# GDA User's Manual

Paul O. Lewis and Dmitri Zaykin

Last updated January 7, 2002

## No Warranty

GDA is distributed AS IS with the hope that it will be useful. You alone are responsible for any results obtained using GDA. GDA comes with absolutely NO WARRANTY.

## This version will not expire

Previous GDA versions expired (quit working) after a certain date. This was to prevent their use when newer versions (containing fewer bugs) were available. GDA still has plenty of glitches, but there don't seem to be major problems any more, so this version will not expire and does not begin with the usual warning about asking the authors' permission before publishing. You should still check the web site to see if a newer version has been posted, of course. This version (1.1) is identical to version 1.0 (d16c) except for its immortality.

## GDA Availability

GDA is a free program and can be obtained most easily through the "Software" section of the web site located at

`http://lewis.eeb.uconn.edu/lewishome/`

Note that GDA now has two menu items (*Help* | *Go to GDA home page...* and *Help* | *Go to GDA download page...*) that start up your default web browser and take you to the GDA Home Page or the software download page.

## Authors' Contact Information

Paul O. Lewis
Department of Ecology and
Evolutionary Biology
The University of Connecticut
75 North Eagleville Road Unit 3043
Storrs, CT 06269-3043
**Tel:** (860) 486-2069
**Fax:** (860) 486-6364
**Email:** `paul.lewis@uconn.edu`

Dmitri Zaykin
Statistical Genetics, MAI.A1112
Glaxo Wellcome, Five Moore Drive
Research Triangle Park, NC 27709
**Tel:** (919) 483-9391
**Email:** `zaykin@statgen.ncsu.edu`

# Contents

# Data File Structure

GDA uses a data file format known as NEXUS [1]. This format was originally created by David and Wayne Maddison (authors of MacClade) and David Swofford (author of PAUP*) for inputting discrete morphological and sequence data for phylogenetic analyses. It has been adopted for use with GDA because of its flexibility and the fact that information can be readily passed between programs supporting this format. One of the goals of the NEXUS format was that it allow researchers to store all of the data for a particular project in the NEXUS data file and yet make it possible to easily choose subsets of the data for particular analyses. This has the advantage of discouraging "tampering" with the data set in order to analyze different parts; too much cutting and pasting eventually leads to the introduction of errors into the data matrix. Thus, the format allows for sections of the matrix to be "commented out" in various ways. A comment in the NEXUS format is a section of the file enclosed in square brackets. Comments will be output to the screen if the next character following the opening square bracket is an exclamation point. For example,

```
[! this is a comment that would appear in gda's output].
```

Particular programs can make it especially easy to analyze subsets of the data. For example, GDA allows you to interactively exclude loci or populations from particular analyses. GDA also allows you to place commands within a "GDA block" in the data file itself, so that in the future you can have a complete record of the series of analyses performed for a given study.

Data files conforming to the NEXUS standard comprise units known as **blocks**. Each block begins with the keyword **begin** and ends with the keyword **end**. Each type of block has its own name (example block names are **gdadata**, **distances**, and **trees**). One feature of the NEXUS data file standard is that the default behavior for a program is to skip over a block if the block's name is not recognized. This provides yet another way to make sections of the data file invisible to GDA; simply add an underscore character (for example) to the block name (e.g., **_gdadata**) and GDA will skip over it as being an unrecognized block type. This feature also allows very different types of data to be combined in a single file. For example, PAUP* does not recognize the **gdadata** block type used by GDA, but it does recognize a **distances** block. Thus, provided with a data file containing both a **gdadata** block and a **distances** block, PAUP* would ignore the **gdadata** block but still read the **distances** block and be able to perform analyses on the pairwise distances contained therein.

## 1.1 Typographical Conventions

The following conventions are employed in both the following sections as well as the accompanying command reference.

**Bold** is used to introduce new terms, to refer to block and command names in the NEXUS file format, and to refer to GDA command names.

*Italic* is used for file and directory names and for occasional emphasis.

Sans Serif is used for names of computer programs (such as GDA and PAUP*) and for menu choices (for example, File | Open...).

`Constant Width` is used in examples to show the contents of files or the output from commands.

## 1.2 GDADATA Block

An example of a NEXUS file containing a **gdadata** block is provided below. This is the primary type of block used with GDA, as it is designed for storing discrete allelic genetic data.

```
#nexus

begin gdadata; [comments are surrounded by square brackets]
  dimensions npops=2 nloci=3;
  format missing=? separator=/;
  locusallelelabels
    1 ’pgi 1’,
    2 ’pgi 2’,
    3  adh / slow fast
  ;
  matrix
    Embudo:
      indiv_1 A/A 100/110 slow/fast
      indiv_2 A/A 75 / 90 slow/slow
      indiv_3 A/a 75/100  fast/Slow
      indiv_4 A/A 100/100 fast/fast,
    Black_Mesa:
      1 a/a 110/100 fast/slow
      2 a/A  75/100 slow/solw
      3 a/a 100/100 fast/fast
  ;
end;
```

Several important points can be made about the NEXUS format and specifically the **gdadata** block using this contrived example:

- NEXUS files are *plain text* files, which means they can be viewed by any program that is capable of displaying standard ASCII text files. NEXUS data files may be created using any word processor, although care must be taken in saving the file so that no formatting information is included (i.e., choose file types such as MS-DOS text, ASCII text, or plain text in the Save As... dialog box of the word processor). Saving the data file as a Microsoft Word document (for example) will make it uninterpretable to GDA. A good check: if the file is readable with the Windows NOTEPAD.EXE program, then it will be readable by GDA.

- NEXUS files are *free-format*, which means that the entire file could conceivably consist of a single, long line of text. It does not matter to GDA where you break lines (as long as you don't split up a keyword or the name of a locus, allele or population), nor does it matter to GDA if you use one space or a dozen spaces to isolate each individual word (**token**) in the file. Tokens may be casually defined as sequences of characters separated by **whitespace** (e.g., spaces, carriage returns, line feeds, tabs, etc.). Exceptions to this definition are common, however. Special punctuation characters may also delineate adjacent tokens in the absence of whitespace; for example, in the genotype designation `A/a` (locus `'pgi 1'` in individual `indiv_2` of the population `Embudo`, above) there are three distinct tokens, because the forward slash (‘`/`’) character serves as a single-character token. Other common single-character tokens include the equal sign (‘`=`’), semicolons (‘`;`’), colons(‘`:`’), and commas (‘`,`’). Tokens may also contain whitespace! If a token begins with a single or double quote character, then every character until the next, matching quote character will be treated as a single token. This is useful as a means of putting blank spaces inside population or locus labels. If you want a population label, for example, to contain an embedded space, you simply enclose it in single quotes. This was done for the locus names '`pgi 1`' and '`pgi 2`' in the example above. An alternative approach is to use the underscore character where you want the space to appear. The underscore character approach was used in the labels for individuals in the Embudo population above (e.g., `indiv_1`). Underscores show up as blank spaces when GDA reports results.

- NEXUS blocks are made up of **commands**, which always end in a semicolon (‘`;`’). The **gdadata** block above comprises a **begin** command, a **dimensions** command, a **format** command, a **locusallelelabels** command, a **matrix** command, and an **end** command.

- NEXUS files are for the most part not case-sensitive by default. The biggest exception to this is in the **matrix** command in the **gdadata** block, where (by default) an allele named ‘`A`’ is treated as being distinct from ‘`a`’. The case-insensitivity in the rest of the data file means that you can, for example, use `gdadata`, `Gdadata`, or `GDADATA` for the block name and GDA will not complain. The number of populations and loci must be specified in the **dimensions** command, but not the number of individuals per population. The end of the data for one population is signified by either a comma or the semicolon indicating the end of the matrix command itself.

- The **locusallelelabels** command provides a way to give names to loci. It is optional: loci will simply be numbered beginning with 1 if this command is absent. The **locusallelelabels** command also provides a way to have GDA do some error checking for you. Notice how, for the locus `adh`, two allele names are provided (`slow` and `fast`). If in the matrix section GDA finds an allele name other than either `slow` or `fast` for locus `adh`, it will stop parsing the data file and report it as an error. Thus, GDA would find two errors in the sample data file above: the first is `Slow` (individual indiv_3 of population `Embudo`) and the second is `solw` (individual 2 of the `Black Mesa` population). If valid allele names are not provided in a **locusallelelabels** command, GDA simply interprets each distinct allele name as a separate allele. Were it not for the **locusallelelabels** command, the above example data set would be interpreted as having four alleles (not two) for the `adh` locus!

### 1.2.1 Haploid Data

There are two ways to include haploid loci in GDA. If all loci are haploid, simply include the keyword **haploid** in the **format** command. See the example data file *haploid.nex* for an example of an all-haploid data file. Note that no separator characters should appear inside the matrix in this case. The second method applies if you wish to include a mixture of haploid and diploid data in the same matrix in a **gdadata** block. In this case, the **hapset** command is used to specify which loci are haploid (diploid is the default). Thus, if there are 8 loci and the last three are haploid, the **gdadata** block would begin something like this:

```
#NEXUS

[Note: first 5 loci are diploid and last 3 are haploid]

begin gdadata;
  dimensions nloci=8 npops=6;
  format tokens labels missing=? datapoint=standard;
  hapset 6-8;
  locusallelelabels
    1 ’dip 1’,
    2 ’dip 2’,
    3 ’dip 3’,
    4 ’dip 4’,
    5 ’dip 5’,
    6 ’hap 1’,
    7 ’hap 2’,
    8 ’hap 3’
  ;
  matrix
    Pop1:
      indiv1 4/4 4/3 4/3 3/3 4/4 3 3 4
      indiv2 4/4 4/4 4/3 3/3 4/4 3 3 4
      .
      .
      .
```

## 1.3 TREES Block

Another type of block used extensively by GDA (as well as phylogenetic analysis programs such as PAUP* and MacClade is the **trees** block. GDA uses a **trees** block to specify the hierarchy to be assumed when estimating F-statistics. An example of such a block is shown below:

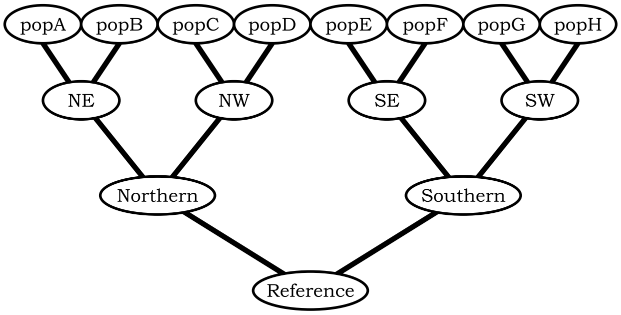
```
begin trees;
  tree two = (popA,popB,popC,popD,popE,popF,popG,popH);
  tree * three = ((popA,popB,popC),(popD,popE,popF,popG,popH));
  tree four = (((popA,popB),(popC,popD)),((popE,popF),(popG,popH)));
end;
```

Three tree descriptions are specified in this one **trees** block. The block contains three **tree** commands (note the semicolons terminating each of these commands). Each **tree** command specifies a name, or label, for the tree as well as the tree's structure. The tree named `two` specifies a basic two-level hierarchy. In this case, the 8 populations are all on the same hierarchical level, with the hypothetical reference population (ancestral to these 8 populations) making up the other level of the hierarchy. The population names `popA`, `popB`, etc., are assumed to have been provided in a preceding **gdadata** block. The tree named `three` specifies a three-level hierarchy consisting of subpopulations within populations. Regardless of the number of levels in the hierarchy tree, the populations delineated in the gdadata block form the lowest level. Thus, in the case of tree three, 8 "populations" were defined and named in the gdadata block, even though they are considered subpopulations in the analysis. The first population defined by tree three consists of the 3 subpopulations `popA`, `popB`, and `popC`; the second population comprises `popD`, `popE`, `popF`, `popG`, and `popH`. The nesting created by parentheses in the tree description defines the hierarchy of subpopulations within populations. Tree `four` defines a balanced four-level hierarchy in which each population has two subpopulations, and each of these subpopulations has two subsubpopulations.

It is possible to assign names to the higher-level groups in the hierarchy, as illustrated in the following alternative specification for tree four:

```
tree four = (((popA, popB)NE, (popC, popD)NW)Northern, ((popE, popF)SE,
  (popG, popH)SW)Southern)Reference;
```

The four-level hierarchy defined by the above tree command is represented graphically below



The hierarchy tree to be loaded when GDA reads the data file is indicated by an asterisk between the command name **tree** and the label of the tree definition. For example, although the definitions of 3 hierarchy trees are given in the **trees** block on the previous page, the tree named `three` (representing the three-level hierarchy) will be the one used in any analyses once the **trees** block has been read. The GDA command **hierarchy** can be used to switch to another predefined hierarchy tree while GDA is running (the **hierarchy** command is described below in the command reference section). A default hierarchy tree is created if no **trees** block is encountered in the NEXUS data file. This default tree represents the basic two-level hierarchy and is named `default`.

*Important: trees defined for purposes of specifying a population hierarchy must have at least two descendant nodes for each ancestral node defined. Thus, in the tree named* `four`*, one could not leave out population* `popD` *since this would leave the node named* `NW` *with only one descendant.*

## 1.4 GDA Block

The final block type discussed is the **gda** block. This is a simple block consisting of GDA commands (see below) of the same sort as one might type in at the command line of the program. The **gda** block allows you to store a complete record of an analysis right in the data file itself. The following simple example opens a log file, reads data from a NEXUS data file, estimates F-statistics, and then closes the log file:

```
begin gda;
  log file=mylog.txt replace;
  fstats ms vc coancestry indivalleles;
  log stop;
end;
```

Note that the **gda** block above cannot be the only block present in a NEXUS data file; a **gdadata** block must precede it in order for any analyses to be possible.

Details of the file format (other than the **gdadata** and **gda** blocks) may be found in the following paper:

> Maddison, David R., Swofford, David L. and Maddison, Wayne P. 1997. NEXUS: an extensible file format for systematic information. *Systematic Biology* **46**: 590-621

# Command Reference

## 2.1 General Notes About Commands

Commands must be correctly spelled. Some abbreviation is possible. For example, the command **execute** may be abbreviated **exe**, but using the abbreviation **ex** is not allowed since it would conflict with another command, namely **exact**. Several commands (e.g., **stats**, **fstats**, **gdist**, and **exact**) do nothing if they are entered without additional information. For these commands, entering the command followed by a space and then a question mark ('?') will provide you with the current settings of variables associated with that command. To initiate an analysis, type the name of the command followed by the word **estimate**.

Some of the commands are not available all of the time. For example, you are not allowed to use the **gdist** command unless either raw allelic data or a matrix of pairwise genetic distances has been read into memory. The **help** command can be used at any time to find out which commands are currently available.

The menu choices necessary to initiate the command in the graphical Windows versions are given below the command name. Vertical lines ('|') are used to separate menu items from submenu items. Thus, *File* | *Save* means "*click on the word 'File' in the main menu, then on the word 'Save' in the submenu that pops up.*" If no menu choice is listed, the command can only be run from the command line.

## 2.2 Commands Only Available From Menu

Most GDA commands are available from either the command line or the menu system. Some Windows-specific commands are only available from the menu system. These include:

***File*** | ***Editor*** This menu command invokes the editor application whose path and filename is specified in the *Misc* | *Preferences...* dialog box. If a NEXUS data file has been opened, this file will be displayed in the editor once it has started.

***Dist*** | ***Invoke TreeView*** This menu command invokes the TreeView application whose path and filename should be specified in the *Misc* | *Preferences...* dialog box. This command is unconditionally grayed out in the 16-bit version of GDA because TreeView is a 32-bit Windows application.

***Dist*** | ***Invoke Paup*** This menu command invokes the PAUP* application whose path and filename should be specified in the *Misc* | *Preferences...* dialog box. This command is unconditionally grayed out in the 16-bit version of GDA because PAUP* is a 32-bit Windows application.

***Help*** | ***How to cite GDA*** Pops up a dialog box explaining how to cite the GDA program in the event that you publish results obtained using GDA. The copy button in this dialog will copy the recommended citation to the Windows clipboard. This dialog has been available for some time, but has been buried in the *Help* | *About...* dialog box, where it was easily overlooked.

***Help*** | ***PDF Documentation...*** Starts the default PDF viewer application (or the one specified in the *Misc* | *Preferences...* dialog box) and opens the *gdahelp.pdf* file containing this documentation.

***Help*** | ***Go to GDA home page...*** Starts the default web browser and attempts to load the URL for the GDA home page.

***Help*** | ***Go to GDA download page...*** Starts the default web browser and attempts to load the URL for the GDA download page.

Also, note the section under the **set** (below) concerning options available in the dialog box corresponding to this command that are not available using the typed command.

## 2.3 Bootloci Command

*Windows version menu choice: Fstats* | *Bootstrap across loci*

The **bootloci** command is used to bootstrap over loci for the purpose of obtaining confidence intervals for F-statistics. To perform 1000 bootstrap resamplings using a pseudorandom number seed of 1234567 and constructing a 95% confidence interval, use:

```
bootloci nreps=1000 rseed=1234567 ci=95;
```

where:

- **nreps** is the number of bootstrap replicates to generate
- **rseed** is the random number seed
- **ci** is the nominal confidence interval

The CI specified is called the *nominal* confidence interval because it may not be possible (given the number of replicates requested) to get the exact CI requested. For example, if only 3 bootstrap replicates were to be generated, it would not be possible to construct a 95% confidence interval exactly. The precise confidence interval obtained is printed out in the output. GDA tries to get as close as possible to the nominal (requested) CI, but is conservative if this goal cannot be met exactly.

Typing the command

```
bootloci ?
```

shows the current settings for the **bootloci** command.

## 2.4 Briefreport Command

*Windows version menu choice:* none

Typing **briefreport** will give a brief description of every type of data currently stored in memory. For example, if the program had just read a data file containing a **gdadata** block, the command **briefreport** would produce output similar to the following:

```
Data matrix has 6 populations, 5 loci, and 44 individuals
  Missing data represented by the symbol ?
  Different genes at one locus separated by the symbol /
  Multicharacter allele names.
  Labels provided for individuals.
  Respecting case for allele names.
  Data matrix not interleaved.
  All loci are diploid.
```

## 2.5 Exact Command

*Windows version menu choice: Diseq | Options... | Exact tests*

The command **exact** is used to initiate shuffling tests for linkage and Hardy-Weinberg disequilibrium [7]. This analysis includes tests for Hardy-Weinberg disequilibrium at individual loci, tests for disequilibrium of any kind for all possible pairs of loci, triplets of loci, etc., and can also test for disequilibrium (of any kind) for all loci considered jointly. The following descriptions illustrate additional keywords and options available.

- **nruns** specifies the number of shufflings to be done for each test
- **missings** can be either **infer** (missing data points will be substituted with a data point chosen to reflect the frequencies of alleles in the remainder of the data set) or **discard** (default).
- **subsets** causes single loci as well as pairs, triplets, etc., to be tested for disequilibrium
- **nosubsets** causes the program to skip directly to testing for disequilibrium at all loci simultaneously
- **upto** specifies the maximum number of loci to test jointly for disequilibrium (e.g., if `upto=3` were specified, single loci, pairs of loci, and triplets of loci would be tested, but not quadruplets of loci or higher combinations)
- **onlyhets** tells the program to base comparisons on the absolute value of heterozygosity excess $(H_o - H_e)/H_e$, averaged over loci, where $H_o$ is observed heterozygosity and $H_e$ is expected heterozygosity

- **noonlyhets** turns off the **onlyhets** option
- **fisher** forces a success to be defined as the event in which a particular shuffling results in a data set (multilocus table) that is as probable or less probable than the original data
- **chisquare** forces a success to be defined as the event in which a particular shuffling results in a more extreme chi-square value than that produced by the original data
- **permutemethod** specifies either the letter **b** or **p** for each locus. If **b** is specified for a particular locus, the genotypes at that locus will be broken up by the shuffling process. If **p** is specified, however, genotypes will be preserved but associations of genotypes across loci will be destroyed. If there are two loci, the first equation at the top of p. 128 in [4] corresponds to `permutemethods=pp`, and the second equation corresponds to `permutemethods=bb`. If no **permutemethod** statement is specified, **b** is assumed to apply to all loci

NOTE: All shuffling is done in the computer's memory; GDA never changes the original data file in any way.

## 2.6 Execute Command

*Windows version menu choice: File | Execute*

The **execute** command is used to specify a data file to be interpreted. For example, to cause GDA to read the file *diploid.nex* and store its contents in memory, you could use the following command:

```
execute file=diploid.nex
```

If you need to input a file from a directory other than the current directory (where the GDA program is located), you should surround the file name with double quotes, as follows:

```
execute file="mydir/diploid.nex"
```

## 2.7 Fstats Command

*Windows version menu choice: Fstats | Options...*

The **fstats** command is used to obtain estimates of parameters related to population structure. To obtain all the output possible, you could use the following command:

```
fstats estimate ss ms vc coancestry noassumehw
```

where:

- **ss** refers to the components making up sums of squares ($C$, $S_0$, $S_1$, etc.)
- **ms** refers to the mean squares ($MSG$, $MSI$, $MSP$, etc.)

- **vc** refers to the variance components ($\sigma^2_G$, $\sigma^2_I$, $\sigma^2_P$, etc.)
- **coancestry** refers to the coancestry coefficients ($\theta_P$, $\theta_S$, $\theta_{SS}$, etc.)
- **noassumehw** means "do not assume Hardy-Weinberg proportions" (i.e., estimate inbreeding the coefficients $F$ and $f$)

Note that to obtain estimates of the inbreeding coefficients $F$ and $f$, you should specify **noassumehw**; **assumehw** forces $f = 0$ and $F = \theta$. For one population, it is not possible to estimate coancestry coefficients or $F$. For several populations in a simple two-level hierarchy (i.e., only populations and no subpopulations), only one coancestry coefficient ($\theta$) can be estimated, which is called `Theta-P` in GDA's output. $\theta_S$ (`Theta-S`) and $\theta_{SS}$ (`Theta-SS`) can be estimated for three-level and four-level hierarchies, respectively. Results can be output for individual alleles at each locus by using the keyword **indivalleles**. As for the **stats** command, issuing just

```
fstats ?
```

provides a listing of which statistics are currently turned on and which are currently turned off.

## 2.8 Gdist Command

*Windows version menu choice: Gdist | Options...*

The **gdist** command is used to obtain a matrix of pairwise estimates of genetic distance. This matrix can be saved in the form of a NEXUS file, which can then be read (for example) by PAUP* for the purpose of searching for the optimal phylogenetic tree. Trees may be obtained by simple cluster analysis (UPGMA or neighbor-joining methods) without leaving GDA. To obtain a distance matrix based on pairwise estimates of $\theta$ (i.e., $F_{ST}$), then build a neighbor-joining tree from that distance matrix, the following command could be used:

```
gdist estimate above=none below=driftdist showmatrix showphenogram
  cluster=nj
```

If you are not interested in outputting the matrix, you could use the **noshowmatrix** keyword, and likewise, if you wanted to avoid building the tree, you could use the **noshowpenogram** keyword. You might also wish to include the keyword **nouselinedraw**, as this will force GDA to use normal ASCII characters to show the tree rather than the special line drawing characters which only look good in an MS-DOS window.

| Option | Choices |
|---|---|
| **above** | **none**, **driftiden**, **nei72iden**, and **nei78iden** |
| **below** | **none**, **driftdist**, **nei72dist**, and **nei78dist** |
| **cluster** | **nj** and **upgma** |

## 2.9 Help Command

*Windows version menu choice:* none

Typing this command at the GDA prompt shows a list of all commands currently available. Note that commands dependent on the presence of certain types of data will not appear unless that type of data has been stored. For example, the **fstats** command is not available unless a **gdadata** block has been read from a NEXUS file and stored in memory.

## 2.10 Hierarchy Command

*Windows version menu choice: Hierarchy | Select*

The **hierarchy** command can be used to change hierarchy trees while GDA is running. For example, if several trees are defined in a **trees** block in the data file, one of which is named `threelevel`, the following command can be used to select `threelevel` as the current hierarchy tree:

```
hierarchy use=threelevel;
```

## 2.11 Ldiseq Command

*Windows version menu choice:*
*Diseq | Options... | Analysis type: Linkage disequilibrium*

The **ldiseq** command estimates linkage disequilibrium coefficients for pairs of loci. The following options are available for this command:

- **assumehw** forces the analysis to assume Hardy-Weinberg genotypic proportions
- **noassumehw** Hardy-Weinberg genotypic proportions will not be assumed
- **collapse** causes all alleles at a given locus except the most common allele to be considered to be a single allelic class
- **nocollapse** should be used if you do not wish to combine the less-frequent alleles.

The analysis performed when **assumehw** is in effect is identical to that performed by the program LD79 in the back of [3]. This analysis calculates the gametic linkage disequilibrium for two loci each with two alleles when Hardy-Weinberg equilibrium is assumed. The assumption of HWE allows one to get around the need for distinguishing between the two types of double heterozygotes. This analysis is described in [3, pp. 64-65, 310], as well as in [5].

The analysis performed when **noassumehw** is in effect is essentially that performed by the program LD86 in [3, pp. 102-103, 323]. This computes the composite linkage disequilibrium coefficient, which gets around the problem of the two forms of heterozygotes by summing the digenic frequencies $p_{AB}$ and $p_{A/B}$. The composite disequilibrium coefficient is then defined as $D_{AB} = p_{AB} + p_{A/B} - 2p_A p_B$. This solution avoids an assumption of Hardy-Weinberg equilibrium. This analysis is also discussed by [6].

## 2.12 Log Command

*Windows version menu choice: File | Log...*

The **log** command starts or stops GDA from logging output to a disk file. The output generated by GDA is not automatically saved to disk. To begin logging output to the file *myanalysis.log*, use the construct

```
log file=myanalysis.log start
```

and to stop logging (i.e., close the file) use

```
log stop
```

In addition, you can use something like

```
log file=myanalysis.log replace
```

to automatically overwrite a previously existing file, and

```
log file=myanalysis.log append
```

to add new output onto the end of a previously existing file. This command is particularly useful when placed in a GDA block at the beginning of the data file:

```
#nexus

begin gda;
  log file=myanalysis.log replace;
end;
```

## 2.13 Quit Command

*Windows version menu choice: File | Exit*

Exits the program and returns you to the operating system prompt.

## 2.14 Set Command

*Windows version menu choice: Misc | Preferences...*

The **set** command is used to change variables related to the display of output. Typing `help set` produces the following output:

```
SEt
  Precision = <integer value>
  Colwidth = <integer value>
  Outwidth = <integer value>
  Verbose
  NOVerbose
  ?
```

To change the output formatting such that decimal numbers are displayed in columns of width 12 with 6 digits precision, with a maximum page width equal to 79 characters, one could use the command:

```
set precision=6 colwidth=12 outwidth=79
```

To get a listing of the current settings of these variables, simply type

```
set ?
```

Issuing the command

```
set verbose
```

causes GDA to issue some descriptive information about some analyses when they are invoked. Note that not all analyses are documented in this way (we're working on this, however). Setting "noverbose" will revert GDA back to its default mode of withholding descriptive information about analyses.
The *Misc | Preferences...* dialog box offers more options than are possible using the command line. Windows-specific options available through the dialog box include the ability to specify the path to several helper applications:

**Editor** You may specify an editor to use in conjunction with the *File | Editor* menu command, which pops up an editor containing the currently-opened NEXUS data file. If nothing is specified here, the Windows NOTEPAD.EXE application is used.

**Paup*** Specify the path to the PAUP* Windows or DOS application in order to use the *Dist | Invoke Paup* menu command, which first saves the current distance matrix as the NEXUS file *tmp.nex*, then starts PAUP* and loads this file into PAUP*'s editor. From this point, you can issue the keystroke *Ctrl-R* to execute the file in PAUP*, then issue commands to perform a search for the best tree under one of the distance criteria (least squares, nj, upgma, or minimum evolution) implemented by PAUP*. The PAUP* web site is

```
http://www.lms.si.edu/PAUP/
```

**TreeView** Specify the path to the TreeView application in order to use the *Dist | Invoke TreeView* menu command, which first saves the current tree (obtained using UPGMA or WPGMA clustering, or the NJ method) as the NEXUS file *tmp.tre*, then starts TreeView and loads this tree file. At this point, you can (using TreeView) print the tree, save it to the clipboard for pasting into a document, and even reroot the tree. The TreeView web site is

```
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
```

**PDF Viewer** If this field is left blank, GDA will use the default PDF viewer for viewing the PDF documentation when you issue the menu command *Help | PDF Documentation...*. Only specify a path and file name here if you wish to use a different application to view the PDF documentation (other than the default PDF viewer).

## 2.15 Showhierarchy Command

*Windows version menu choice: Hierarchy | Show current*

The **showhierarchy** command produces a diagrammatic representation of the current hierarchy tree (the one that will be assumed for estimation of F-statistics).

## 2.16 Stats Command

*Windows version menu choice: Descr | Options...*

This is the command used to obtain descriptive statistics for the data currently in memory. To include all the descriptive statistics that can be computed by GDA, use the command:

```
stats estimate samplesize ppl al ap he ho f
```

where:

- **samplesize** refers to sample sizes (number of individuals sampled)
- **ppl** is the proportion of polymorphic loci
- **al** is the mean number of alleles per locus
- **ap** is the mean number of alleles per polymorphic locus
- **he** is the expected heterozygosity (based on Hardy-Weinberg expectations)
- **ho** is the observed heterozygosity
- **f** is a method-of-moments estimate of the inbreeding coefficient

To include information about each locus individually, add the keyword **indivloci** to the command. To specify a cutoff frequency for the proportion of polymorphic loci other than the default of 99 (i.e., loci for which the frequency of the most common allele is less than 0.99 are considered polymorphic), include a statement such as `cutoff=90`. Finally, to exclude a statistic from the output, use the prefix `no`. For example, to exclude the statistic "mean number of alleles per polymorphic locus", use the keyword **noap**. The command `stats ?` will provide a listing of which statistics are currently turned on (i.e., will be reported in the output) and which are currently turned off.

## 2.17 Useloci Command

*Windows version menu choice: Misc | Include/Exclude loci...*

The **useloci** command is for selecting subsets of loci for analysis. To use only locus number 3, locus number 4, locus number 5, and locus number 8, issue the following command:

```
useloci 3-5 8
```

Alternatively, you could use the following form:

```
useloci 3 4 5 8
```

If the loci above had been labeled `adh` (locus 3), `'mdh 1'` (locus 4), `'mdh 2'` (locus 5), and `6pgd` (locus 8), you could use either of these versions of the command:

```
useloci adh 'mdh 1' 'mdh 2' 6pgd
useloci adh-'mdh 2' 6pgd
```

Finally, to reactivate all loci again, you can use the command:

```
useloci all
```

## 2.18 Usepops Command

*Windows version menu choice: Misc | Preferences...*

The **usepops** command works in the same way as the **useloci** command, except the units being activated and deactivated are populations rather than loci.

# Tutorial

## A.1 Introduction

This tutorial is designed to acquaint you with most of the analyses that can be accomplished using GDA. This tutorial presumes you are using the graphical Windows version of the program, but the necessary commands are also provided for each step to make this tutorial useful regardless of the version of GDA you are using. For a more detailed explanation of any of the commands used, see Chapter 1.4 starting on page 12.

Launch the GDA application by double-clicking its icon gda to begin the tutorial.

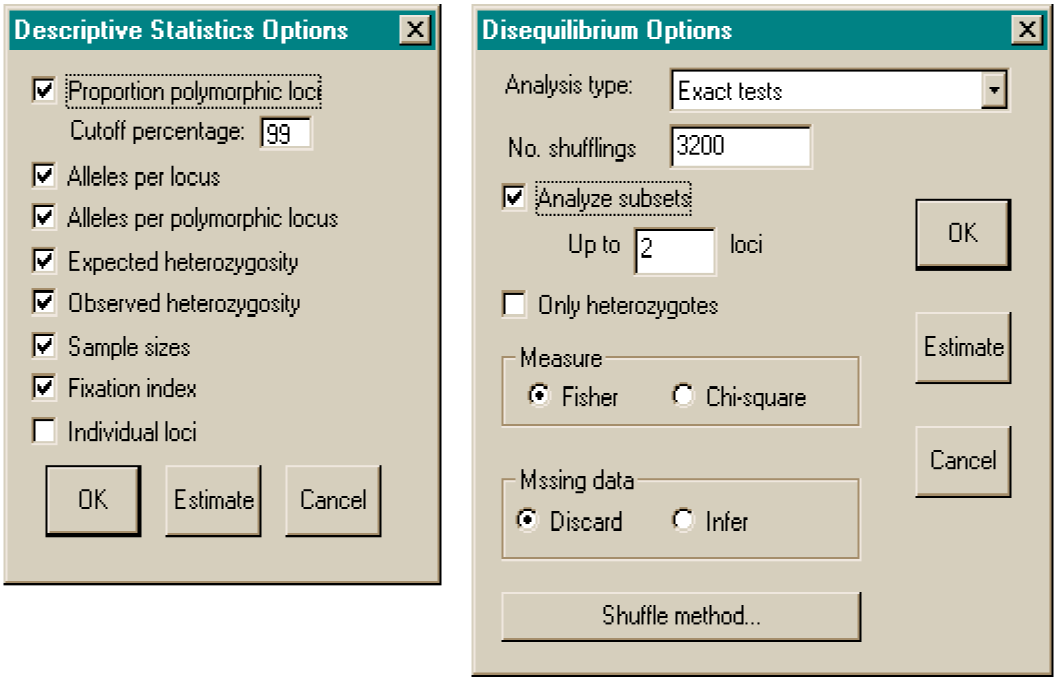
## A.2 Executing the sample data file

**Executing** a data file means that GDA will try to interpret the contents of the file as genetic data. To execute the sample data file *diploid.nex*, select the file using the File | Open... menu command, or type:

```
exe file=diploid.nex;
```

The output should now resemble that shown below.

```
Opening the NEXUS data file C:\GDA1\diploid.nex

Sample data set on p. 338-339 of Weir, B. S. 1990. Genetic
Data Analysis. Sinauer, Sunderland, Mass.

Data matrix has 6 populations, 5 loci, and 44 individuals
  Missing data represented by the symbol ?
  Different genes at one locus separated by the symbol /
  Multicharacter allele names.
  Labels provided for individuals.
  Respecting case for allele names.
  Data matrix not interleaved.
  All loci are diploid
```



Descriptive Statistics Options

☑ Proportion polymorphic loci

Cutoff percentage: 99

☑ Alleles per locus

☑ Alleles per polymorphic locus

☑ Expected heterozygosity

☑ Observed heterozygosity

☑ Sample sizes

☑ Fixation index

□ Individual loci

Disequilibrium Options

Analysis type: Exact tests

No. shufflings: 3200

☑ Analyze subsets

Up to 2 loci

□ Only heterozygotes

Measure: ● Fisher ○ Chi-square

Mssing data: ● Discard ○ Infer

Figure A.1: The Descriptive Statistics Options and Disequilibrium Options (Exact Tests) dialog boxes

## A.3 Logging output to a file

In order to permanently save the output of this analysis, we will open a log file called *mylog.txt*. To do this, use the File | Log... menu item or type:

```
log file=mylog.txt replace;
```

The **replace** keyword instructs GDA to overwrite the file *mylog.txt* if it already exists. If this keyword is not specified, GDA would ask whether it was Ok to overwrite the file before proceeding. The following output line should appear after this command is entered:

```
Logging to file C:\GDA1\mylog.txt
```

## A.4 Producing a table of descriptive statistics

Now we will produce a table of descriptive statistics for this data set. Choose Descr | Options... to bring up the **Descriptive Statistics Options** dialog box (see Figure A.1). Ensure that all options are checked except for **indivloci** checkbox and press the **Estimate** button. The corresponding command for this action is:

```
stats est samplesize ppl al ap he ho f;
```

The output should be similar to that shown below. This table presents, for each population: the mean sample size ($n$) over all loci, the proportion of polymorphic loci ($P$), the mean number of alleles per locus ($A$), the mean number of alleles per polymorphic locus ($A_p$), the expected heterozygosity ($H_E$)[1], the observed heterozygosity ($H_O$), and an estimate of the fixation index ($f$)[2].

```
Descriptive statistics (by population):

  Population           n           P           A          Ap
  ----------  ----------  ----------  ----------  ----------
       Pop 1    7.400000    0.600000    1.800000    2.333333
       Pop 2    8.000000    0.600000    1.800000    2.333333
       Pop 3    5.000000    0.400000    1.800000    3.000000
       Pop 4    7.000000    0.400000    1.400000    2.000000
       Pop 5    8.800000    0.600000    2.000000    2.666667
       Pop 6    7.000000    0.400000    1.800000    3.000000
  ----------  ----------  ----------  ----------  ----------
        Mean    7.200000    0.500000    1.766667    2.555556

  Population          He          Ho           f
  ----------  ----------  ----------  ----------
       Pop 1    0.245556    0.350000   -0.485969
       Pop 2    0.306667    0.175000    0.446328
       Pop 3    0.266667    0.240000    0.111111
       Pop 4    0.186813    0.200000   -0.076923
       Pop 5    0.211046    0.238889   -0.142079
       Pop 6    0.195604    0.200000   -0.024390
  ----------  ----------  ----------  ----------
        Mean    0.235392    0.233981    0.008092

Descriptive statistics (by locus):

       Locus           n           P           A          Ap
  ----------  ----------  ----------  ----------  ----------
     locus 1   44.000000    0.000000    1.000000         ***
     locus-2   44.000000    1.000000    2.000000    2.000000
     locus-3   41.000000    1.000000    3.000000    3.000000
     locus-4   43.000000    1.000000    4.000000    4.000000
     locus-5   44.000000    0.000000    1.000000         ***
  ----------  ----------  ----------  ----------  ----------
         All   43.200000    0.600000    2.200000    3.000000
```

[1] The expected heterozygosity calculated by GDA is the unbiased estimator formed by multiplying the sample expected heterozygosity $(1 - \sum_u p_u^2)$ by the factor $(2n)/(2n-1)$

[2] Unlike the mean for the other statistics (simple average), the mean for $f$ is computed by summing the numerator and denominator of the estimator (eq. 2.28, p. 80, in [4]) separately, then forming a quotient from these two sums.

```
     Locus          He          Ho           f
----------  ----------  ----------  ----------
   locus 1    0.000000    0.000000    0.000000
   locus-2    0.128527    0.090909    0.295082
   locus-3    0.426076    0.439024   -0.030780
   locus-4    0.625171    0.604651    0.033201
   locus-5    0.000000    0.000000    0.000000
----------  ----------  ----------  ----------
       All    0.235955    0.226917    0.038771
```

## A.5 Testing for Hardy-Weinberg and pairwise disequilibrium

Choose Diseq | Options... to bring up the **Disequilibrium Options (Exact Tests)** dialog box (see Figure A.1). Making sure that the **Exact tests** choice is visible in the **Analysis type** drop-down list, click on the **Estimate** button to begin the exact tests for Hardy-Weinberg and pairwise linkage disequilibrium. The command for the above menu action is:

```
exact est upto=2;
```

The statement `upto=2` specifies the maximum number of loci to be involved in tests of disequilibrium. If `upto=1` had been specified, each individual locus would be tested for Hardy-Weinberg disequilibrium. Since we specified `upto=2`, tests will be performed for Hardy-Weinberg disequilibrium at individual loci, but linkage disequilibrium will also be tested for all possible pairs of loci. Specifying `upto=3` would add analyses of all possible locus triplets. The probabilities shown are estimates (obtained using shuffling tests) of the exact significance levels. An exact probability less than 0.05 indicates a statistically significant amount of disequilibrium (using the normal 5% rule of thumb). An example of output (first two populations only) from this command is shown below.

```
Exact tests for linkage and Hardy-Weinberg disequilibrium:

  Subsets of loci will be analyzed
  Subsets will be comprised of up to 2 loci
  Individuals with missing data will be discarded
  Number of runs: 3200
  Measure: Fisher
  Shufflings will break up genotypes for all loci

Population # 1 (Pop 1) of 8 individuals
        Runs        Prob            Locus combination
------------------------------------------------------------
        3200    1.000000            locus 1
        3200    1.000000            locus-2
        3200    0.181875           locus-3
        3200    0.495625           locus-4
        3200    1.000000           locus-5
        3200    1.000000           locus 1/locus-2
        3200    0.177500           locus 1/locus-3
        3200    0.500938           locus 1/locus-4
        3200    1.000000           locus 1/locus-5
        3200    0.695312           locus-2/locus-3
        3200    0.303125           locus-2/locus-4
        3200    1.000000           locus-2/locus-5
        3200    1.000000           locus-3/locus-4
        3200    0.179063           locus-3/locus-5
        3200    0.480625           locus-4/locus-5

Population # 2 (Pop 2) of 8 individuals
        Runs        Prob           Locus combination
------------------------------------------------------------
        3200    1.000000           locus 1
        3200    0.395313           locus-2
        3200    0.082187           locus-3
        3200    0.368437           locus-4
        3200    1.000000           locus-5
        3200    0.387188           locus 1/locus-2
        3200    0.077813           locus 1/locus-3
        3200    0.362187           locus 1/locus-4
        3200    1.000000           locus 1/locus-5
        3200    0.068750           locus-2/locus-3
        3200    0.256250           locus-2/locus-4
        3200    0.390937           locus-2/locus-5
        3200    0.016875           locus-3/locus-4
        3200    0.075625           locus-3/locus-5
        3200    0.358438           locus-4/locus-5
```

The output indicates that no locus shows a significant departure from Hardy-Weinberg expectations in any population, although locus 3 in population 2 and locus 4 in population 5 come close (with estimated exact probabilities of 0.082187 and 0.074688, respectively). Population 2 shows significant pairwise disequilibrium between locus 3 and locus 4 (probability 0.016875), although this disequilibrium could largely be due to the Hardy-Weinberg disequilibrium present in locus 3 in this population. This is because the pairwise measures include all types of disequilibrium (i.e., within-locus as well as between-locus disequilibrium).

We can prevent the within-locus disequilibrium from affecting the significance of disequilibrium in pairwise (and higher order) comparisons by telling GDA to preserve genotypes. To tell GDA to perform the exact tests again, but this time preserving genotypes for all loci when performing the shuffling tests, open up the **Disequilibrium Options (Exact Tests)** dialog box once more (Diseq | Options...), but this time click the **Shuffle Method** button. After making sure all loci are on the "Preserve" side of the dialog box (click the **All** button on the "Break up" side, then press the **Preserve** button to move all selected loci to the "Preserve" side), click the **Ok** button to close the **Shuffle Method** dialog box and then **Estimate** button to begin the exact tests again. The following command is equivalent to the above set of menu actions:

```
exact est upto=2 permute=ppppp
```

Here are the results for just population 2. Note that there is no evidence for linkage disequilibrium between locus 3 and locus 4 now that the effects of Hardy-Weinberg disequilibrium have been removed (probability 0.41875):

```
Population # 2 (Pop 2) of 8 individuals
        Runs        Prob            Locus combination
------------------------------------------------------------
        3200    1.000000            locus 1/locus-2
        3200    1.000000            locus 1/locus-3
        3200    1.000000            locus 1/locus-4
        3200    1.000000            locus 1/locus-5
        3200    0.764062            locus-2/locus-3
        3200    1.000000            locus-2/locus-4
        3200    1.000000            locus-2/locus-5
        3200    0.418750            locus-3/locus-4
        3200    1.000000            locus-3/locus-5
        3200    1.000000            locus-4/locus-5
```

## A.6 Excluding loci and populations

To exclude the monomorphic loci (locus 1 and locus 5), choose Misc | Include/exclude loci (see Figure A.2) from the menu to invoke the **Include/exclude loci** dialog box, then double-click locus 1 and locus 5 to move them from the "Included" side of the dialog box to the "Excluded" side. To do this using commands, you exclude loci by telling GDA to use all loci except the ones you wish to exclude:

```
useloci 2-4;
```

In either case (dialog box method or using the typed command), GDA responds with:

```
Excluding 2 loci

Active loci:
  locus-2
  locus-3
  locus-4
```

To exclude both the first and the second *population* from further consideration, choose Misc | Include/exclude populations (see Figure A.3) from the menu to invoke the **Include/exclude populations** dialog box. This time, instead of double-clicking the population names, click each only once, holding down the Ctrl key to enable multiple selection. Once both population 1 and

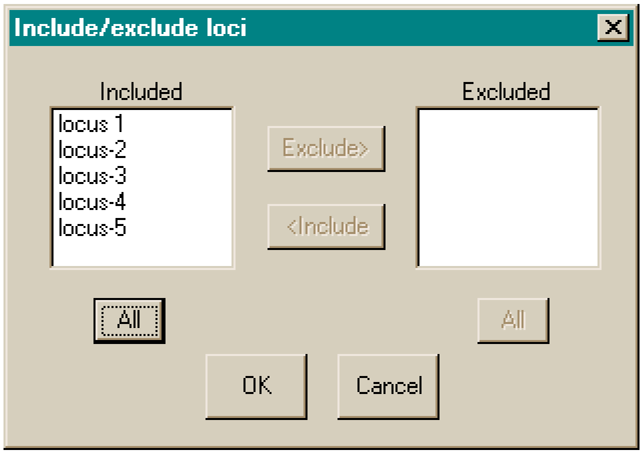


Figure A.2: The Include/exclude Loci dialog box

population 2 have been selected (as shown in Figure A.3), click the **Exclude** to move them both simultaneously from the "Included" side of the dialog box to the "Excluded" side. You can always choose either the double-click or the single-click method when using either of these "Include/exclude" dialog boxes. To do the same thing using a command, we tell GDA that we only want to use populations 3 to 6:

```
usepops 3-6;
```

GDA responds with:

```
Excluding 2 populations

Active populations:
  Pop 3
  Pop 4
  Pop 5
  Pop 6
```

## A.7 Computing composite linkage disequilibrium coefficients

Choosing Diseq | Options... and selecting the **Linkage disequilibrium** choice in the **Analysis type** drop-down list will bring up the **Disequilibrium Options (Linkage Diseq.)** dialog box (see Figure A.4), and then pressing the **Ok** button will produce estimates of linkage disequilibrium coefficients. The command equivalent of this menu choice is **ldiseq** command (you can shorten it to just **ld**):

```
ld est;
```

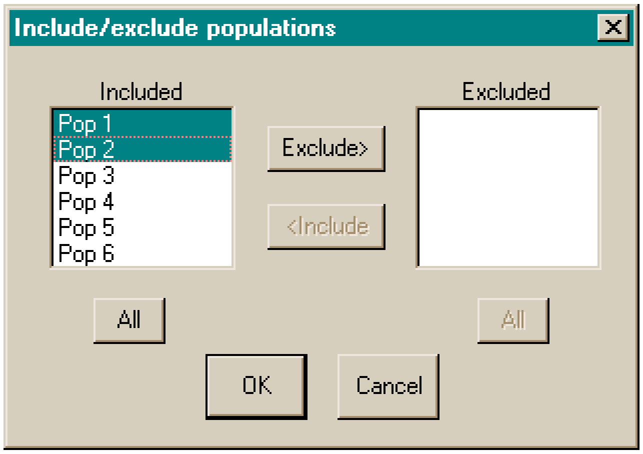
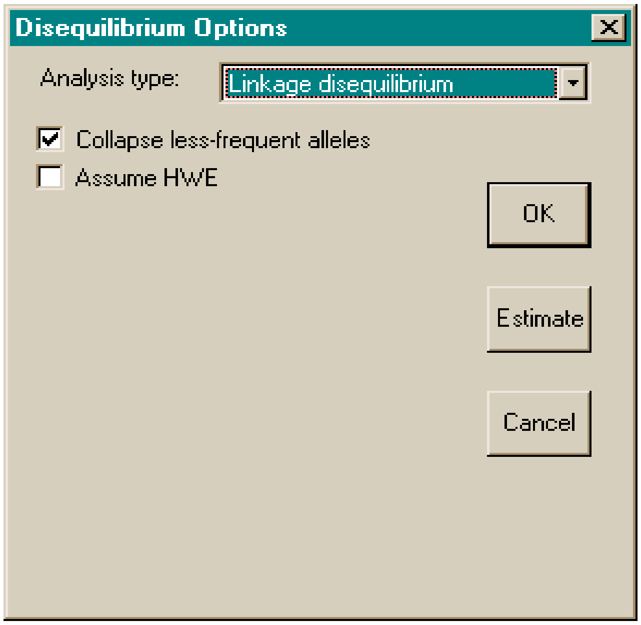


Figure A.3: The Include/exclude Populations dialog box



Analysis type: Linkage disequilibrium

☑ Collapse less-frequent alleles

□ Assume HWE

Figure A.4: The Disequilibrium Options (Linkage Disequilibrium) dialog box

Note that, by default, this analysis does not assume Hardy-Weinberg equilibrium and collapses alleles into the most common allele versus the rest. The collapsing does not affect the stored data matrix (i.e., a virtual, not actual, collapsing of the data is performed). A portion of the output produced by the above command is reproduced below.

```
Composite disequilibrium measures:
   Comparing only most common alleles at each locus
  Not assuming Hardy-Weinberg equilibrium (composite disequilibrium analysis)


Population # 3 (Pop 3) of 5 individuals

loci: locus-3/locus-4; alleles: 4/4
counts:    1    1    0    0    0    2    0    1    0

     coeffs:           D_a           D_b          D_ab         D_aab
     estims:      0.040000      0.040000      0.120000      0.068000
     stdevs:      0.107331      0.107331      0.087636      0.028510
     chisqu:      0.138889      0.138889      0.782609      3.648990

     coeffs:         D_abb        D_aabb
     estims:      0.032000     -0.044800
     stdevs:      0.040160      0.032843
     chisqu:      0.533333      0.863341
```

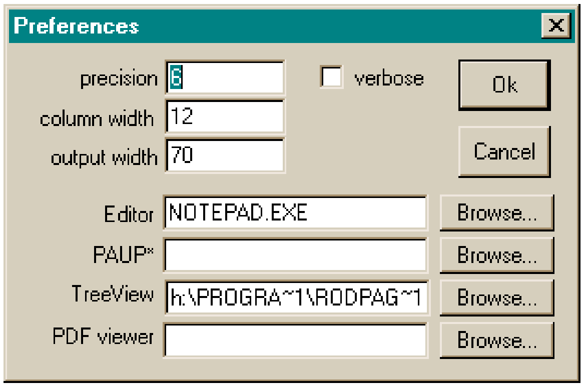
We can note from the output ("`loci: locus-3/locus-4; alleles: 4/4`") that the most common allele at both of the loci compared (locus-3 and locus-4) was allele 4. The `counts` line shows the counts of each possible combination of a genotype at the first locus with a genotype at the second locus. The counts shown in the output above correspond to the following table, where $\bar{4}$ means "not allele 4" in genotype designations:

| | | locus-4 | | |
|---|---|---|---|---|
| | | 4/4 | 4/$\bar{4}$ | $\bar{4}$/$\bar{4}$ |
| locus-3 | 4/4 | 1 | 1 | 0 |
| | 4/$\bar{4}$ | 0 | 0 | 2 |
| | $\bar{4}$/$\bar{4}$ | 0 | 1 | 0 |

The disequilibrium coefficients labeled `D_a` and `D_b` represent the Hardy-Weinberg disequilibrium measure $D_A$ (described in [4, p. 95]) for the two loci represented in the comparison, and the coefficient labeled `D_ab` is the two-locus linkage disequilibrium measure $D_{AB}$ (described in [4, pp. 112-114]).

## A.8 Changing the output formatting

It is possible to change the formatting of the output in some ways. For example, it is possible to change the column width, numerical precision, and output width using the Misc | Preferences...



Preferences

precision 6

column width 12

output width 70

□ verbose

Editor NOTEPAD.EXE

PAUP*

TreeView h:\PROGRA~1\RODPAG~1

PDF viewer

Figure A.5: The Preferences dialog box

menu choice. This invokes the **Preferences** dialog box, which is shown in Figure A.5. The column width can be increased to 20 and the number of decimal places (numerical precision) increased to 12 using the commands:

```
set colwidth=20 precision=12
```

or by changing the appropite quantities in the dialog box. Rerun the previous command (the linkage disequilibrium estimates) after setting the column width to 20 and the precision to 12 produces the following output (again, only the first portion of the output is shown):

```
General preferences:
  Precision = 12
  Colwidth  = 20
  Outwidth  = 70
  Verbose   = yes

Composite disequilibrium measures:
  Comparing only most common alleles at each locus
  Not assuming Hardy-Weinberg equilibrium (composite disequilibrium analysis)


Population # 3 (Pop 3) of 5 individuals

loci: locus-3/locus-4; alleles: 4/4
counts:    1    1    0    0    0    2    0    1    0

             coeffs:                   D_a                 D_b
             estims:        0.040000000000      0.040000000000
             stdevs:        0.107331262920      0.107331262920
             chisqu:        0.138888888889      0.138888888889
             coeffs:                  D_ab                 D_aab
             estims:        0.120000000000        0.068000000000
             stdevs:        0.087635609201        0.028509647490
             chisqu:        0.782608695652        3.648989898990

             coeffs:                 D_abb                D_aabb
             estims:        0.032000000000       -0.044800000000
             stdevs:        0.040159681274        0.032842820829
             chisqu:        0.533333333333        0.863341041735
```

It is also possible to change the total width (in terms of the number of characters) of the output as well:

```
set outwidth=50;
```

Now reissuing the linkage disequilibrium command results in output squeezed down to a single column (note that once again only partial output is shown below to save space):

```
General preferences:
  Precision = 12
  Colwidth  = 20
  Outwidth  = 50
  Verbose   = yes

Composite disequilibrium measures:
  Comparing only most common alleles at each locus
  Not assuming Hardy-Weinberg equilibrium (composite disequilibrium analysis)


Population # 3 (Pop 3) of 5 individuals

loci: locus-3/locus-4; alleles: 4/4
counts:    1    1    0    0    0    2    0    1    0

             coeffs:                   D_a
             estims:        0.040000000000
             stdevs:        0.107331262920
             chisqu:        0.138888888889

             coeffs:                   D_b
             estims:        0.040000000000
             stdevs:        0.107331262920
             chisqu:        0.138888888889

             coeffs:                  D_ab
             estims:        0.120000000000
             stdevs:        0.087635609201
             chisqu:        0.782608695652

             coeffs:                 D_aab
             estims:        0.068000000000
              stdevs:        0.028509647490
             chisqu:        3.648989898990

             coeffs:                 D_abb
             estims:        0.032000000000
             stdevs:        0.040159681274
             chisqu:        0.533333333333

             coeffs:                D_aabb
             estims:       -0.044800000000
             stdevs:        0.032842820829
             chisqu:        0.863341041735
```

There are a couple of other features available in the **Preferences** dialog box. Checking the **verbose** checkbox sets a flag that causes explanatory text to be issued whenever some analyses are chosen. The text editor invoked whenever the File | Editor menu item is chosen may also be specified here (it is the standard Windows text editor NOTEPAD.EXE by default). Other helper applications that can be specified here include PAUP*, TreeView and the preferred PDF viewer. PAUP*[3] is useful for performing analyses on genetic distances computed by GDA (see page 39), and TreeView[4] is useful for viewing trees generated by cluster analyes in GDA (see page 39). The PDF viewer is convenient if you want to invoke this PDF file from within GDA.

## A.9 Estimating F-statistics

Now we will produce a table of estimates of Wright's F-statistics: $\theta$ (or $F_{ST}$), $F$ (or $F_{IT}$), and $f$ (or $F_{IS}$). First, reinclude Population 1 and Population 2 and restore the original output preferences. You can do this either through the menu choices (see section A.6 on page 27 for the former and section A.8, page 30, for the latter), or use the following commands:

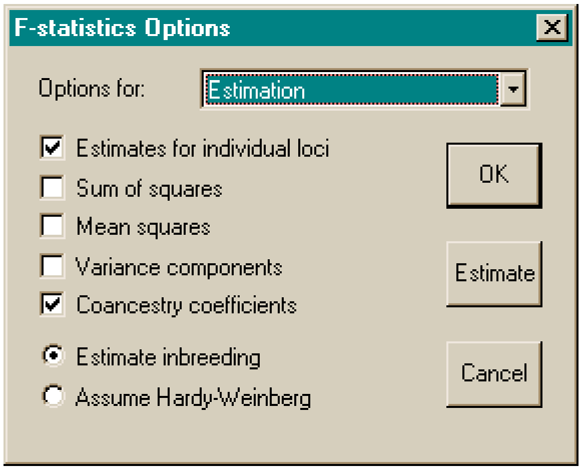
```
usepops all;
set colwidth=12 precision=6 outwidth=70;
```

If all goes well, you should see output similar to this:

```
Including 2 populations

Active populations:
  Pop 1
  Pop 2
  Pop 3
```

[3] For information about the program PAUP*, see http://www.lms.si.edu/PAUP/

[4] To download TreeView, see http://taxonomy.zoology.gla.ac.uk/rod/treeview.html



Options for: Estimation

☑ Estimates for individual loci
□ Sum of squares
□ Mean squares
□ Variance components
☑ Coancestry coefficients

● Estimate inbreeding
○ Assume Hardy-Weinberg

Figure A.6: The F-statistics Options dialog box

```
  Pop 4
  Pop 5
  Pop 6

General preferences:
  Precision = 6
  Colwidth  = 12
  Outwidth  = 70
  Verbose   = yes
```

To obtain estimates of the F-statistics, choose F-stats | Options..., making sure that the resulting **F-statistics Options** dialog box looks like that shown in Figure A.6, then click the **Estimate** button to tell GDA you are ready. The command equivalent to this menu action is:

```
fstats est;
```

and the output should appear identical to the following (the initial description of the F-statistics analysis will only appear if you have checked the **verbose** checkbox in the **Preferences** dialog box):

```
F-statistics analysis (written by Paul O. Lewis). This analysis
estimates the parameters f (=Fis), F (=Fit), and Theta (=Fst) for up
to a four-level hierarchy of populations. Notes: (1) f is only
parameter estimable if there is only one active population; (2) F and
Theta-P are estimated for the simplest (2-level) hierarchy as well as
3-level or 4-level hierarchies; (3) Theta-S is estimated for 3-level
and 4-level hierarchies; (4) Theta-SS is estimated for 4-level
hierarchies only. Based on chapter 5 in Weir, B. S. (1996. Genetic
Data Analysis II. Sinauer) and Weir, B. S., and Cockerham, C. C.
```

```
(1994. Evolution 38: 1358-1370).

 Analysis of variance
  There are 3 active loci
  Two-level analysis
    6 active populations

       Locus      Allele           f           F     Theta-P
  ----------  ----------  ----------  ----------  ----------
     locus-2         All    0.250514    0.302529    0.069401
                       3    0.250514    0.302529    0.069401
                       4    0.250514    0.302529    0.069401

     locus-3         All   -0.034807   -0.030052    0.004595
                       2   -0.032448    0.005674    0.036924
                       3   -0.006242   -0.016518   -0.010212
                       4   -0.063455   -0.045283    0.017087

     locus-4         All    0.012751    0.036738    0.024297
                       1   -0.079968   -0.029778    0.046473
                       2    0.176678    0.185832    0.011118
                       4    0.025000    0.026754    0.001799
                       3   -0.042608    0.007194    0.047767

  ----------  ----------  ----------  ----------  ----------
     Overall         ---    0.020232    0.041954    0.022171
```

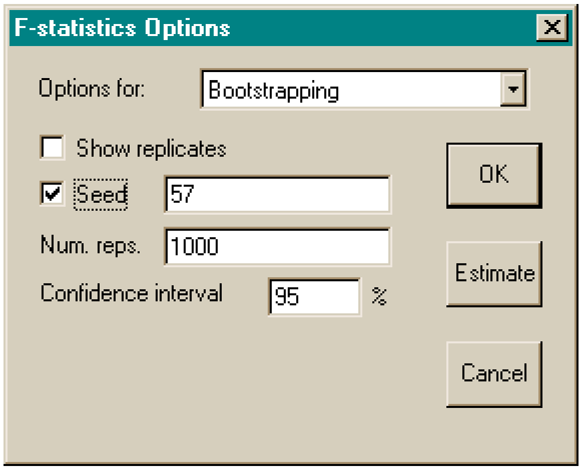
A word of explanation is in order. The columns are labeled `f`, `F`, and `Theta-P`, rather than `Fis`, `Fit`, and `Fst` (as you might expect). In [3, 4], the parameters underlying Wright's F-statistics are called $F$, $f$, and $\theta$ rather than $F_{IT}$, $F_{IS}$, and $F_{ST}$ because of the confusion that surrounds the usage of the latter symbols. `Theta-P` is specified (rather than just `Theta`), because in a hierarchical analysis there is a need to distinguish between the $\theta_P$ for populations (`Theta-P`) and the $\theta_S$ for subpopulations (`Theta-S`). While this is not a hierarchical analysis (i.e., there are no subpopulations), `Theta-P` is used here simply for consistency.

## A.10 Viewing the population hierarchy

The population hierarchy tree used for obtaining estimates of F-statistics earlier can be viewed either by typing the command below:

```
showhierarchy;
```

or using the menu choice Hierarchy | Show current. In either case, GDA responds with the following output, indicating that all 6 populations defined in the data file are assumed to be on the same hierarchical level. This is what [4] refers to as a two-level F-statistics analysis: the first level is the reference population (called Root by default), from which the 6 populations on the second level are assumed to have been derived.



Options for: Bootstrapping

□ Show replicates

☑ Seed 57

Num. reps. 1000

Confidence interval 95 %

Figure A.7: The Bootstrapping Options dialog box

```
Current hierarchy tree

+---------------------------------------------------------------Pop 1
|
|---------------------------------------------------------------Pop 2
|
|---------------------------------------------------------------Pop 3
Root
|---------------------------------------------------------------Pop 4
|
|---------------------------------------------------------------Pop 5
|
+---------------------------------------------------------------Pop 6
```

## A.11 Bootstrapping over loci to obtain confidence intervals for F-statistics

To assess the significance of the F-statistics estimates obtained above, bootstrapping over loci may be used. First, reinstate all loci to active status (see section A.6 on page 27). Now bring up the **Bootstrapping Options** dialog box (shown in Figure A.7) using the menu command F-stats | Options..., choosing the **Bootstrapping** choice in the **Options for** drop-down list. Make sure that the number of replicates is set to 1000 and the confidence interval is set to 95%, then hit the **Estimate** button[5]. Typing the commands

```
useloci all;
```

[5] If you want your bootstrapping results to be identical to those shown in the example output, check the **seed** checkbox and type `57` in the accompanying edit control before hitting the **Estimate** button

```
bootloci est nreps=1000 ci=95 rseed=57;
```

will accomplish the same thing. Either method causes GDA to reactivate the loci previously excluded and then bootstrap across loci, presenting a table similar to the following:

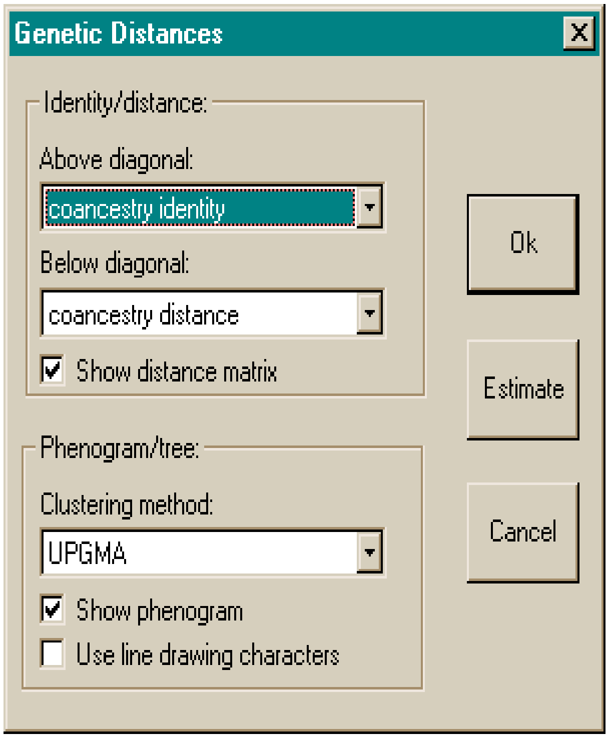
```
F-statistics bootstrap analysis (written by Paul O. Lewis). This
analysis obtains bootstrap confidence intervals for the parameters f
(=Fis), F (=Fit), and Theta (=Fst). Notes: (1) f is only parameter
estimable if there is only one active population; (2) F and Theta-P
are estimated for the simplest (2-level) hierarchy as well as 3-level
or 4-level hierarchies; (3) Theta-S is estimated for 3-level and
4-level hierarchies; (4) Theta-SS is estimated for 4-level
hierarchies only. Based on chapter 5 in Weir, B. S. (1996. Genetic
Data Analysis II. Sinauer).

Bootstrapping over loci
  Number of replicates = 1000
  Nominal confidence interval = 95%
  Random number seed = 57 (specified)

          Bound           f           F     Theta-P
  -------------  ----------  ----------  ----------
          Upper    0.250514    0.302529    0.069401
          Lower   -0.034807   -0.030052    0.004595
      No. reps.        1000        1000        1000
  CI (realized)   95.000000   95.000000   95.000000
```

which shows that only $\theta$ is significantly greater than zero (the 95% confidence regions of $F$ and $f$ both span 0.0). Note that a brief description of the analysis is included before the results. Analysis descriptions will eventually be included for all analyses performed by GDA as a means of pointing the user to further reading on the topic. This option can be turned off with the command `set noverbose` or by unchecking the **verbose** checkbox in the **Preferences** dialog box (Misc | Preferences...).

## A.12 Obtaining pairwise genetic distances between populations

A genetic distance appropriate for the case in which genetic drift is the force responsible for changing gene frequencies over time can be computed using the value of $\theta$ between two populations (see [4, pp. 194-195]). The following command will output a square matrix in which estimates of $\theta$ for each pair of populations are above the main diagonal and the genetic distances based on those $\theta$ values ($d = -\ln(1-\theta)$) are below the main diagonal. In addition, the distances will be clustered using the UPGMA (Unweighted Pair Group Method using Arithmetic averaging) and a tree depicting these phenetic relationships among populations will be drawn.

```
gdist above=driftiden below=driftdist showmatrix cluster=upgma est;
```



Identity/distance:

Above diagonal: coancestry identity

Below diagonal: coancestry distance

☑ Show distance matrix

Phenogram/tree:

Clustering method: UPGMA

☑ Show phenogram

☐ Use line drawing characters

Figure A.8: The Genetic Distances Options dialog box

To do the same thing via the menu, first choose Dist | Options... to bring up the **Genetic Distances Options** dialog box (shown in Figure A.8), ensure that the **Show distance matrix** checkbox and the **Show phenogram** checkbox are both checked, select the **Coancestry identity** choice from the **Above diagonal** drop-down list, the **Coancestry distance** choice from the **Below diagonal** drop-down list, and the **UPGMA** choice from the **Clustering method** drop-down list, then press the **Estimate** button to tell GDA to start. The output is shown below:

```
Distance matrix constructed
  Matrix has 6 taxa
  Distances/identity measures based on 3 loci
  coancestry identity used above diagonal
  coancestry distance used below diagonal

Distance matrix
  Pop 1                0.020744    0.066491    0.110078    0.070373    0.106847
  Pop 2    0.020962               -0.038961    0.011930   -0.018475   -0.013095
  Pop 3    0.068804   -0.038221               -0.000557    0.014659   -0.010426
  Pop 4    0.116621    0.012002   -0.000557                0.034238    0.037801
  Pop 5    0.072972   -0.018306    0.014767    0.034838               -0.052535
  Pop 6    0.112998   -0.013010   -0.010372    0.038534   -0.051202

Cluster analysis
  Matrix elements above diagonal ignored
  Node 7 created ( level = 0.000000)
  Node 8 created ( level = 0.000000)
  Node 9 created ( level = 0.004922)
  Node 10 created ( level = 0.021343)
  Node 11 created ( level = 0.078471)

Note!!
  Ties were encountered - tree is not unique
  Negative matrix elements were set to zero

UPGMA phenogram

+--------------------------------------------------------------------Pop 1
|
|                                                                  +-Pop 2
|                                                                  +
11                                                             +---8-Pop 3
|                                                              |   |
|                                                +-------------9   +-Pop 6
|                                                |             |
+------------------------------------------------10            +-----Pop 5
                                                 |
                                                 +-------------------Pop 4

|----------------|----------------|----------------|----------------|
0.039            0.029            0.020            0.010            0.000
```

To view this tree in TreeView, choose the menu item *Dist | Invoke TreeView.* This will start the TreeView program and display the UPGMA tree inside TreeView. Using TreeView, you can print the tree or save it to the clipboard for purposes of pasting it into a manuscript. You can also reroot the tree in TreeView, but this makes sense only for NJ trees.

To instead use Nei's (1978) genetic distances [2] (see also [4, p. 195]) and cluster using the neighbor-joining method, use the following command:

```
gdist above=none below=nei78dist showmatrix cluster=nj est;
```

or choose the **Nei (1978) identity** choice from the **Above diagonal** drop-down list, the **Nei (1978) distance** choice from the **Below diagonal** drop-down list, and the **Neighbor-joining** choice from the **Clustering method** drop-down list in the **Genetic Distances Options** dialog box.

## A.13 Using GDA and PAUP* to estimate phylogenies from genetic distances (optional)

If your goal in obtaining genetic distances is to obtain a phylogeny or phenogram, and if you have access to the phylogenetic software PAUP*, you can obtain a much better estimate of phylogeny by outputting distances from GDA to a file, then reading them into PAUP* to perform a search for the best tree. If PAUP* is available, try the procedure outlined below. *This is an optional exercise, so if you are running short on time, or do not have access to* PAUP*, *feel free to skip to the next section.*

1. The *Dist* | *Invoke Paup* GDA menu command should automatically save your current distance matrix as a file (*tmp.nex*), then start PAUP* and load this file into PAUP*'s editor. All you need do to cause PAUP* to read the file is issue the keystroke *Ctrl-R* or choose the menu item *File* | *Execute "tmp.nex"* from within PAUP*. Alternatively, you can save the distance matrix under a name of your choosing by issuing the GDA menu command *Dist* | *Save matrix to file...*, choosing a filename for the file. The corresponding typed command is

   ```
   putdist file=distances.nex;
   ```

   where *distances.nex* is the name of the resulting NEXUS file containing the distances in a DISTANCES block. Now start the Windows version of PAUP* by double-clicking its icon. When PAUP* starts up, it will display a dialog box asking for the name of a file to open. Specify the name you provided for the file containing the distance matrix (*distances.nex* in the example above).

2. Now type the command `upgma` in PAUP*'s command edit control and press the Enter key or the **Execute** button to obtain the UPGMA tree.

3. Enter the PAUP* command `nj` to get a Neighbor-joining tree.

4. To really harness the power of PAUP*, try entering the following commands (one after another, pressing return after entering the semicolon terminating each command). These commands will cause PAUP* to perform an exhaustive search[6] for the best tree under the least squares criterion (also known as the Fitch-Margoliash criterion). After the search, the tree will be displayed (the `describetrees` command) and the tree will be saved in a file named *distances.ls.tre*.

   ```
   set criterion=distance autoclose;
   dset objective=lsfit power=2;
   alltrees;
   describetrees 1 / plot=phylogram;
   savetrees file=distances.ls.tre brlens;
   ```

[6] An exhaustive search examines every possible tree, and thus is guaranteed to find the globally optimal phylogeny. The number of possible trees is a function of the number of populations in your distance matrix, however, and already exceeds 2 million possible trees at only 10 populations! Thus, do not use PAUP*'s `alltrees` command if you have more than 10 or 11 populations in your distance matrix, otherwise you will be in for a *very* long wait.

5. As one final example of the use of PAUP* to estimate phylogenetic trees from a distance matrix created in GDA, enter the following sequence of commands into PAUP*. These will find the best tree under the minimum evolution criterion (the same criterion used by neighbor-joining), but using an heuristic search rather than an exhaustive search. Heuristic searches are not guaranteed to find the globally optimal tree, but are the only practical search strategy when the number of populations being analyzed is larger than about 10 or 11 (see footnote).

   ```
   dset objective=me;
   hsearch start=nj swap=tbr;
   savetrees file=distances.me.tre brlens;
   ```

6. The MacIntosh version of PAUP* provides a very nice facility for viewing and printing trees (as well as a very user-friendly, menu-driven interface!), but these capabilities have not yet been incorporated into the Windows version. You can view and/or print the trees saved in the files *distances.ls.tre* and *distances.me.tre* using the free program TreeView by Rod Page, however.

7. Exit PAUP* by entering the command `quit`, pressing the key combination Ctrl-Q, or choosing File | Exit from the menu.

## A.14 Closing the log file

Back in GDA again (if you've been using PAUP*), to stop logging the output to the file *mylog.txt*, choose File | Log... or use the command

```
log stop;
```

Either action closes the log file. Before the file is closed, some of the output that has been generated may still be waiting in a buffer in the computer's memory. Closing the file forces this buffered output to be written to the file.

## A.15 Quitting the program

You can terminate the program by choosing File | Exit or by simply typing the command `quit` and pressing the Enter key.

## A.16 FSim and ThetaSim

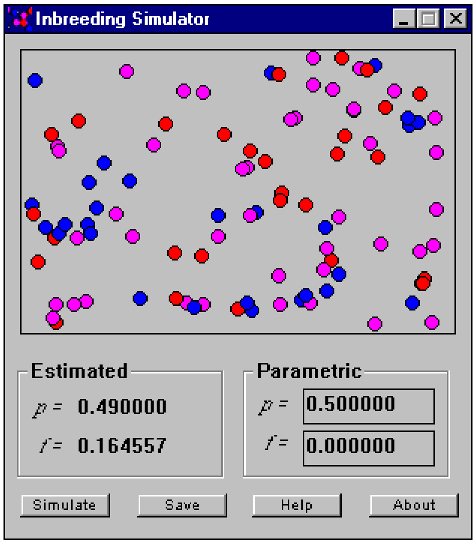
FSim and ThetaSim are two graphical Windows applications designed to teach concepts related to estimation and interpretation of Wright's F-statistics (and specifically the parameterization of these quantities used in [4]. You can download these programs over the web for your personal use or to use in your own courses. They are available along with GDA at Paul Lewis' software download web page:

`http://lewis.eeb.uconn.edu/lewishome/software.html`

In both FSim and ThetaSim, diploid individuals are represented by randomly-placed spots whose color indicates the individual's genotype. The spatial position of each individual is random, and thus has no particular meaning; what is important is proportions of the three possible genotypes.

### A.16.1 FSim

FSim (Figure A.9) simulates populations showing varying degrees of inbreeding. The opening screen shows 100 randomly arranged circles in three different colors. *Blue* circles can be taken to represent individuals with genotype **AA** at a particular locus, *red* circles represent **aa** individuals, and *pink* circles are heterozygotes (**Aa**) at this locus.



Estimated

$p$ = 0.490000

$f$ = 0.164557

Parametric

$p$ = 0.500000

$f$ = 0.000000

Figure A.9: The FSim program running.

Under the panel with the colored circles are two panels labeled *Estimated* and *Parametric*. The values in the *Parametric* panel are the true values used to simulate the population. The values in the *Estimated* panel are the method-of-moment estimates of the same parameters. The values in the *Estimated* panel would be the estimates that GDA would produce.

You can modify the parametric values and perform a new simulation using those values. Try changing the value for $f$ to 1.0 and pressing the *Simulate* button. You should find there are now no heterozygotes in the simulated population since the inbreeding coefficient $f$ is at its maximum value. Now try changing the value of $f$ to 0.0 (as it was at the beginning). Pressing the *Simulate* button now shows a population with no inbreeding (i.e., the numbers of each genotype correspond to what would be predicted from considering only the frequencies of the individual alleles making up that genotype).

Now change the frequency $p$ to 0.1 (it should have been 0.5 at the start). The parameter $p$ is the frequency of the allele **A**, so setting $p$ to a value close to zero will result in a population dominated by *red* (homozygous **a**) and *pink* (heterozygous) individuals. Setting $p$ to a value close to 1.0 will result in mostly *blue* (homozygous **A**) and *pink* individuals.

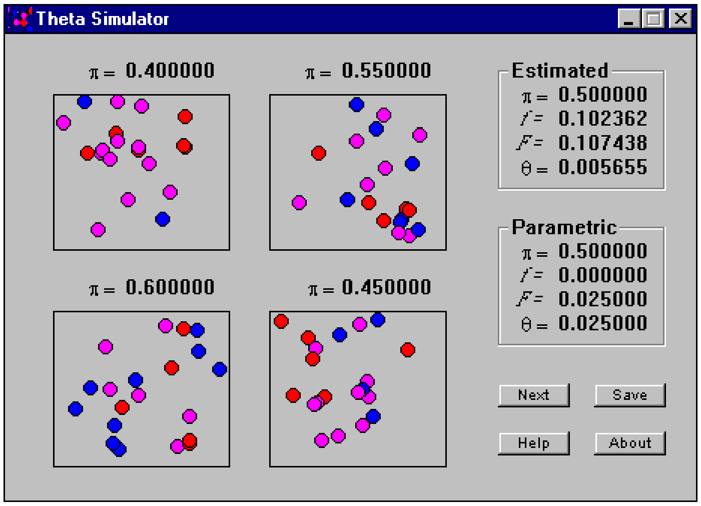
Pressing the *Save* button saves the current population in the form of a NEXUS data file (the file is automatically named *fsim.nex*). This file can be read into GDA and analyzed just like a real data file. You might try, for example, estimating the descriptive statistics (see section A.4 on page 23) and comparing the estimate of the fixation index produced by GDA to the value shown in FSim.

It is possible to change the sample size from 100 to some other value. With the program NOT running, edit the file *fsim.ini*, which (if the program has been used at least once already) will be present in the same directory as the application itself. Changing the value of `n` to some other value will result in that new value being used as the population size when the program is next started. Future versions of this program will allow the population size to be changed interactively.

The equations used in this simulation relating expected genotype frequencies to the parameters $f$ and $p$ may be found in [4, p. 42].

### A.16.2 ThetaSim

ThetaSim (Figure A.10) is similar to FSim in using colored circles to represent individuals in populations and the colors used in the two programs have identical meanings. The main difference is that ThetaSim simulates four populations rather than just one and lets you simulate genetic drift through time. Once again, there are both *Estimated* and *Parametric* panels showing the method-of-moments estimators (such as would be produced by GDA) and the true population values, respectively, of the parameters governing the distribution of genotypes at this simulated locus. The parameter $p$ from FSim is called $\pi$ in ThetaSim, but its meaning is the same – it



π = 0.400000
π = 0.550000
π = 0.600000
π = 0.450000

Estimated
π = 0.500000
f = 0.102362
F = 0.107438
θ = 0.005655

Parametric
π = 0.500000
f = 0.000000
F = 0.025000
θ = 0.025000

Figure A.10: The ThetaSim program running.

represents the proportion of **A** (*blue*) genes, and there is a $\pi$ shown for each subpopulation as well as for the total population. The $\pi$ values shown for each subpopulation are the *sample* values of this parameter. Each subpopulation is *expected* to have the true value for $\pi$, which is shown in the *Parametric* panel.

Besides the parameter $f$ (which corresponds to Wright's $F_{IS}$), two other parameters are listed in both the *Estimated* and *Parametric* panels. The parameter $F$ is the total inbreeding coefficient, corresponding to Wright's $F_{IT}$, and the parameter $\theta$ is the coancestry coefficient, corresponding to Wright's $F_{ST}$.

ThetaSim is designed to simulate the effects of pure genetic drift, and thus the value of $f$ is fixed at 0.0 (i.e., all inbreeding comes from drift and none from non-random mating). Pressing the *Next* button simulates one generation of random mating. That is, all individuals are replaced by offspring whose parents are determined solely by random chance from the previous generation. Thus, it is possible that the same parental individual will be both mother and father to one of the offspring.

Pressing the *Next* button repeatedly allows one to watch the long-term effects of random mating on small populations. Over time, individuals within subpopulations become more related to each other than they were at the start, even though all mating is at random within the subpopulations. At the start of the simulation each individual was generated using Hardy-Weinberg genotypic proportions based on $p = 0.5$. The $\theta$ parameter is initially zero as are $F$ and $f$. The $\theta$ and $F$ parameters increase with each generation of random mating due to the fact that, by chance, some individuals leave no offspring and thus the offspring that are produced have a reduced set of parents and are thus inbred to some extent.

The increasing $\theta$ eventually manifests itself as fixation of either the **A** or **a** allele in a particular subpopulation. When all subpopulations are fixed for one allele or the other, note that the overall frequency of allele **A** is still expected to be 0.5 (i.e., the parametric value of $p$ is still 0.5)! The difference is that now half the subpopulations will be fixed for **A** and the other half fixed for **a**, making the frequency of **A** in the total population one-half (in expectation anyway).

Note that all inbreeding is due to the increase in the coancestry coefficient. This is why the parametric value of the total inbreeding coefficient $F$ is identical to the parametric value of $\theta$ each generation. The *Save* button works as it does in FSim, saving the current data in a NEXUS data file (called *theta4.nex*) so that it may be read into GDA and analyzed. The most relevant analysis to try in GDA with the *theta4.nex* file is the F-statistics analysis (see section A.9 on page 33), which will provide estimates of $f$, $F$, and $\theta$ (it is not possible to bootstrap over loci to test these F-statistics estimates since there is only one locus).

# Bibliography

[1] Maddison, D. R., D. L. Swofford and W. P. Maddison. 1997. NEXUS: an extensible file format for systematic information. *Systematic Biology* **46**: 590-621.

[2] Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583-590.

[3] Weir, Bruce S. 1990. *Genetic Data Analysis* Sinauer, Sunderland, Massachusetts.

[4] Weir, Bruce S. 1996. *Genetic Data Analysis II* Sinauer, Sunderland, Massachusetts.

[5] Weir, B. S., and C. C. Cockerham. 1979. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **42**: 105-111.

[6] Weir, B. S., and C. C. Cockerham. 1989. Complete characterization of disequilibrium at two loci. Pages 86-110 in *Mathematical evolutionary theory* (M. E. Feldman, ed., Princeton University Press, Princeton.

[7] Zaykin, D., L. Zhivotovsky and B. S. Weir. 1995. Exact tests for association between alleles at arbitrary numbers of loci. *Genetica* **96**: 169-178.