(1) DATA PREPARATION COMMANDS

LOAD MARKERS Command

Summary: Load marker-locus data
Argument: <file name>

This command reads in the marker-locus data (allele frequencies for each genetic marker, frequency and penetrance information for the disease). The format of this file must be identical to the Linkage parameter file (output from the PREPLINK program). See the file linkloci.dat as an example of this file format or consult Linkage documentation for further help.

After 3 header lines (only the number of loci on line 1 and the marker order specified on line 3 are relevant and need to be changed), this file must begin with one (and only one) affectation locus describing the disease allele frequencies and penetrances. Following this should be entered the information for each marker as in the following example:

```
3  6  #  D1S1234
.20 .15 .15 .40 .05 .05
```

The 3 on the first line is obligatory, followed by the number of alleles for the marker. If desired a '#' followed by the name of the marker may be entered and this name will then appear on the Postscript output of the 'total' command and can be used to enter marker orders using the 'use' command. The second line for each marker simply contains the allele frequencies for alleles 1 through 6 in this case. Map distances (interlocus distances in the marker order specified on line 3) may be entered on the second to last line in this file format.

The 'load markers' command should occur at the beginning of every session as the information loaded here is required by every subsequent step in the analysis process.

USE Command

Summary: Select the current map for analysis
Argument: <genetic map>
Default: displays the current map selected

The 'use' command is used to select the current map that the 'scan' command will operate on. It is called in the following manner:

```
use <marker> <distance> <marker> <distance> <marker> ...
```

Markers may be specified numerically (1 being the first listed in the marker locus file - the affectation locus does not count in this numbering scheme as it does in the Linkage parameter file) or by the names specified in the comment area for each marker. If a map is specified in the Linkage parameter file, it will be entered automatically during the "load markers" step. Enter "use" without arguments to see what current linkage map has been entered. IF THERE IS NO LINKAGE MAP IN THE LINKAGE PARAMETER FILE, ONE MUST BE ENTERED USING THE "USE" COMMAND BEFORE ANY ANALYSIS CAN TAKE PLACE.

Distances may be specified as either recombination-fractions or centiMorgans, with the necessary assumption that if EVERY distance is less than 0.5, they are all assumed to be recombination-fractions,

otherwise (if ANY distance is greater than 0.5) they are interpreted as centiMorgan distances.

(2) GENEHUNTER MAPPING COMMANDS

SCAN PEDIGREES Command

Summary: Analyze pedigree data
Argument: <file name>

The main analysis command in GENEHUNTER is the "scan" command. For each pedigree found in the file indicated, the "scan" command will compute LOD scores and NPL sharing statistics at many positions in the genetic map (entered in the locus parameter file or via the "use" command). In addition, if the "count recs" option is turned on, observed recombinations will be displayed for each map interval at the end of the scan for each pedigree. This can be useful in highlighting likely positions of errors in the data.

The pedigree should be in the Linkage pedigree input format (before running MAKEPED or doing any preprocessing!). Each line of this file must have the following structure:

```
 3    12   8   9   1   2   1      1 2   8 3   0 0   4 6   1 3 ...
(a)   (b) (c) (d) (e) (f) (g)     (h ------------------------)
```

(a) pedigree name
(b) individual ID #
(c) father's ID #
(d) mother's ID #
(e) sex (1=MALE, 2=FEMALE)
(f) affectation status (1=UNAFFECTED, 2=AFFECTED)
(g) liability class (OPTIONAL) - classes specified in marker data file
(h) marker genotypes

A 0 in any of the disease phenotype or marker genotype positions (as in the the genotypes for the third marker above) indicates missing data. See the file linkped.pre as an example.

In this file format, you may enter as many pedigrees as you wish in a single file. If a pedigree is too large to be computed using a reasonable amount of time and memory, some individuals that provide less information will be discarded and warnings will be printed. Unaffected individuals with no descendents in the pedigree may be discarded with minimal loss of information and these will be the first eliminated should the pedigree be too large. See the "discard" option if you wish to utilize this speed-up in general.

The scan output of each pedigree consists of up to 5 columns of information (depending on the setting of 'analysis type') as follows:

cM position in the scan
LOD score (computed using the disease model given in the parameter file)
NPL statistic
exact computed significance (p-value)
information content of the genotype data

The "total stat" command may be run after a successful "scan" to see the total scores for the entire data set.

*** IMPORTANT ***
Keep in mind when creating files that there must be a one-to-one correspondence (IN ORDER AND NUMBER) between the markers described in the marker data file and the markers that have genoptypes listed for them in the pedigree file.

TOTAL STAT Command

Summary: Show total scores from a scan of multiple pedigrees
Arguments: <'het'> <fixed-alpha>

The "total" command can only be used after a successful "scan" command of multiple pedigrees. It will display the same 5 columns of output as the "scan" command produced for each pedigree, only now the columns will display the combined values of each statistic (sum of LODscores, combined NPL score, average information content, and p-values of the raw NPL score total). In addition to the screen display of this information (if the "postscript output" option is turned on) postscript graphs of the total NPL statistic (stored in npl_plot.ps), total LOD score (lod_plot.ps), and total information content (info_content.ps) will be created.

In addition two optional arguments may be entered. If the first argument is the word "het" then LODscores under heterogeneity will also be calculated alongside the regular LODscore sum. If a second numeric argument is provided after the word het, the LODscores under heterogeneity will be calculated assuming a fixed alpha (fraction of pedigrees linked - a number between 0.0 and 1.0). If this second argument is not provided, alpha will be allowed to vary until the HLOD is maximized.

SINGLE POINT Command

Summary: activate/deactive single-point analysis
Argument: <'on' or 'off'>
Default: displays the current setting

Turning the 'single point' option on instructs subsequent 'scan' and 'total' commands to calculate and display single-point LOD and NPL scores for each marker in the data set individually rather than the usual multi-point analysis. This command will ignore the linkage map set with the 'use' command and will not produce haplotype output or recombination counts for obvious reasons. 'Single point' is 'off' when GENEHUNTER is initiated.

COUNT RECS Command

Summary: turn recombination counting on
Argument: <'on' or 'off'>
Default: displays the current setting

Turning this option on activates the recombination-counting mechanism in the "scan" command. After each pedigree is scanned, the observed recombinations (and resulting distances) are shown for each map interval alongside the actual distance of the interval. When there are significantly more recombinants than expected in an interval or set of intervals, this can often indicate an error or errors in the genotype data.

At the end of the scan of multiple pedigrees, the overall count of recombinants in each interval is displayed along with the expected value for the entire data set. Recombination counts significantly higher than expected here can be an indication of a marker that is error-prone over multiple pedigrees or of an error in the entered genetic map (either in order or distance).

'Count recs' is ON when GENEHUNTER is started.

HAPLOTYPE Command

Summary: determine likely haplotypes for individuals
Argument: <'on' or 'off'>
Default: displays the current setting

When the 'haplotype' option is turned on, the 'scan' command will report the most likely inferences made regarding the haplotypes of the individuals in each pedigree. The haplotypes for founders will be displayed on the screen and the haplotypes for all individuals analyzed will be stored in a file called haplo.dump. In addition, if the 'postscript output' option is 'on', the entire pedigree (with haplotypes and recombinations indicated) will be drawn in a postscript file suitable for printing and displaying.

The haplotypes displayed represent the maximum-likelihood set of inheritance vectors that explain the data. After all markers have been scanned in a pedigree the most likely path through all of the markers is recreated - thus yielding the most likely pattern of inheritance at each marker and likely positions of recombinants. Among nearby markers that show no recombination, these haplotypes are usually unambiguous, but in cases where recombinants are present (especially in small sibships of 2 or 3 individuals), the haplotypes may be imperfect and simply represent the most likely choice out of several valid choices. For example, the most likely position of recombinants is shown in the PostScript output but other placements may be possible but simply less likely due to considerations of map interval size and allele frequency at certain markers.

Haplotypes can be invaluable tools both analytically (in searching for shared genomic regions of distantly related affected individuals and indicating linkage disequilibrium between markers) and practically (in searching for errors in genotyping which usually manifest themselves as excessive obligate recombination in an individual or pedigree). In cases where two original parents are both untyped for all loci, haplotypes will be displayed for them as usual but it must be noted that the assignments could be reversed (i.e., the two haplotypes assigned to the original father could actually belong to the original mother and vice-versa).

N.B. - at this time the drawing code is not yet complete and while nearly complete, certain pedigree structures (such as those containing marriage loops, inbreeding loops, or individuals with many spouses) may not always be drawn properly. Refer to the results in the haplo.dump file if it appears the pedigree has not been drawn properly.

'Haplotype' is ON when GENEHUNTER is started.

DISCARD Command

Summary: eliminate less informative individuals
Argument: <'on' or 'off'>
Default: displays the current setting

As noted in the "scan" command, some larger pedigrees can be quite time consuming to analyze. To speed this up, some less informative individuals can be discarded without significant loss of information. When the "discard" option is turned on, unaffected individuals that have no descendents in the pedigree and have informative parents (i.e, genotyped) are discarded from analysis. This will alter results somewhat (LOD scores more than NPL statistics since the unaffected individuals are not considered in NPL statistics which measure the degree of sharing among affected individuals) and should only be used

if you are interested in obtaining a fast approximation of the results or if your pedigrees are extremely large and cannot be fully analyzed by GENEHUNTER.

MAX BITS Command (abbreviation 'mb')

Summary: determine how large a pedigree may be analyzed
Argument: <number of bits>
Default: displays the current setting

Because of the time and memory requirements of the mapping algorithms in GENEHUNTER, a maximum pedigree size must be set to keep the computations within the ability of the computer it is running on. The memory and time required are directly proportional to the number of bits in the inheritance vector (number of meioses being examined). This number is 2N - F where F is the number of founders in the pedigree and N is the number of non-founders. For example, a pedigree consisting of two parents and their 4 children would have a size = 2N-F = 6. Entirely uninformative individuals such as individuals in the last generation of a pedigree that are ungenotyped are not included in this figure as they will not be analyzed.

On most workstations, setting the value to 15 or 16 will be a reasonable limit. If pedigrees exceed the size that may be computed under the current 'max bits' setting, individuals may be dropped or the pedigree may be skipped (depending on the setting of 'skip large' - see below). The default setting of 'max bits' is 16.

SKIP LARGE Command

Summary: determine how large pedigrees are dealt with
Argument: <'on' or 'off'>
Default: displays the current setting

Because of the memory and time limitations described in the 'max bits' section, certain pedigrees may not be able to be computed. In this instance a warning message is displayed and one of two things will happen:

if 'skip large' is ON - the pedigree will be skipped over entirely and the computation will continue with the next pedigree in the data set

if 'skip large' is OFF - pedigree individuals will be trimmed off until the pedigree is small enough to be analyzed within the current setting of 'max bits'. This trimming is done such that the maximum amount of linkage information is retained - the first individuals to be eliminated will be unaffected individuals at the bottom of the pedigree as these individuals add very little to the NPL statistic (which measures sharing among affected individuals) and will affect the LOD score somewhat depending on the proposed penetrance of the disease allele.

In either case, it is recommended that for very large pedigrees (where a large number of individuals are not being analyzed) you consider dividing the pedigree into two or more reasonably sized pedigrees that can be analyzed in full.

ANALYSIS Command

Summary: select what type of linkage analysis to perform
Argument: <'NPL', 'LOD', or 'BOTH'>
Default: displays the current setting

The 'analysis' command allows the user to select the method of linkage analysis employed by the scan command. One may select one of three options:

NPL: the 'scan' and 'total' commands will produce only the non-parametric sharing statistics

LOD: the 'scan' and 'total' commands will produce only parametric LOD scores based on the model specified in the locus information file

BOTH: both NPL and LOD scores will be produced

The 'analysis' option is set to BOTH when GENEHUNTER is started.

SCORE Command

Summary: select NPL scoring function
Argument: <'pairs' or 'all'>
Default: displays the current setting

The 'score' command allows the user to select the NPL scoring function to be used during analysis with the 'scan' command. These functions offer a measurement of the degree of sharing among affected individuals and are not dependent on the specific model proposed for the disease as the parametric LOD score is. The statistic reported will represent the deviation from Mendelian expectation observed and will roughly follow the normal distribution.

The 'pairs' function computes a score based on the degree of sharing among all pairs of affected individuals in a pedigree. This statistic is similar to those used in non-parametric sib-pair or APM analyses.

The 'all' function examines all individuals simultaneously and assigns a higher score when more of them share the same allele by descent. It is our experience in extensive simulations and analysis of real pedigree data that the 'all' statistic provides a more powerful test.

POSTSCRIPT OUTPUT Command (abbreviation 'ps')

Summary: activate Postscript graphing capability
Argument: <'on' or 'off'>
Default: displays the current setting

When the "postscript output" option is turned on, the "total stat" command will prompt the user for filenames in which to store postscript graphs for total LOD score, total NPL statistic, and total information content. These files are ready for printing on any Postscript printer and can be displayed by many screen browsers such as Ghostscript. In addition, if the 'haplotype' option is 'on', the scan command will produce pedigree drawings with most likely haplotypes of original individuals and most likely placements of recombinations.

DRAWING SCALE Command (abbreviation 'ds')

Summary: set scale of Postscript 'total' drawings

The 'drawing scale' command allows the user to select the type of scaling used to draw the total NPL, LOD, and information content pictures during the 'total' command. The two options are to have the genetic map (along the x-axis) fill the page, or to set a constant numeric scale in dots per cM. The latter option may be used if you are interested in having the same scale used among different runs of GENEHUNTER for later comparison of output. There are roughly 650 dots available for drawing so a good choice for scale would be roughly 650/(length of largest chromosome). By default, the Postscript drawings will fill the page.

OFF END Command

Summary: Select how far to compute scores beyond ends of map
Argument: <distance>
Default: displays the current value

This command controls how far before the first marker and after the last marker in a map scores will be calculated. For example, if off-end is set to 10.0, then subsequent scan commands will begin calculating scores 10 cM before the first marker and continue stepping through until 10 cM after the last marker. The default value of 'off end' is 0.0 cM. Calling 'off end' with no arguments causes GENEHUNTER to report the current value.

Distances may be specified as either recombination-fractions or centiMorgans, with the necessary assumption that any distance below 0.5 is assumed to be a recombination-fraction and any greater than or equal to 0.5 is assumed to be in centiMorgans.

INCREMENT Command

Summary: Choose the scan step size
Arguments: <'distance' or 'step'> <number>

If 'increment distance 2.0' is entered, the 'scan' command will calculate LODscores and NPL statistics every 2.0 cM throughout the genetic map selected (regardless of the position of markers in that map) as follows (in this example the off end distance is set to 6.0 cM):

-6.0 (6 cM before the first marker), -4.0, -2.0, 0.0 (the position of the first marker), 2.0, 4.0, ...etc...until 6.0 cM after the last locus.

If 'increment step 5' is selected, the scan command will calculate scores at 5 equally spaced positions between each marker. For example, with a three-locus map with 10 and 15 cM intervals and 'off-end' set to 5.0 cM, maps will be computed at the following positions:

-5.0, -4.0, -3.0, -2.0, -1.0 (equally spaced in the 5cM before the first marker)
0.0, 2.0, 4.0, 6.0, 8.0 (equally spaced in the 10 cM interval)
10.0, 13.0, 16.0, 19.0, 22.0 (equally spaced in the 15 cM interval)
25.0, 26.0, 27.0, 28.0, 29.0, 30.0 (equally spaced in the 5cM after the map)

The default value of 'increment' is 'step 5'. Calling 'increment' with no arguments causes GENEHUNTER to report the current value.

Note that the first ('distance') method is not guaranteed to hit every marker position and should be considered inferior to the second ('step') method, which will compute a map at every marker position.

MAP FUNCTION Command

Summary: Choose a cM <-> rec-frac conversion function
Argument: <'haldane' or 'kosambi'>
Default: displays the current value

This command controls which mapping function is used to convert centiMorgans to recombination-fractions and back again both in the input and output of the program and in the internal calculations. Currently only Haldane and Kosambi map functions are available. The default 'map function' is Kosambi.

UNITS Command

Summary: Choose whether scan output is in cM or rec-frac
Argument: <'cM' or 'rec-frac'>
Default: displays the current setting

The 'units' command enables the user to select whether the output from the 'scan' command appears in recombination-fractions (rf) or centiMorgan distance (cM). The conversion function for centiMorgans to recombination fractions can be set using the 'map function' command. When GENEHUNTER is started up, Kosambi centiMorgans are selected as output units.

(3) ADDITIONAL COMMANDS

There are several basic features which GENEHUNTER provides to make the program more friendly and useful. These include on-line help (’help’), the ability to record session output (’photo’), and the ability to accept input from a batch file (’run’).

HELP Command (abbreviation ’?’)

Summary: GENEHUNTER on-line help facility
Argument: <command or topic>

’Help’ displays on-line help information for GENEHUNTER commands and features. Typing ’help’ alone produces a list of available topics and commands. For a general description of a numbered topic, type ’help <number>’, where <number> is the displayed number of the topic. For help on a more specific command or feature, type ’help <name>’, for example:

```
npl:1> help haplotype
```

The on-line help is an exact duplicate of the Postscript reference manual (gh.ps) which accompanies the distribution.

PHOTO Command

Summary: record the output of a session in a file
Argument: <file name>

The "photo" command is used to save a copy of the current GENEHUNTER session (input and output) in a text file. If you type "photo <file name>", for example,

```
npl:1> photo sample.out
```

all input and output from that point on will be copied into the specified file (here, the file named "sample.out"). Typing "photo off" or quitting GENEHUNTER terminates this process and closes the photo file. The default extension for a transcript file is ".out". The ’photo’ command will append program output to the specified file, so output from several sessions may be collected in the same file if desired.

RUN Command

Summary: instruct GENEHUNTER to take input from a file
Argument: <file name>

The "run" command instructs GENEHUNTER to take a series of commands from any text file. This file should contain lines of commands and other input just as they would be typed into GENEHUNTER interactively.

For example, you might want to use a ’run’ file to save setup commands for loading your data:

```
load markers test.loci
increment step 5
postscript on
count recs on
```

```
    haplotype off
```

and could be run with the command

```
    npl:1> run setup.in
```

where 'setup.in' is the name of the file containing the 5 lines of commands above. This feature is especially useful for providing input to GENEHUNTER during long runs on data files with many pedigrees which you may wish to let run overnight or at least without any user input.

SYSTEM Command

Summary: execute a command under the operating system
Argument: <system command>

The 'system' command is used to temporarily interrupt GENEHUNTER and start up a new command interpreter from the operating system. Commands which are normally typed to the operating system may then be issued. You can return to GENEHUNTER by typing 'exit' or control-D in most operating systems. If an argument is supplied to 'system', the argument is interpreted just as a normal command issued to the operating system. For example:

```
    npl:4> system lp results.out
```

would execute the printing command on your operating system and then return control immediately to GENEHUNTER.

CHANGE DIRECTORY Command (abbreviation 'cd')

Summary: change the current directory
Argument: <new directory>

The 'cd' command works essentially the same way it does under Unix. By default, all files are read or written from the current directory unless specified otherwise.

TIME Command

Summary: display the current time
No Arguments

Display the current time from the system clock.

QUIT Command (abbreviation 'q')

Summary: exit session
No Arguments

Assures that the program exits properly.

GENEHUNTER 1.0 COMMAND REFERENCE:

(1) DATA PREPARATION COMMANDS 1
LOAD MARKERS Command 1
USE Command 1

(2) GENEHUNTER MAPPING COMMANDS 3
SCAN PEDIGREES Command 3
TOTAL STAT Command 4
SINGLE POINT Command 4
COUNT RECS Command 4
HAPLOTYPE Command 5
DISCARD Command 5
MAX BITS Command 6
SKIP LARGE Command 6
ANALYSIS Command 7
SCORE Command 7
POSTSCRIPT OUTPUT Command 7
DRAWING SCALE Command 8
OFF END Command 8
INCREMENT Command 8
MAP FUNCTION Command 9
UNITS Command 9

(3) ADDITIONAL COMMANDS 10
HELP Command 10
PHOTO Command 10
RUN Command 10
SYSTEM Command 11
CHANGE DIRECTORY Command 11
TIME Command 11
QUIT Command 11

GENEHUNTER 1.0 QUICK REFERENCE:

(1) DATA PREPARATION COMMANDS
load markers.....Load marker-locus data
use.....Select the current map for analysis

(2) GENEHUNTER MAPPING COMMANDS
scan pedigrees.....Analyze pedigree data
total stat.....Show total scores from a scan of multiple pedigrees
single point.....activate/deactive single-point analysis
count recs.....turn recombination counting on
haplotype.....determine likely haplotypes for individuals
discard.....eliminate less informative individuals
max bits.....determine how large a pedigree may be analyzed
skip large.....determine how large pedigrees are dealt with
analysis.....select what type of linkage analysis to perform
score.....select NPL scoring function
postscript output.....activate Postscript graphing capability
drawing scale.....set scale of Postscript 'total' drawings
off end.....Select how far to compute scores beyond ends of map
increment.....Choose the scan step size
map function.....Choose a cM <-> rec-frac conversion function
units.....Choose whether scan output is in cM or rec-frac

(3) ADDITIONAL COMMANDS
help.....GENEHUNTER on-line help facility
photo.....record the output of a session in a file
run.....instruct GENEHUNTER to take input from a file
system.....execute a command under the operating system
change directory.....change the current directory
time.....display the current time
quit.....exit session

* = reference information only - not a command