(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.

See 'help variance components' for information on how phenotype and covariate data should be specified in this file.

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 ...  4.10 0.374
    (a)   (b) (c) (d) (e) (f) (g)     (h ------------------------)     (i -------)
```

(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
(i) phenotype/covariate data (OPTIONAL)

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.

A - in the phenotype/covariate data indicates missing data - NB: 0 is a real value that a phenotype may take on and DOES NOT represent missing phenotype data

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.

LETTERS Command

Summary: controls allele display in Postscript output
Argument: <'on' or 'off'>
Default: displays the current setting

When the 'haplotype' and 'postscript' options are both turned on, the 'scan' command produces postscript pedigree drawings with the most likely haplotypes of original individuals displayed. If 'letters' is on, these haplotypes will be drawn as letters representing the founder chromosome inherited rather than the numeric genotypes themselves. Upon startup of Genehunter, 'letters' is off and these drawings will display the actual alleles inherited.

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.

DISPLAY SCORES Command

Summary: activate screen display of scores and haplotypes
Argument: <'on' or 'off'>
Default: displays the current setting

'Display scores' is ON when GENEHUNTER is started.

(3) SIBS QUALITATIVE TRAIT MAPPING COMMANDS

Commands to map loci using affectation status.

ESTIMATE Command

Summary: maximum likelihood estimate of IBD sharing
No Arguments

Usage -- To run the command just type 'estimate'; no arguments are needed. GENEHUNTER will first ask you if you want to analyze your data under the assumption of no dominance variance, or under the assumption of dominance variance where Holman's triangle is applied:

```
    analyze under the assumption of no dominance variance? y/n [n]
```

The default is to perform the analysis with the assumption of dominance variance.

GENEHUNTER will then query you for the filenames to store the text and postscript output respectively. You will be alerted if either of the chosen filenames already exist.

Output -- The text file consists of the columns:

```
           <position>    <z0>    <z1>    <z2>  <loglike>
```

where the z-values are the calculated maximum likelihood proportions. At the end of the text output is a time-stamped summary of the session settings when the analysis was run. The first postscript output file is a graph of position vs. loglike and the second is a plot of how the maximum likelihood sharing proportions change across the region. The marker names are given along the x-axis and the distance examined in the analysis is given at end of the x-axis (this may be larger than the map distance if you have specified an off-end distance).

Background -- 'estimate' scans the selected map region and identifies regions of significant excess allele sharing.

Note that the LOD score is never negative, because the maximum likelihood solution for z0, z1, and z2 can never be worse than the Mendelian segregation expectation.

EXCLUDE Command

Summary: exclusion mapping
Arguments: <relative risk ratio hypotheses>

Usage -->
command line:

```
    npl:6> exclude
```

You are then given the option of inputting a set of z's or relative risk values (the input queries will be different depending on whether you want to analyze your data under the assumption of no dominance variance or not).

Output --> The text file consists of tabbed columns in the format:

```
    position    z2-1    z2-2    z2-3 ... etc.
```

(You should be able to use this file as input to a plotting program if you don’t have access to a postscript printer.) At the end of the text file is a time-stamped summary of the session settings. The postscript file consists of multiple y-axis LOD score plots for each relative risk value/set of z’s and gives the distance examined in the analysis at end of the x-axis (this may be larger than the map distance if you have specified an off-end distance). A horizontal dashed line is drawn at the traditional exclusion criterion of Z < -2.

Background --> Exclusion mapping is used to identify and exclude regions unlikely to have a major effect on the trait you are mapping. GENEHUNTER does this by comparing the likelihood of the observed sharing proportion of 0, 1 and 2 alleles between affected sibs (z0,z1,z2), to the likelihood under the Mendelian expectation of a0=1/4, a1=1/2, and a2=1/4. When using GENEHUNTER under the asssumption of no dominance the sharing proportions are given by:

$$z0 = a0/Ls$$
$$z1 = a1$$
$$z2 = a2((2Ls-1)/Ls)$$

where Ls = lambda-sub-S, the relative risk ratio for a sib, defined as:

$$\frac{\text{prevalence of the trait in siblings of affected individuals}}{\text{prevalence of the trait in the population at large}}$$

Note that Ls = 1 when there is no observed difference in prevalence of sibs vs the population (z0=a0, z1=a1, z2=a2 and LOD = 0). If Ls < 1, it would imply that there was some protective advantage in having an affected sib. Since neither of these cases are interesting and/or reasonable, only Ls values > 1 are allowed. (The no dominance variance assumption allows us to simplify the sharing proportions above to the one variable Ls. With dominance variance Ls = Lo where Lo = relative risk ratio for an offspring, and Lm-1=2(Ls-1) where Lm is the relative risk ratio for a monozygotic twin.)

The likelihood under Bayes theorem is:

$$L(pos) = (z0*p0+z1*p1+z2*p2) / (a0*p0+a1*p1+a2*p2)$$

and the LOD score is calculated by summing log10(L(pos)) across pedigrees for each position.

The relations for z0, z1, and z2 above hold if multiple loci are involved in the trait, provided that the loci interact multiplicatively and the lambda values are defined as the component of the relative risk attributable to the locus.

More details on the analytical method are present in the publication

(4) SIBS QUANTITATIVE TRAIT LOCI (QTL) MAPPING COMMANDS

Commands to map loci using numerical phenotype scorings.

HASEMAN ELSTON Command

Summary: traditional & EM Haseman-Elston analysis
No Arguments

Usage -- After typing the command you will be queried as to which phenotype you want to analyze (if you have loaded more than one), and then queried for files to store the text output for the traditional haseman-elston and EM haseman-elston analyses, as well as the filename for the postscript output.

Output -- The traditional and EM haseman-elston output files have the columns:

```
   <position>     <beta>     <LOD>     <t>
```

At the bottom of each of these text output file is a time-stamped summary of the session variables when the command was run. This summary will also list which phenotype was selected and in the case of the EM algorithm, the convergence limit that was used. The postscript output file has a plot of both the traditional and EM results.

Note1: The EM algorithm has been found to have very rare instabilities in large intervals between markers; if there is a sudden peak in the EM plot make sure a similarly shaped peak also appears in the traditional haseman-elston results. (The nonparametric method does not have these instabilities either and can also be used to verify your results.)

Note2: In order to run this command you must have selected more than two pedigrees/pairs -- which shouldn’t be a problem since it won’t be very significant using any less!

ML VARIANCE Command

Summary: maximum likelihood QTL variance estimation
No Arguments

Usage -- After typing the command you will be queried as to which phenotype you want to analyze (if you have loaded more than one), and then queried for a files to store the text and postscript output.

Output -- The text output file has the format:

```
   <position>     <LOD>     <sigsq0>     <sigsq1>    <sigsq2>
```

At the bottom of each of these text output file is a time-stamped summary of the session variables when the command was run. This summary will also list which phenotype was selected and the convergence limit that was used. The postscript output file is a plot of position vs. LOD.

Note: This EM-based algorithm has been found to have very rare instabilities in large intervals between markers; if there is a sudden peak in the plot you can verify it by checking it against the results of the nonparametric method, which is not subject to the same instabilities.

NO DOM VAR Command

Summary: maxlike QTL variance est. under no-dominance assmp.
No Arguments

Usage -- After typing the command you will be queried as to which phenotype you want to analyze (if you have loaded more than one), and then queried for a files to store the text and postscript output.

Output -- The text output file has the format:

   <position>   <LOD>   <sigsq0>   <sigsq1>   <sigsq2>

At the bottom of each of these text output file is a time-stamped summary of the session variables when the command was run. This summary will also list which phenotype was selected and the convergence limit that was used. The postscript output file is a plot of position vs. LOD.

Note: This EM-based algorithm has been found to have very rare instabilities in large intervals between markers; if there is a sudden peak in the plot you can verify it by checking it against the results of the nonparametric method, which is not subject to the same instabilities.

NONPARAMETRIC Command

Summary: non-parametric QTL analysis
No Arguments

Usage -- After typing the command you will be queried as to which phenotype you want to analyze (if you have loaded more than one), and then queried for a files to store the text and postscript output.

Output -- The text output file has the format:

   <position>   <Z-score>

At the bottom of each of these text output file is a time-stamped summary of the session variables when the command was run. This summary will also list which phenotype was selected and the convergence limit that was used. The postscript output file is a plot of position vs. Z-score.

of position vs. Z-score.

PAIRS USED Command

Summary: select what pair combinations will be used
No Arguments

If you have loaded more than two sibs in any of your sibships this command allows you to include the extra sibs in the analysis commands (all sibs are automatically included for phase information if parents are missing). When using 'all pairs', each pair is considered as an independent pedigree but a weight (2/num_affecteds) is factored in to counteract inflation of significance due to the statistical dependence among these pairs.

Simply type 'pairs used' and indicate which pair setting you would like to use:

```
sibpair:1> pairs used

the current pair setting is: *first affected/phenotyped sibpair only*

  Possible pair options:
   1. First pair of affected/phenotyped sibs
   2. All independent pairs of affected/phenotyped sibs*
   3. All pairs of affected/phenotyped sibs*
   4. All pairs pf affected/phenotyped sibs-UNWEIGHTED

Enter the index of the analysis you want to use [1]: 2
```

*"independent" pairs of sibs are created by taking the first sib paired with sibs 2...n (for a three-sib sibship this will mean the sharing for pairs 1-2 & 1-3 will be computed). Therefore, the results can be different if you rearrange the order of the sibship. "all" pairs are created by taking the first sib paired with sibs 2...n, the second sib paired with 3...n, etc. For a four-sib sibship this means the sharing for pairs 1-2, 1-3, 1-4, 2-3, 2-4 and 3-4 will be computed. The sibs are considered as part of a whole family when inheritance vectors are determined and then each pair is treated as a essentially a separate pedigree for the purposes of analysis.

You DO NOT need to re-scan for a change in the pair setting to take effect.

The default is to use the first pair of affected/phenotyped sibs.

DUMP IBD Command

Summary: dump the ibd distribution to a text file
No Arguments

This command allows you to output the calculated likelihood of sharing 0, 1 or 2 alleles for each relative pair within each pedigree, possibly for use in another program. (You will be queried for the filename to store it in.)

The output format is:

```
<pos> <pedigree> <indiv1-indiv2> <priorz0> <priorz1> <priorz2> <z0> <z1> <z2>
```

This command has been expanded from the original MAPMAKER/SIBS command to include all non-founder relative pairs (regardless of relationship or affected status).

(6) VARIANCE COMPONENTS
or affected status).

VARIANCE COMPONENTS Command

Summary: run variance components analysis
No Arguments

This command looks for evidence of quantitative trait loci (QTLs). At each scan position, the program determines whether a significant amount of the variance in a quantitative trait can be attributed to a QTL at that position. Specifically, it calculates maximum likelihood values for the mean trait value (separately for each sex, if desired), additive and dominance variance components for the QTL, additive and dominance variance components for other, unlinked loci, and an environmental variance component. One or both of the dominance components can be optionally excluded. In addition, the program can incorporate covariate effects by estimating the regression of the trait value on a given covariate value. The significance of the QTL effects is tested by comparing the maximum likelihood model with another one in which the QTL variance components are constrained to equal zero. The likelihood ratio of the two models is used to calculate a LOD score which can be compared to a chi-squared distribtion as in classical methods of QTL analysis.

Data preparation
Phenotype values should be included in the pedigree file, after the genotype values for each individual. Covariates should be listed immediately after the phenotypes. Multiple values may be entered, up to the numbers given by the constants MAX_PHENOTYPES and MAX_COVARIATES in npl.h. Each phenotype should be indicated in the map file by a single line reading "0 2" followed by five empty lines (These are needed to maintain consistency with the LINKAGE file format. The data expected by LINKAGE in these lines are not used by Genehunter and can be excluded). Each covariate should be indicated with a single line reading "4 0". In addition, the total number of loci (the first number on the first line of the map file) should include the number of phenotypes and covariates, as well as the number of markers and qualitative traits.

Running the program
When the "variance components" command is entered, the user is prompted for names for the output files and is then asked whether to include dominance variance components for the unlinked loci and for the QTL. The user is then given the option of entering starting estimates for the model's parameters, rather than letting the program come up with it's own estimates. This option is provided because the program's ability to converge on the maximum likelihood values is sometimes sensitive to the starting guesses. Trying out different starting values and seeing whether the same result is obtained provides a check on the correctness of the results. This should probably be done with all analyses, but is especially needed if the program is yielding odd results, such as negative or unrealistically high LOD scores. If manual input is chosen, the program first displays the total variance of the trait value being examined, as well as the mean trait value (separately for males and females if this option has been chosen). These figures can be helpful in choosing starting values.

Output
The output file shows a LOD score for each scan position, along with estimates of the means, variance components, and covariate regression coefficients for that position. The corresponding estimates for the null model are also reported. Because the program can sometimes fail to converge on an estimate for some positions, the output indicates for each position whether convergence occurred. When it does not occur, the output shows the estimates for the last position which did converge. If the program frequently fails to converge, it may be necessary to raise

the maximum number of iterations allowed in the estimating algorithm (MAXITS in the file varcom.c). If postscript output is switched on, the program also produces two graphics files. One contains a plot of LOD score versus position, and the other a plot of the proportion of total phenotypic variance accounted for by each component of the maximum likelihood model, versus position.

SET STARTING VALUES Command

Summary: choose method for initial estimates of parameters
No Arguments

This command determines how the variance components command makes its initial parameter estimates. To use, enter "set starting values" and choose the desired option:

```
   npl:1> set starting values

   Genehunter currently uses a constant fraction of total phenotypic variance.

 Possible starting values:
  1. ML estimate from adjacent position
  2. Constant fraction of total phenotypic variance
   Enter the index of the start values you want to use [2]:
```

The first method simply divides the total trait variance evenly among the variance components of the model. Thus it uses the same initial values for each position. The second method does this for the first position, but thereafter uses the maximum likelihood values for the last position. The second method is generally faster, because adjacent postions usually have similar maximum likelihood estimates, hence the algorithm requires fewer iterations to converge when it starts near its destination. However, the method can sometimes prevent convergence on the true maximum likelihood estimate, instead settling on a local maximum near the maximum likelihood values of the adjacent position. The first method, slower but more reliable, is the default.

MEANS BY SEX Command

Summary: choose whether to estimate means by sex
No Arguments

This command determines whether the variance components command estimates means separately by sex. To use, simply enter "means by sex" and indicate the desired setting:

```
   npl:1> means by sex

   Genehunter currently estimates male and female means separately.

 1. Estimate a single mean
 2. Estimate male and female means separately
   Enter the index of the option you want to use [2]:
```

The default setting of [2] should improve the method’s power to detect linkage when sex actually has an effect on the trait’s value. Setting [1] can slightly speed things up when no such effects are thought to exist, or when insufficient data exist for one sex.

(7) TDT COMMANDS

GENEHUNTER now contains a standard implementation of the transmission disequilibrium test (TDT) along with several extensions for using missing data and estimating significance via simulation.

TDT Command

Summary: standard single locus TDT
Argument: <file name>

The 'tdt' command performs the traditional Transmission Disequilibrium Test (Spielman, McGinnis, and Ewens, Am J Hum Genet. 1993 Mar;52(3):506-16.) on the Linkage-style pedigree file specified as the argument. Transmissions from homozygous parents are not counted (the obligately provide a transmitted and untransmitted copy of the same allele) and cases where one parent is missing are used only when the genotyped parent and the proband are both distinct heterozygotes (Curtis and Sham, Am J Hum Genet. 1995 Mar;56(3):811-2.) In addition, the transmissions and non-transmissions are stored for use by multi-locus TDT commands (tdt2, tdt3, tdt4). The case where the both parents and the proband have the same heterozygous genotype are counted (as a transmission and non-transmission of each allele) but are not stored for use in the multi-locus test.

TDT2 Command

Summary: two locus TDT
Argument: <offset between markers to examine>
Default: analyze adjacent markers (offset=1)

The 'tdt2' command computes the two-locus version of the TDT. The identical rules for counting transmissions and non-transmissions are employed and as in the standard single marker TDT. If an offset is provided as an argument, the analysis will be done on pairs of markers as follows (1 and 1+offset, 2 and 2+offset, 3 and 3+offset, etc.). By default, offset is set to 1 so with no argument specified, 'tdt2' will produce a two-locus TDT test for marker pairs in map order (1 and 2, 2 and 3, 3 and 4, etc.) By nature, this model assumes there is no recombination between adjacent markers (or at least not a significant amount) which would interfere with the detection of potential founder haplotypes. Therefore it is probably most useful on closely spaced markers and/or in more recently founded populations. This command is only available after the 'tdt' analysis of a pedigree file.

TDT3 Command

Summary: three locus TDT

Computes a three-locus TDT (see 'tdt2' for a more details about multi-locus TDTs). This command is only available after the 'tdt' analysis of a pedigree file.

TDT4 Command

Summary: four locus TDT

Computes a four-locus TDT (see 'tdt2' for a more details about multi-locus TDTs). This command is only available after the 'tdt' analysis of a pedigree file.

PERM1 Command

Summary: permutation test for determining TDT significance
Argument: <number of simulations>

Since the standard application of the TDT usually involves the analysis of numerous alleles at numerous markers, a significant correction is required to interpret the significance of any one result. Treating the num_markers x num_alleles tests as independent is extremely conservative since a) the tests of the alleles at each marker are not independent and b) there will be very rare alleles which will penalize an additional degree of freedom without any chance of providing results of interest. A better test of significance is provided by a permutation method in the 'perm1' command as follows:

* create a new data set by taking each pair of transmitted and untransmitted alleles and arbitrarily (at p=0.50) reversing the assignment of which was transmitted

* tally and store the results of the TDT for this new data set

* repeat 1000 or more (the number of simulations is indicated in the argument to 'perm1'), comparing each simulated data set to the actual results observed in the real data set

After the simulations are completed, a report indicating

* how many of the permuted data sets had a higher maximum value and
* how many of the permuted data sets had more results above certain thresholds (.01, .001)

is displayed providing a better estimate of the significance of the observed data. This command is only available after the 'tdt' analysis of a pedigree file.

```
PERMUTATION SUMMARY:
12 of 1000 simulations had a larger maximum value than the real best (15.42)
48 of 1000 simulations had as many tests (22) exceeding p=.01
19 of 1000 simulations had as many tests (3) exceeding p=.001
```

PERM2 Command

Summary: permutation test for determining TDT significance
Argument: <number of simulations>

This test performs the same permutation test as in 'perm1' but instead examines all permutations of all two locus haplotypes formed by adjacent markers and markers separated by 1 in the current map order. The results can be interpreted as in the 'perm1' command. This command is only available after the 'tdt' analysis of a pedigree file.

(8) 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:

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
letters..................controls allele display in Postscript output
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
display scores...........activate screen display of scores and haplotypes

(3) SIBS QUALITATIVE TRAIT MAPPING COMMANDS
estimate.................maximum likelihood estimate of IBD sharing
exclude..................exclusion mapping

(4) SIBS QUANTITATIVE TRAIT LOCI (QTL) MAPPING COMMANDS
haseman elston...........traditional & EM Haseman-Elston analysis
ml variance..............maximum likelihood QTL variance estimation
no dom var...............maxlike QTL variance est. under no-dominance assmp.
nonparametric............non-parametric QTL analysis

(5) OTHER SIBS COMMANDS
pairs used...............select what pair combinations will be used
dump ibd.................dump the ibd distribution to a text file

(6) VARIANCE COMPONENTS
variance components......run variance components analysis
set starting values......choose method for initial estimates of parameters
means by sex.............choose whether to estimate means by sex

(7) TDT COMMANDS
tdt......................standard single locus TDT
tdt2.....................two locus TDT
tdt3.....................three locus TDT
tdt4.....................four locus TDT
perm1....................permutation test for determining TDT significance
perm2....................permutation test for determining TDT significance

(8) 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
```