as strata, when the number of populations corresponds to the size of the matrix. When an ecological data set is used, one of its variables (columns) can be selected, the data of which will be used as strata (after rounding to integers).

A Partial Mantel test

A Partial Mantel test (Smouse at al., 1986) is used to test the association between two resemblance matrices, while controlling for the influence of a third matrix. This test can e.g. be used find out whether patterns of population differentiation are only due to Isolation by Distance or also to ecological processes. To perform a Partial Mantel test, at least three similar-sized resemblance matrices have to be open.

How to perform a Mantel test:

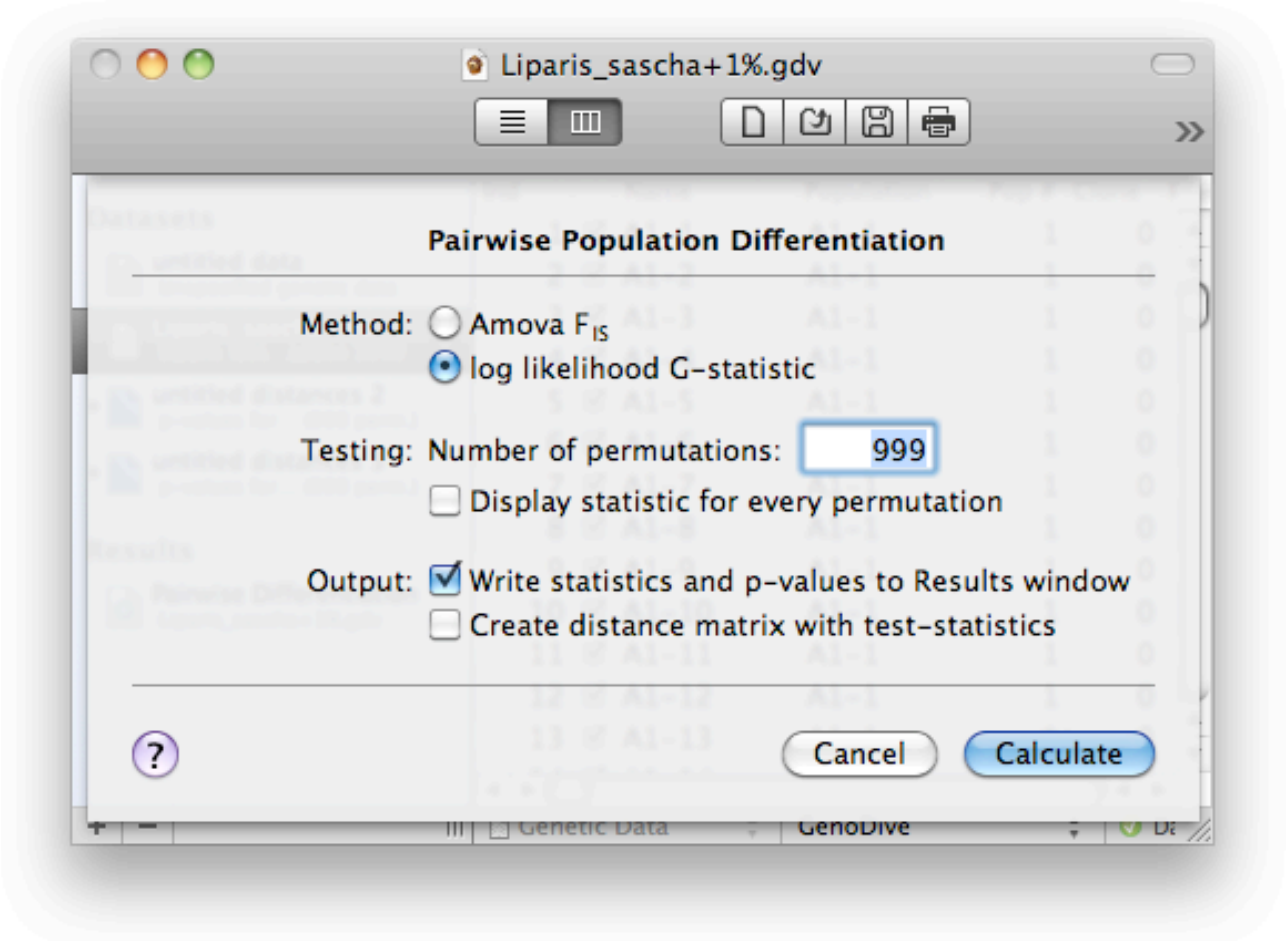
1. Make sure that the frontmost window contains a distance matrix.
2. The Mantel test can only be performed when at least one other distance matrix of equal size is open. Note that the size of a matrix depends on how many observations are included.
3. Select Mantel Test... from the Analysis menu.
4. Choose the type of Mantel test that you want to perform.
5. From the popup menu labelled "Correlate with matrix", select the matrix for which you want to test the association with the current document.
6. If you are performing a Stratified Mantel test, use the second popup menu to select the dataset which you want to use as strata for permutation. If you are performing a Partial Mantel test, use the second popup menu to select the matrix for which you would like to control the association between the first two matrices.
7. Select the test statistic, either Mantel's r or Spearman's r. Mantel's r is most suitable in situations where there is approximately a linear relationship between the two matrices, whereas Spearman's r is non-parametric, based on ranks, and most suitable when the relationship is strongly non-linear or when their distributions are strongly non-normal.
8. Select whether you want a matrix with residuals to be created.
9. Type in the number of permutations to test the significance.
10. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations as it will make the Result file overly large.
11. Click Calculate.

Pairwise population differentiation

This analysis tests for population differentiation between all pairs of populations, the siginificance of the which is assessed using permutations. There is a choice between two test statistics:

- F_{st} from an Analysis of Molecular Variance (AMOVA, Excoffier, 1992, Michalakis & Excoffier, 1996), which is for this purpose exactly equivalent to Weir & Cockerham's (1984) Theta. The F_{st} values obtained are the same as those from the F_{st} Genetic Distances option.
- The log-likelihood G-statistic, a well-known statistic that is often used to test goodness-of-fit or contingency tables. The G-statistic is in principle distributed as Chi-square with the number of alleles found in the two populations, minus one, as the degrees of freedom. However, this possibility is not offered in GenoDive as I think that a permutation approach gives more reliable results, given the often extremely low allele frequencies found in marker data (Goudet, 1996).

The analysis involves multiple tests, and therefore a Bonferroni correction should be applied to the p-values. GenoDive does not apply such a correction so you have to do it yourself (see e.g. Rice, 1989). The number of tests quickly increases with the number of populations so that you may not have much statistical power even for moderately large datasets. In general, rather than to test differentiation between all pairs of populations, it is adviseable to perform an overall test of population differentiation, possibly using a hierarchical population structure (see Amova). Alternatively, the overall population structure can be assessed using a clustering approach (see K-Means clustering).



To perform a test for pairwise population differentiation:

1. Make sure the selected dataset contains genetic marker data
2. Choose Pairwise Differentiation... from the Analysis menu
3. Choose whether you want to use F_{st} or the log-likelihood G as a test statistic.
4. Type in the number of permutations to test the significance.
5. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations, and for datasets with many populations or loci, as it will make the Result file overly large.
6. Choose the desired output options. A distance matrix with pairwise p-values is always created. You can also choose to write all the results of the test, both the values of the test-statistic and the p-values, to the Results window. However, this is not adviced for a large number of populations. In addition, you can choose to create a new distance matrix containing the values of the test statistic.
7. Click Calculate.

Population Assignment

This analysis tries to assign individuals to populations, i.e. it looks what population an individual is most likely to come from. It does this by calculating, for every population, the likelihood that the individual's genotype is found in a population given the allele frequencies in the population (Paetkau et al., 1995).

Usually, population assignment is performed by calculating the allele frequencies from the same data file that contains the individual genotypes that should be assigned. To avoid bias in the assignment, a targeted individual is removed from its source population before the allele frequencies are calculated. Furthermore, to protect from strong errors in teh estimation of allele frequencies, individuals can only be assigned to populations that have more than five individuals.

GenoDive also has the option to use a separate file with allele frequencies for the assignment test. This is for example handy if you want to assign genotyped juveniles using population frequencies calculated from adults. The file with allele frequencies should be an "ecological data" file, with as many columns as there are alleles in the file with individual genotypes. The first column should contain the frequency of the allele at the first locus with the lowest identifier (for example the shortest microsatellite band). The second column should contain the frequency of the allele at the first locus with the second lowest identifier, and so on.

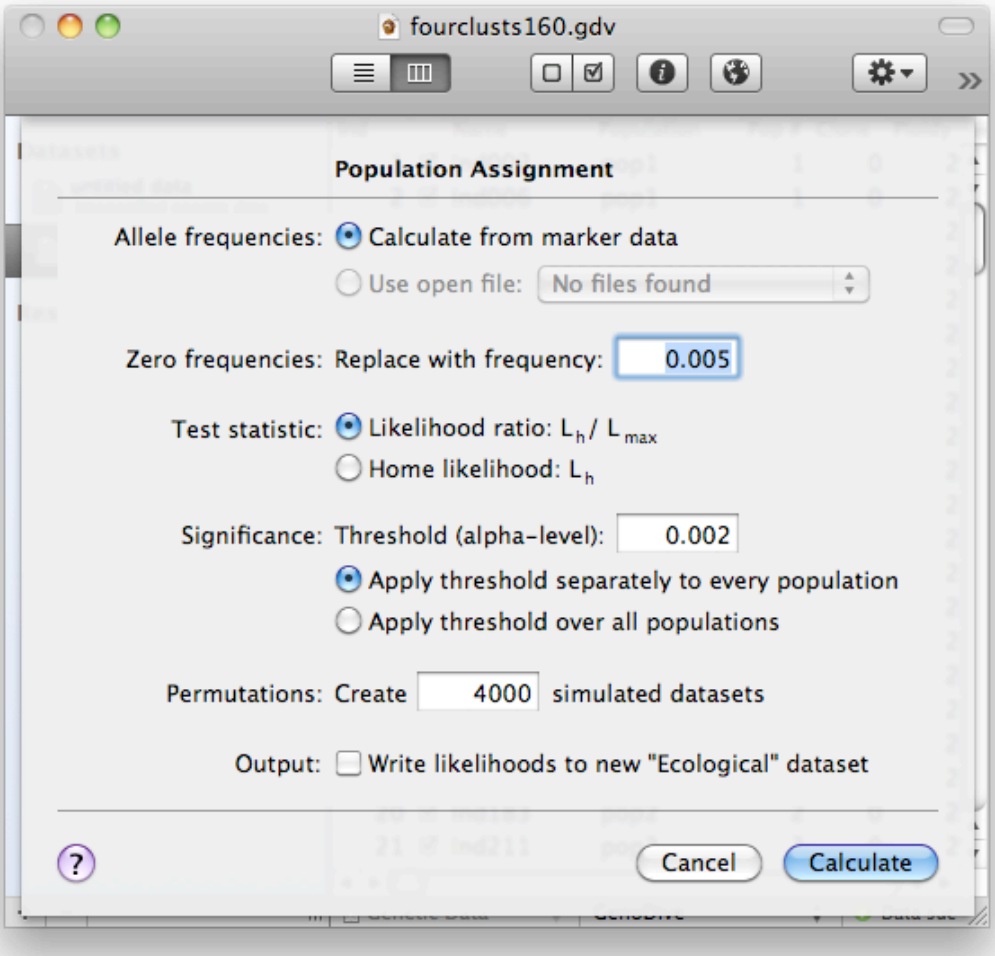
One problem with this analysis is that when a certain allele is present in an individual, but not in a certain population, the likelihood for this population becomes zero. This may be problematic since an allele may in reality be present in a population at a low frequency, but simply not have been sampled. To avoid this problem, any allele frequencies that are found to be equal to zero are replaced by a fixed, low, allele frequency (0.005 was found to work rather well, Paetkau et al., 2004).

Testing

A Monte Carlo test (Cornuet et al. 1999) can be used to generate a null-distribution of likelihood values with which the values for the sampled individuals are compared. A threshold value is then determined from this distribution based on a predefined alpha-value. Individuals with a likelihood value lower than the threshold are thought to be migrants. In GenoDive, the threshold can either be calculated separately for every population, or over all populations.

Two different statistics can be used: 1) The ratio between the likelihood that the individual comes from the population where it was sampled and the maximum likelihood found over all samples. 2). The likelihood that the individual comes from the population where it was sampled. The former gives greater power, but assumes that all possible source locations of migrants have been sampled. The latter is more appropriate if only a part of all possible source populations have been sampled.

In every permutation of the Monte Carlo test, a whole new dataset is recreated with the same number of populations and the same total size as the original dataset. New genotypes are created by randomly picking two parents and creating an offspring genotype from two randomly gametes that are randomly drawn from those two parents (Paetkau et al., 2004).

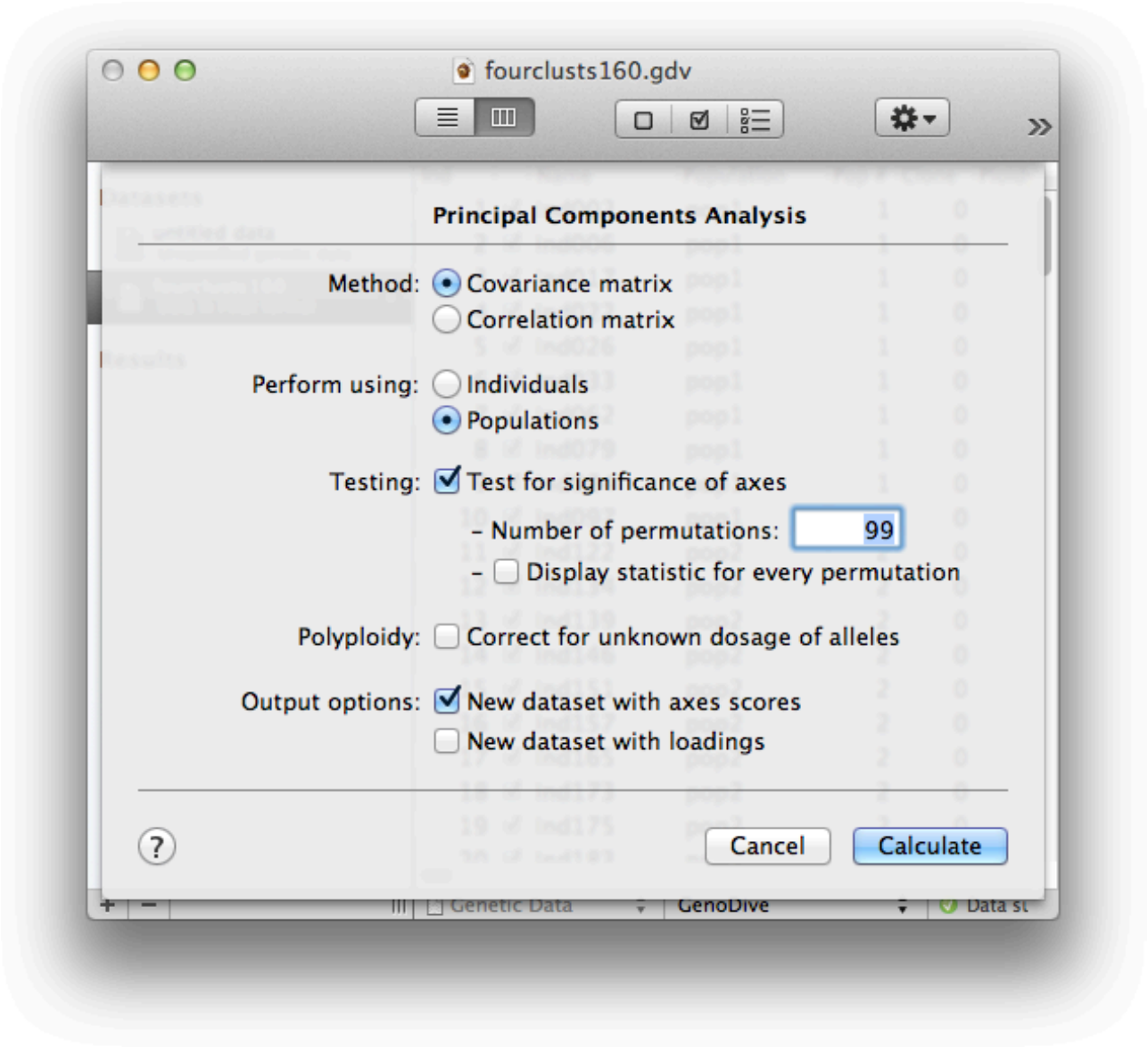


To perform an analysis of population assignment:

1. Make sure the selected dataset contains genetic marker data
2. Choose Population Assignment... from the Analysis menu
3. If you want the allele frequencies to be calculated from the same file that contains the individual genotypes, select "Calculate from marker data". If you want to use an external file with allele frequencies, click "Use open file" and select the appropriate file from the popup menu.
4. Fill in the baseline frequency with which you want to replace any allele frequencies that are found to be equal to zero (or smaller than the given baseline). Enter 0.0 if you do not want to replace zero frequencies.
5. Choose the summary statistic you would like to use for the Monte Carlo test.
6. Select a significance threshold (alpha) to use for testing. This value should be low enough to prevent false positives, in general, it should be lower than the expected migration rate. The default value of 0.002 has been shown to work well Paetkau et al., 2004.
7. Give the number of datasets to simulate for the Monte Carlo test. For every permutations a whole new dataset is recreated, so the total number of permutated individuals depends on the sample size.
8. Select if you want GenoDive to display the likelihood-scores for every combination of individual and population as a new "Ecological" data file.
9. Click Calculate.

Principal Components Analysis

A Principal Components Analysis (PCA) is used to reduce the complexity of multivariate data. It does so by rotating the variable axes in such a way that the greatest possible amount of variance is explained by the first new component axis, the second greatest amount by the second axis, and so on. Mathematically, the analysis is performed through an eigenanalysis on a covariance or correlation matrix. The eigenvectors represent the axis loadings, and the eigenvalues indicate the relative amount of variance explained by the axes.



GenoDive can perform a PCA on either ecological variables or on the allele frequencies from genetic marker data. In the latter case, the allele frequencies can either be calculated for individuals (so 0, 0.5 and 1 for diploids) or for populations. When population allele frequencies are used, it is possible to perform a permutation test for the significance of the population differentiation represented by the PCA-axes (Goudet, 1999). The test randomises individuals among populations and then recalculates the PCA. The percentage of variance explained by the axis is used as a test statistic.

There are two methods of calculating a PCA. The first uses a matrix of covariances between pairs of variables or allele frequencies, the second uses a matrix of correlations between them. Which method is most suitable depends on the data. If all variables have been measured on the same scale (such as is the case for allele frequencies, which are always bound between zero and one) a

covariance matrix is thought to be most suitable. When variables are measured on different scales (for example different ecological variables such as pH, soil moisture, and temperature), it is generally better to use a correlation matrix as this means that the variables are standardised to unit mean and variance. However, a correlation matrix can also be useful with allele frequencies as it tends to put more weight on rare alleles.

When there are a lot of missing values in a dataset, a PCA can be biased as individuals with missing data tend to be grouped together. Therefore, it is advised to first replace the missing data with random values (see Fill in Missing Data) before performing the PCA.

How to perform a Principal Components Analysis:

1. Make sure that the frontmost window contains either genetic marker data or ecological data.
2. Select Principal Components... from the Analysis menu.
3. Choose whether you want to perform the analysis using a matrix of the covariances or the correlations between pairs of allele frequencies or variables
4. If you are working with genetic data, choose whether you want to perform the analysis on individuals or on populations.
5. When you perform the analysis on populations, it is possible to perform a permutation test whether the axes present significant among-population differentiation. Switch on the "Test for significance of axes" button, fill in the number of permutation and choose whether you want the full results of the permutation test.
6. For polyploid data it is possible to correct the allele frequencies for the unknown dosage of alleles.
7. Select the output options. You can choose to create new ecological documents containing the axis-scores (the "new and improved" variables) for all observations, and/or the loadings (the weights that the old variables are given for constructing the axes).
8. Click Calculate.

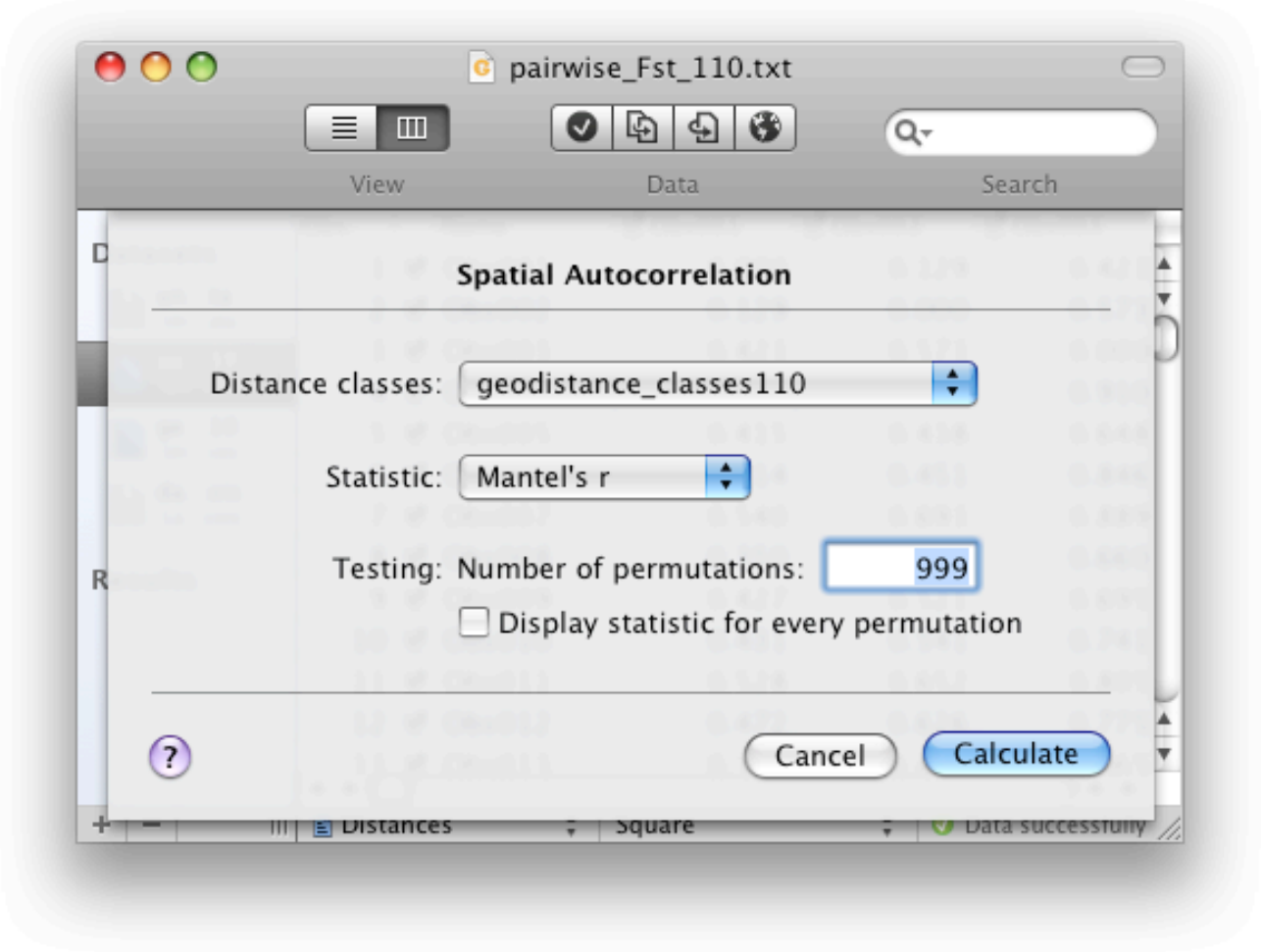
Spatial Autocorrelation

Spatial Autocorrelation is the often-observed phenomenon that the similarity between observations depends on the geographical distance between them. A well-known example from population genetics is so-called Isolation-by-distance, where populations that are geographically close are genetically more similar than populations that are more distant. The same mechanism can work within populations with individuals that grow/live close together being genetically more similar.

The test makes use of a matrix of distance classes, see here for information on how to construct a matrix of distance classes. The output of the test is a so-called correlogram, where for every distance class the value of the test statistic is given. A typical correlogram has high values for the distance classes that correspond to short distances, and low values for the classes that correspond to longer distances. However, the shape of the correlogram depends on the chosen test-statistic and the data that is used as input. Significance testing is performed separately for every distance class. Testing takes place through permutation, by randomising the locations of the observations.

The way GenoDive performs a test for spatial autocorrelation depends on the type of data:

- **Genetic data.** A correlogram is calculated over individuals for every allele that is present in the included loci. The test statistic, either Moran's I or Geary's C, is calculated using the within individual allele frequency (see also Hardy & Vekemans, 1999). The matrix of distance classes should therefore be based upon the distances between individuals rather than between populations. To test for spatial autocorrelation between populations, or for a multilocus test of spatial autocorrelation on individuals, first calculate genetic distances between them, and perform the test for spatial autocorrelation on those distances (see below).
- **Distance matrix.** A Mantel correlogram (Sokal, 1986) is used. The method performs a standard Mantel test for every distance class, using a different model matrix for every distance class. This model matrix contains a one where a distance point lies inside that distance class and a zero otherwise. As a test statistic, there is the choice between Mantel's r, Spearman's r, and the average value of the target distance within a distance class. So if the genetic distance matrix contains Fst-values and the upper limit of the first distance class is 1 km, the statistic gives the average Fst value for populations that are separated by less than one km.
- **Ecological data.** A correlogram is calculated for every included variable. As a test statistic, either Moran's I or Geary's C can be used.



How to test for spatial autocorrelation:

1. Make sure you have created a matrix with distance classes.
2. Select the dataset on which you want to perform the test from the sourcelist on the left.
3. Select Spatial Autocorrelation... from the Analysis menu.
4. Select the matrix of distance classes from the popup menu labeled Distance classes.
5. Select the test statistic, see above for which statistics are available for the different types of data.
6. Type in the number of permutations to test the significance.
7. If you would like to inspect the null-distribution of the test statistic, switch on the "Display statistic for every permutation" button. This is not recommended for a very large number of permutations as that will make the Result file overly large.
8. Click Calculate.

Miscellaneous

How to cite GenoDive:

This version of GenoDive is a major overhaul of the first version, to such an extent that I might consider writing a new paper about it. In the mean time you can cite GenoDive as follows:

- Meirmans, P.G., and P.H. Van Tienderen: (2004), GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms, Molecular Ecology Notes 4 p.792-794.

Most of the analyses have been described in papers written by other people than me; please remember to cite these papers as well.

Requirements

Though GenoDive should be able to run just fine on any reasonable modern Mac, there are some specific requirements.

What do you need to run GenoDive:

- The minimum version of Mac OS X that GenoDive can run on is OS 10.4 (Tiger). Given that Tiger is now more than five years old, I hope that this is not much of a problem to anyone.
- The amount of RAM-memory required depends mostly on your datasets. Especially distance matrices tend to become fairly large, so if you have a computer with only a small amount of RAM, you may not be able to handle extremely large matrices. That said, my iMac with 1 GB of memory was able to calculate distances between almost 10.000 individuals, resulting in a matrix of close to 100 million distances (equivalent to 18.657 pages in Word). I also successfully performed AMOVA's with close to 100.000 individuals, though this took a lot of time. In most cases, GenoDive will give a warning when an analysis will take more memory than available.
- The maximum ploidy level of codominantly scored marker data is 16 (hexadecaploid). As scoring markers codominantly is very difficult for any ploidy level higher than two, I doubt that this restriction is problematic to anyone.
- The maximum ploidy level for the Hybrid Index analysis is 8 (octaploid). The programming gets more difficult for higher ploidy levels, and the algorithm will take very long even for small datasets, so I didn't bother to implement support for higher ploidy levels than octaploid.
- GenoDive is compiled as a so-called "universal binary" which means that it can run both on PPC and on Intel-based Macs. Unfortunately, I do not have any PPC-based Macs available, so most testing has been done on Intel-based Macs running OS 10.4, 10.5, 10.6 and 10.7

License

Included source code

GenoDive contains source code written by third party authors and kindly made available for general use. The licenses and copyrights of the included source code can be found in the Help files.

Version History

GenoDive 2.0b23:

- Added correction for missing dosage information for polyploids to multiple analyses.
- Filling of missing data can now also correct the dosage of polyploids.
- Added stratified Mantel test, removed constrained Mantel test.
- Better calculation of Smouse & Peakall distances.
- Fixed crashing bug in opening of Results.
- Fixed crashing bug in calculation of distance classes.
- Fixed bug in results of clone assignment.
- Principal Components Analysis now uses Gst instead of rho.
- Testing clonal propagation now works when replacing missing data with average allele lengths.
- K-means clustering no longer flags best clustering according to AIC.
- Now less stringent in requiring enough free memory to calculate distances.

GenoDive 2.0b22:

- Genetic diversity now has bootstrapped confidence intervals that do not just consist of zeroes.
- Corrected eror in calculation of heterozygosity for polyploid populations of small size.
- Now automaticaly parses Results files on opening, for better compatibility with Lion.
- Added calculation of confidence intervals using bootstrapping for AMOVA and G-statistics.
- AMOVA no longer removes individuals with missing data, but fills the missing data with random alleles.
- Calculation of genetic diversity now works better for datasets with lots of missing data.
- The number of decimals in the results can now be set by the user.
- Fixed bug in special include of populations.
- Correct writing of individual names in Genetix format.
- Fixed small bug in interface of Transformation.
- Better error checking in reading of population names.

GenoDive 2.0b21:

- Improved handling of missing data in K-means (frequency-based), PCA, spatial autocorrelation, and some distance calculations.
- Introduced multiple new methods of redefining populations in the transformation of marker data.
- Added support for data files in "Convert"-format.
- Fixed bug in calculation of some between-individual distances.
- Fixed popup menu in Import and Convert windows.
- Removed cruft.

GenoDive 2.0b20:

- Added calculation of genetic diversity.
- Added per-population testing to assignment test.
- Fixed error in Monte Carlo permutations of assignment test.
- Fixed bug in making clone corrected dataset using Special Include.
- Added Nei's Gst, Meirmans & Hedrick's G''st, and Jost's D as pairwise distances.
- GenoDive now correctly gives the p-value for negative values of Gis.
- Demoted "Matrix with Residuals" from separate analysis to a part of the Mantel test.

- Updated toolbar; turned "Analysis" item into "Recent analyses".
- Fixed bug in checkboxes to include or exclude loci.
- Fixed bug that could lead to a crash after saving results.
- Fixed crashing bug in Clonal Diversity.
- Subtly changed adaptive starting temperature in K-means' simulated annealing.
- Fixed small interface bug in Inspector tables.
- Changed all Phi_{st} into F_{st} for generality and consistency with the literature.
- Removed cruft.

GenoDive 2.0b19:

- Fixed error in calculation of AMOVA for datasets with mixed ploidy levels.
- Disabled permutation tests of AMOVA with mixed ploidy levels, due to deep-level bug.
- Fixed error in the calculation of the observed gametic heterozygosity for non-diploid data.
- Assignment test now only assigns individuals to populations of more than 5 individuals.
- Improved testing of Hardy-Weinberg equilibrium; added the possibility of using Gis as statistic.
- Added permutation test of Gis to G-statistics.
- Reduced memory consumption of AMOVA-based clustering of populations.
- Major internal restructuring of analyses code.
- GenoDive can now read ecological data with "nan" values (such as can be produced by GenoDive).
- Now correctly saves a selection of the results.
- Can now add new populations by changing individuals' population numbers in Matrix View.
- Fixed bug that occurred when converting polyploid data to a new diploid format.
- Added population and locus numbers to inspector.
- Removed cruft.

GenoDive 2.0b18:

- Introduced calculation of G-statistics, including Nei's Gst, Hedrick's G'st, and Jost's D.
- Changed main Help page for easier access to most important topics.
- Removed support for old style preferences (pre 2.0b13).
- Fixed bug in Hybrid Index that occurred when a population only had missing data at a locus.
- Miscellaneous small bug fixes.

GenoDive 2.0b17:

- Changed clone assignment, added more info to the histogram and fixed a crashing bug.
- Fixed bug where the program would sometimes not quit.

GenoDive 2.0b16:

- Improved compatibility with Mac OS 10.6 "Snow Leopard".
- Removed possibility to show splash screen at startup, as this could lead to crashes.
- Fixed bug in reading of GenePop format.
- Internal restructuring of assignment procedures.

GenoDive 2.0b15.1:

- Fixed bug in Geographical distances introduced in 2.0b15

GenoDive 2.0b15:

- New icon, the big orange G is replaced with a diving helmet.

- Added calculation of ploidy independent Rho-statistic to the Amova.
- Amova can now be calculated with individuals at the lowest level, which is e.g. handy for asexuals.
- Amova now also supports the use of custom distance matrices.
- Added Monte Carlo permutation test to Population Assignment.
- Better support for population names in GenePop format.
- Enabled Growl notifications to inform the user that an analysis has finished.
- The source list now stays the same size when the window is resized.
- Fixed bugs in conversion of marker data.
- Fixed bug in permutations for constrained Mantel test.

GenoDive 2.0b14:

- Added calculation of standard error of F-statistics in Amova through jackknifing over loci.
- Amova does not crash anymore when used on haploid datasets; instead the haploid data is converted to diploid homozygotes.
- Kinship coefficient now follows a more appropriate method for dominant data.
- Fixed error in calculation of regression equation for Matrix with Residuals.
- Fixed bug in pairwise Phi_st and Phi'_st calculation with unbalanced population sizes (again).
- Fixed small bug in reading files in GenePop format.
- Fixed error in output of Phi'_st in AMOVA.
- Fixed crash in Spatial Autocorrelation.
- Internal restructuring of analysis dialogs.
- Miscellaneous small bug fixes.

GenoDive 2.0b13:

- Major speed increase in K-means clustering using simulated annealing; made it possible to repeat the algorithm a certain number of times.
- Changed interface for K-means clustering to make better use of speedier simulated annealing.
- Fixed bug in K-means clustering using hill-climbing, is now slightly slower as a result.
- Expanded the options in the transformation of marker data.
- Added calculation of pairwise Phi'_st (Phi_st standardised relative to the within population diversity).
- Added calculation of pairwise Rho, an Fst-analogue that is independent of ploidy and double reduction.
- Fixed bug in pairwise Phi_st calculation with unbalanced population sizes.
- Fixed bug in Population Assignment, where excluded populations were sometimes taken into account.
- Changed output of Population Assignment, now the likelihood of the current population is also displayed.
- New bundle identifier for preferences-file that matches my new homepage http://www.patrickmeirmans.com.
- Fixed problem that occurred when running genoDive for the first time from a freshly downloaded copy.

GenoDive 2.0b12:

- Implemented Shuffle Data: permutation of data using randomisations or bootstraps.
- K-means clustering now also shows Akaike Information Criterion, which can also be calculated for a single cluster (k=1).

- Pairwise Phi_st-values can now correctly be calculated for haploid data.
- Fixed bug in output of clone assignment.
- Miscellaneous small bug fixes.

GenoDive 2.0b11:

- Added two additional example files to show formats of distance and eco data.
- Fixed bug in k-Means UI, introduced in b10.
- Fixed bug in reading of Ecological data with only a row of titles.
- Hopefully fixed drawing error in header of tables with only few columns.
- Created proper license-page in the Help-files.

GenoDive 2.0b10:

- Implemented Principal Components Analysis.
- Implemented filling in of missing data.
- Inspector panel now also shows some info when results are selected
- Mantel Correlograms can now really use Spearman's r (previously, Mantel's r was returned when Spearman's r was selected).
- Population Assignment now always outputs the maximum likelihood value.
- Fixed bug in calculation of overall variance components for AMOVA.
- Fixed very rare bug in reading of Fstat files.
- Fixed drawing glitch that occurred when opening a large number of files.

GenoDive 2.0b9:

- Replaced Mantel Correlogram by Spatial Autocorrelation. Can now also calculate Moran's I and Geary's C on alleles and ecological variables.
- Made interface of Mantel test and Matrix with Residuals more in line with other analyses.
- Improved estimation of number of clusters in AMOVA-based k-Means clustering, by changing the calculation of r-squared.
- Added possibility of copying series of population groups from other documents.
- Fixed bug where user was not always prompted to save results.
- Fixed bug where stopping AMOVA-based clustering could lead to a non-stoppable analysis.
- Updated Help-screenshots: they now all match the current look of the program.

GenoDive 2.0b8:

- Better compatibility with OS 10.5 Leopard.
- Added Population Assignment.
- Expanded help on k-means clustering, now gives more details on the methods.
- Added temperature to full output of results from k-means clustering by simulated annealing.
- Fixed bug in listing of matrices in Mantel Test and Mantel Correlogram where files would just disappear and reappear.
- Fixed minor bug in the k-means clustering of individuals.
- Special Include should now work as it was intended in beta 6.
- Improved parsing of previous results.
- Improved parsing of ecological data.
- More friendly welcome screen for new users.
- Several small interface improvements and bug fixes.

GenoDive 2.0b7:

- Fixed bug in the recoding of alleles within population groups.

GenoDive 2.0b6:

- Implemented renaming of open files by double-clicking their name in the source-list.
- Fixed bugs in Distance Classes option, where the limits were not displayed in the results.
- Fixed bug where allele counts were not always displayed in the results.
- More options in Special Include.
- Changed default items in toolbar.
- Moved pop-up menus for the data type and data format to the statusbar at the bottom of the window.
- No more pinstripes in analyses windows and sheets.
- Made preference setting to control the checking for double names.

GenoDive 2.0b5:

- Removed some diversity indices from bootstrap test due to their bad performance in the test.
- Fixed bug in pairwise differentiation test, where permuted statistics were always zero or undefined.
- When opening an existing results file, GenoDive now can parse the results into the sidebar.
- Implemented copying of selected row in Matrix View.

GenoDive 2.0b4:

- Major overhaul of user interface.

GenoDive 2.0b3:

- Added calculation of Genetic Distances.
- Added testing of Pairwise Differentiation between populations.
- Implemented faster and more robust pseudo random number generator: the awesome Mersenne Twister.
- Streamlined opening of files, including a fancy progressbar for large files.
- Fixed bugs in Transformation of Distances and Ecological data.
- Fixed bug in performing some analyses with existing distance matrices.
- Improved speed of Hardy Weinberg test.

GenoDive 2.0b2:

- Added test for Hardy Weinberg equilibrium.
- Added k-Means clustering of individuals and of ecological data.
- Added Sparkle framework for automatic updating.
- Implemented calculation of Spearman's r for all Mantel tests.
- Added support for Structure and BayesAss+ file formats.
- Added preference to prevent computer from sleeping during analyses.
- Automatically filters out individuals with missing data in AMOVA.
- Fixed bug where converting data lead to empty files.
- Fixed bugs in interface of Allele Frequencies.
- Added references to Help-files, and updated Help-files.

GenoDive 2.0b1:

- First public beta-release.

Cited References:

- Beerli, P., and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics. 152:763-773.
- Belkhir K., P. Borsa, L. Chikhi, N. Raufaste, and F. Bonhomme. 1996-2004 GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier (France).
- Buerkle, C.A. 2005. Maximum-likelihood estimation of a hybrid index based on molecular markers. Molecular Ecology Notes 5:684-687
- Bruvo, R, N.K. Michiels, T.G. D'Souza, and H. Schulenburg. 2004. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. Molecular Ecology 13:2101-2106
- Calinski, R.B., and J. Harabasz. 1974. A dendrite method for cluster analysis. Communications in statistics 3:1-27.
- Cavalli-Sforza, L. L. and W. F. Bodmer. 1971. The Genetics of Human Populations. W. H. Freeman, San Francisco.
- Chao, A., and T. J. Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics 10:429-443.
- Cornuet, J.-M., S. Piry, G. Luikart, A. Estoup, and M. Solignac. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153:1989-2000.
- De Silva, H. N., A. J. Hall, E. Rikkerink, and L. G. Fraser. 2005. Estimation of allele frequencies in polyploids under certain patterns of inheritance. Heredity 95:327–334
- Douhovnikoff, V., and R. S. Dodd. 2003. Intra-clonal variation and a similarity threshold for identification of clones: application to Salix exigua using AFLP molecular markers. Theoretical and Applied Genetics 106:1307-1315.
- Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes - application to human mitochondrial-DNA restriction data. Genetics 131:479-491.
- Glaubitz, J.C. 2004. convert: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. Molecular Ecology Notes 4:309-310.
- Gomez, A., and G.R. Carvalho. 2000. Sex, parthenogenesis and genetic structure of rotifers: microsatellite analysis of contemporary and resting egg bank populations. Molecular Ecology 9:203-214.
- Goudet, J. 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. Journal of Heredity. 86:485-486.
- Goudet, J., M. Raymond, T. De Meeus, and F. Rousset. 1996. Testing differentiation in diploid populations. Genetics 144:1933-1940.
- Goudet, J. 1999. PCA-Gen for Windows. available from: http://www2.unil.ch/popgen/softwares/pcagen.htm
- Halkett, F., J.-C. Simon, and F. Balloux. 2005. Tackling the population genetics of clonal and partially clonal organisms. Trends in Ecology and Evolution 20:194-201.
- Hardy O.J., Vekemans X. 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. Heredity 83:145–154.

- Hardy, O. J., and X. Vekemans. 2002. SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes. 2:618-620.
- Hardy O.J. 2003. Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. Molecular Ecology 12:1577-1588.
- Hedrick, P.W. 2005. A standardized genetic differentiation measure. Evolution 59:1633-1638.
- Jost, L. 2008. GST and its relatives do not measure differentiation. Molecular Ecology 17:4015-4026
- Legendre, P., and L. Legendre. 1998. Numerical Ecology. Elsevier, Amsterdam.
- Loiselle, B. A., V. L. Sork, J. Nason, and C. Graham. 1995. Spatial genetic-structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). American Journal of Botany 82:1420-1425.
- MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 :281-297. Univ. of Calif. Press. Berkeley
- Manly, B. F. J. 1991. Randomization and Monte Carlo methods in biology. Chapman & Hall, London.
- Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27:209-220.
- Meirmans, P.G. 2006. Using the AMOVA framework to estimate a standardized genetic differentiation measure. Evolution 60:2399-2402.
- Meirmans, P.G., and P. H. Van Tienderen. 2004. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. Molecular Ecology Notes 4:792-794.
- Meirmans, P.G., and P.W. Hedrick. 2011. Measuring differentiation: G_{st} and related statistics. Molecular Ecology Resources 11:5-18.
- Meirmans, P.G. 2012. AMOVA-based clustering of population genetic data. Journal of Heredity 103:744-750.
- Meirmans, P.G. 2012. The trouble with Isolation by Distance. Molecular Ecology 21:2839-2846.
- Michalakis, Y., and L. Excoffier. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061-1064.
- Moody, M.E., L.D. Mueller, and D.E. Soltis. 2003. Genetic-variation and random drift in autotetraploid populations. Genetics 134:649-657.
- Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583-590.
- Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.
- Paetkau D, W. Calvert, I. Stirling, and C. Strobeck. 1995. Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology 4:347-354.
- Paetkau, D, R. Slade, M. Burdens, and A. Estoup. 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.
- Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.
- Raymond, M., and F. Rousset. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. Journal of Heredity 86:248-249
- Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution 43:223-225.

- Rogers, J.S. Measures of genetic similarity and genetic distance. In: Studies in Genetics VII. University of Texas Publication 7213, Austin.
- Rogstad, S. H., B. Keane, and J. Beresh. 2002. Genetic variation across VNTR loci in central North American Taraxacum surveyed at different spatial scales. Plant Ecology 161:111-121.
- Ronfort, J., E. Jenczewski, T. Bataillon, and F. Rousset. 1998. Analysis of population structure in autotetraploid species. Genetics 150:921-930.
- Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F- statistics under isolation by distance. Genetics 145:1219-1228.
- Slatkin, M. 1995. A Measure of Population Subdivision Based on Microsatellite Allele Frequencies. Genetics 139:457-462
- Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Systematic Zoology 35:627-632.
- Smouse, P. E., and R. Peakall. 1999. Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity 82:561-573.
- Sokal, R. R. 1986. Spatial data analysis and historical processes in E.E.A. Diday, ed. Data analysis and informatics, IV. North-Holland, Amsterdam.
- Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6:461–464.
- Weir, B. S., and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.
- Wilson, G., and B. Rannala. 2003. Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163:1177-1191.