Allegro Version 2.0
Manual

Daniel F. Gudbjartsson
Decode Genetics, Reykjavik, Iceland
10th October 2005

Introduction

Allegro is a computer program for multipoint genetic linkage analysis and related calculations. Allegro can do both classical parametric linkage analysis and analysis based on allele sharing models. In addition, Allegro estimates total number of recombinations between markers, computes posterior IBD sharing probabilities, reconstructs haplotypes and does two types of simulation. Thus Allegro includes the basic functionality of the well known Genehunter program (Kruglyak et al. 1996). It can analyse pedigrees of moderate size, and it can handle many markers (as opposed to programs such as Linkage and Fastlink, which do parametric analysis of large pedigrees, but only with a few markers). The biggest advantages of Allegro over Genehunter are the allele sharing models that it provides and a much shorter execution time.

If you publish results obtained with Allegro, please cite: 'Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsdottir A (2005) Allegro version 2. Nature Genetics. 37:1015–1016'.

This manual is accompanied by a report, describing the computational algorithms employed in Allegro and giving an overview of the statistical models, 'Fast multipoint linkage analysis and the program Allegro' by Kristjan Jonasson, Daniel F. Gudbjartsson, Michael L. Frigge and Augustine Kong. It should have been delivered with the manual; if not send electronic mail to allegro@decode.is. It will be referred to as the 'Allegro report' in the remainder of this manual.

Allele sharing models were described by Kong and Cox (1997), and together with Michael L. Frigge, they created a modified version of Genehunter, called Genehunter-Plus to calculate allele sharing LOD scores. In addition to allele sharing LOD scores, Allegro calculates parametric LOD scores and for compatibility with Genehunter, NPL scores.

The computational engine of Allegro is based on the same inheritance formulation as Genehunter, and it uses the same Fourier transforms (see Kruglyak and Lander 1998) to calculate convolutions. Several improvements to the algorithms used in Genehunter are described in the Allegro report and used in Allegro. As a result, Allegro runs much faster than Genehunter (20–100 times depending on the data and the analysis performed). In addition considerable reduction in required memory may be achieved through automatic recalculation and/or disk-swapping (at a slight cost in run-time).

Allegro has several other advantages over Genehunter. An effort has been made of making the input and output flexible. There is no fixed limit on the number of markers or the number of non-founders, whereas Genehunter has a limit of 50 markers and 16 non-founders. Allegro offers the scoring functions S_{pairs} and S_{all} as Genehunter, but in addition a scoring function for homozygosity mapping and two scoring functions described by McPeek (1999) are supplied. Allegro can also run almost any number of models in a single run.

This manual is in two parts; part 1 describes all the basic functionality of Allegro through a simple example and part 2 is an alphabetic reference of Allegro options. The file 'options.h' that is part of the distributed Allegro

source code begins with a header comment containing condensed information on all the options of Allegro, including some that are not documented in this manual.

1 Example application

The example presented here is quite simple, it consists of the analysis at two markers of a single nuclear family with two sibs. The files used are found in the directory '`examples/ex1`' where Allegro was set up. Originally, this directory contains four files, `README`, with some documentation, an option file, `ex1.opt`, and two input files, `ex1.pre` and `ex1.dat`. To run Allegro, make `ex1` the working directory and issue the command `allegro ex1.opt`. Allegro will just take a second to run, and 18 output files will be created, as described in the following sections. Allegro will print some information on the run to the screen and to a log file named `allegro.log`.

1.1 Option file

The tasks that Allegro performs are controlled through the option file. The option file is also used to change parameters, such as the maximum amount of memory Allegro is allowed to allocate or the number of inter-marker locations where statistics (such as LOD scores) are calculated.

Option files may contain comment lines, which are preceded by a % sign, and lines may also be completely blank. Each non-comment and non-blank line starts with a *keyword* which is usually followed by one or more *arguments*. In this manual the keywords are typeset in an upper case typewriter font and the arguments are typeset in a lower case typewriter font. Arguments are either restricted to a set of fixed values (e.g. restricted to be '`mpt`' or '`spt`') or free (e.g. numbers or file names). Restricted arguments may take parameters (e.g. '`freq:0.1`'). Keywords and restricted arguments are not case sensitive, but parameters and free arguments, such as file names, are case sensitive. Finally, a comment may be placed at the end of a line.

The sample option file, `ex1.opt`, is as follows:

```
% Read input in Linkage style format:
PREFILE ex1.pre
DATFILE ex1.dat

% Linkage analyses to be performed:
MODEL mpt par het
MODEL spt exp pairs equal
MODEL mpt exp pairs equal
MODEL mpt lin all power:0.5

% Other statistical analysis to be performed:
HAPLOTYPE
CROSSOVERRATE

% Other options:
UNINFORMATIVE   % Write uninformative markers to uninformative.out
MAXMEMORY 100   % Maximum memory set to 100 Mb
```

The input is read from the Linkage style input files `ex1.pre` and `ex1.dat` as further described in the next section. The four '`MODEL`' lines specify the linkage analyses to be performed, one classical parametric analysis and three allele sharing analyses. They are described in section 1.3. The '`HAPLOTYPE`' option instructs Allegro to reconstruct haplotypes, '`CROSSOVERRATE`' specifies that the actual number of crossovers between adjacent markers should be estimated from the genotype data and written to a file, and '`UNINFORMATIVE`' instructs Allegro to write a list of uninformative markers to a file. Finally, the '`MAXMEMORY`' option specifies that

Allegro should allocate no more than 100 Megabytes of CPU memory. The effect of these options is described in sections 1.4 and 1.6.

1.2 Linkage style input files

Allegro accepts input in either of two styles; 'Linkage format' and 'new format'. The Linkage format is described in detail in the book by Terwilliger and Ott (1994), and it is the same format as that used by Genehunter (Kruglyak et al. 1996). A rough description of the Linkage format is given in the reference part of this manual, and this section contains a simple example. Currently, however, the 'new format' is not described in this manual (this will hopefully be remedied soon), but the previously mentioned comments in the file `options.h` in the Allegro source contain some information. Thus Linkage format will be assumed in the rest of this manual.

The Linkage format input consists of two files, a *prefile*, and a *datfile*. The prefile contains the pedigree structure, affection status and genotypes, and the datfile contains marker map, allele frequencies, and parameters for the parametric model (disease allele frequency and penetrances).

The example prefile contains information on a single affected sib pair at two markers, and is as follows:

```
f1  father   0       0      1 0  0 0  0 0
f1  mother   0       0      2 0  0 0  0 0
f1  daughter father  mother 2 2  1 2  1 2
f1  son      father  mother 1 2  1 1  1 2
```

The first column gives the family identifier, the next three columns give the family structure, the fifth column specifies the sex, and the sixth column contains an affection status (0 is unknown, 1 is unaffected and 2 is affected). The rest of the file contains genotype data. Here the father and the mother are not genotyped, the daughter is heterozygous with alleles 1 and 2 at both markers, and the son is homozygous with alleles 1 and 1 at the first marker, and heterozygous with alleles 1 and 2 at the second marker.

The datfile used in the example is:

```
3 0 0 5
0 0.0 0.0 0
2 3
1 2
0.9 0.1
1
0.09 0.90 0.90
3 5 # Marker1
0.10 0.20 0.30 0.20 0.10
3 3 # Marker2
0.10 0.30 0.60
0 0
1.7
1 0.10000 0.450000
```

In this example the first marker has 5 alleles with frequencies 0.1, 0.2, 0.3, 0.2 and 0.1, the second marker has 3 alleles with frequencies 0.1, 0.3 and 0.6 and the markers are spaced 1.7 centi-Morgans apart. The disease allele frequency is 0.1 and the penetrances are 0.09, 0.9 and 0.9, with no disease allele, one disease allele and two disease alleles respectively (i.e., a dominant disease). For a few more details, see the DATFILE entry in the reference part of this manual.

To specify a sex-specific marker distances, two lines containing distances should be given and the `SEXSPECIFIC on` option should be given in the options file. The first line gives the distances for males and the second line gives the distances for females.

1.3 Linkage analysis

The first MODEL line in the option file of section 1.1 instructs Allegro to perform classical parametric analysis. The argument '`mpt`' specifies that multipoint analysis is to be performed. Its alternative, '`spt`', specifies single point analysis. Multipoint means that all the markers are analyzed in conjunction; the LOD-score at a marker is calculated taking genotypes at neighbouring markers into account. Single point, however, means that each marker is analyzed independently of all the other markers; a LOD-score corresponding to the disease gene being at the marker is calculated based only on the genotype data at the marker (ignoring data at other markers). Single point analysis is really two point analysis, one point being the the marker of interest and the other the disease locus. The argument '`het`' specifies that a heterogeneity parameter is to be estimated, and a heterogeneity LOD score evaluated.

The combined results of the analysis will be written to a file named `param.mpt`. An additional file, called `fparam.mpt`, which contains the results broken down by families, will also be written. To override this default naming of the output files add the desired names at the end of the MODEL line (e.g., '`MODEL mpt par result.out fresult.out`').

Allele sharing models are run using the other three MODEL lines. The arguments '`exp`' and '`lin`' specify that an allele sharing model is to be run; '`lin`' specifies the linear model and '`exp`' the exponential model of Kong and Cox 1997. The exponential and linear models are similar for small genetic effects, but can differ substantially for large effects.

The argument '`pairs`' specifies that the S_{pairs} scoring function should be used, and '`all`' similarly that the S_{all} scoring function should be used. These two scoring functions (which are also offered by Genehunter) are discussed by Whittemore and Halpern (1994), McPeek (1999) and in the Allegro report. Other scoring functions currently implemented in Allegro are S_{homoz}, S_{robdom}, S_{mnallele} and S_{ps}. S_{homoz} is like S_{pairs}, except that when there is inbreeding and an individual is homozygous IBD he may contribute to the scoring. The definition of S_{mnallele} and S_{robdom} may be found in McPeek (1999). She compares all these scoring functions (and others), and in short she recommends S_{pairs} as a compromise choice for a scoring function that performs well over all disease models, S_{robdom} when the disease is dominant and S_{mnallele} when it is recessive. S_{ps} is described in Karason et al. (2002) and allows the user to put weights on different types of meioses.

The argument '`equal`' specifies that the families should be weighted equally. The argument '`power:0.5`' specifies that each family should be weighted proportionally to the standard deviation of the score function used, under the null hypothesis of no linkage, to the power 0.5. The power is the s of section 5.1 in the Allegro report. The third possibility is to give the name of a file containing weights for all the families in the analysis.

The results of the three allele sharing analyses are written to six files in all, named `exppairs.spt`, `fexppairs.spt`, `exppairs.mpt`, `fexppairs.mpt`, `linall.0.5.mpt` and `flinall.0.5.mpt`. The files beginning with `f` contain results broken down by families, and the other files contain combined results (note that in this simple example there is only one family). The 1 in the name of the last two files comes from the value of the power. As with the parametric analysis, one may override the file name defaults by adding alternative file names at the end of the MODEL lines.

1.4 Haplotyping and crossover rate

The haplotype reconstruction that Allegro offers is described shortly in the Allegro report. The modelling used is the same as that of Genehunter and the program Simwalk2 (Sobel et al. 1996), but the algorithm of Simwalk2 finds an approximate solution to the model, and the algorithm in Genehunter are much slower than Allegro's.

The line '`HAPLOTYPE`' in the options file instructs Allegro to do haplotyping. Four output files will be written, their default names are: `haplo.out`, `ihaplo.out`, `founder.out` and `inher.out`. The names of the files can be changed by specifying alternative names as arguments.

Allegro reconstructs haplotypes by first finding the most likely inheritance vector path (a specification of the inheritance vector at each marker), and then finding the most likely genotype assignment given the likeliest path. This method is not guaranteed to produce the most likely haplotype, but should in most cases either do that, or produce a good approximation.

The output file `inher.out` lists the most likely inheritance vector path and `founder.out` has similar information; in it each founder chromosome is enumerated and, under the most likely inheritance vector path, the inheritance of each founder chromosome down through the pedigree is shown. The file `haplo.out` contains the haplotype assignment everywhere where there are genotypes in the input prefile. The file `ihaplo.out` contains, in addition, haplotypes elsewhere, if they can be imputed. For more details, see 'HAPLOTYPE' in the reference part.

Allegro calculates the 'observed crossover rate' between markers, which actually means that Allegro calculates the expected total number of crossovers that did occur between the markers, given all the data, and divides in to this by the number of meioses that are informative for crossovers. This is a little more consistent than Genehunter's 'observed map'; see the Allegro report for details.

The line '`CROSSOVERRATE`' instructs allegro to calculate the observed crossover rate. This rate is calculated both per family, and for all the families combined; the per family rate is written to the file fxover.out and the combined rate is written to xover.out. As before, these default file names may be overridden by providing file names as arguments to `CROSSOVERRATE`.

1.5 IBD sharing

Computation of pairwise IBD sharing probabilites is effected with the '`PAIRWISEIBD`' option. For these probabilities a pair of individuals is conidered to share 1 if their (IBD) genotypes are a/b and a/c, or if they are a/b and a/a; they share 2 if the genotypes are a/b and a/b or if they are a/a and a/a; otherwise they share 0. Note that this is not the same definition of sharing as that underlying the S_{pairs} scoring function, for that the sharing is 2 for genotypes a/b and a/a, and 4 for a/a and a/a. Both prior and posterior probabilities are calculated, the prior probability of a given sharing depends only on the pedigree (for sibs $P_{\text{pairs}}(0) = 0.25$, $P_{\text{pairs}}(1) = 0.25$ and $P_{\text{pairs}}(2) = 0.25$), but the posterior probability of a given sharing is the probability of this sharing given the pedigree and the genotype data. It is possible to calculate both single point and multipoint sharing probabilities, and this may be done for all pairs or a selection of pairs, for example all genotyped pairs. Each '`PAIRWISEIBD`' line causes two files to be created, one with the prior probabilities and the other with the posterior probabilities. Their default names are `prior.mpt`/`prior.spt` and `posterior.mpt`/`posterior.spt`.

1.6 Other options

The option '`MAXMEMORY 100`' sets a roof on the amount of memory Allegro can allocate to 100 Mb. Allegro estimates how much memory it needs and compares it to the roof set by `MAXMEMORY`. If Allegro needs too much memory it reduces its need by recalculating some statistics and, if necessary, swapping to disk (see `SWAPDIRNAME` in the reference part).

In a normal run, it is likely that some markers are completely uninformative for estimating the distribution of inheritance vectors (and thus the IBD sharing between relatives), at least in some of the families. The option '`UNINFORMATIVE`' instructs Allegro to write all marker-family pairs that are uninformative to a file, by default named `uninformative.out`.

1.7 Output

As already said, running Allegro with the option file of section 1.1 produces 18 output files in all (a log file, 2 output files for each of the four models, 4 haplotyping files, 2 crossover rate files and the list of uninformative

markers). All these files except the log file have already been mentioned, but in this section their format will be covered in some detail. The log file is called `allegro.log`, and it should contain the same information as Allegro prints to the screen, but with a date-time stamp at the beginning of each line. If a log file already exists Allegro appends to it. It is neither possible to change the name of the log file nor prevent Allegro from writing it.

The combined results from the parametric analysis (in `param.mpt`) are as follows:

```
location    LOD     alpha    HLOD    marker
  0.000  -0.0113   1.0000   0.0000   Marker1
  0.567  -0.0014   1.0000   0.0000   -
  1.133   0.0083   0.0000   0.0083   -
  1.700   0.0177   0.0000   0.0177   Marker2
```

The per-family parametric results for the example run (in `fparam.mpt`) are:

```
Family   location    LOD    marker
f1         0.000  -0.0113   Marker1
f1         0.567  -0.0014   -
f1         1.133   0.0083   -
f1         1.700   0.0177   Marker2
```

In these files, the location is measured in centi-Morgans (cM), '`alpha`' is the heterogeneity parameter, '`HLOD`' is the heterogeneity LOD score, and the marker columns contain a `-` at locations between markers. Otherwise, the column-headings are self-evident.

The four files with the combined results from the allele sharing model analysis all have the same format. For example, `exppairs.mpt` is:

```
location     LOD     dhat      NPL      Zlr     info    marker
  0.000    0.0004  -0.0457  -0.0434  -0.0448   0.9604   Marker1
  0.567    0.0001   0.0256   0.0215   0.0234   0.8364   -
  1.133    0.0022   0.1160   0.0863   0.0997   0.7409   -
  1.700    0.0075   0.2337   0.1512   0.1862   0.6432   Marker2
```

As before, the location is in cM and the last column contains the marker name or `-` between markers. The other columns all contain quantities defined in the section 5.1 of the Allegro report. The '`LOD`' column is the allele sharing LOD score, '`dhat`' is $\hat{\delta}$, the maximum likelihood estimator of δ in the allele sharing model, '`NPL`' is the NPL score ('non parametric linkage' score), '`Zlr`' is $Z_{\text{lr}} = \text{sign}(\hat{\delta})\sqrt{(2 \ln 10 \cdot \text{LOD})}$, and info is the measure of information. Note that as the LOD is asymptotically χ^2_1, Z_{lr} is asymptotically normal with mean zero and variance one. A negative dhat means that there is less sharing observed than would be expected under the null.

The four files with the allele sharing results broken down by family contribution also have the same format. In our example `fexppairs.mpt` is:

```
Family   location     LOD      NPL      Zlr     info    marker
f1         0.000   -0.0004  -0.0434  -0.0448   0.9405   Marker1
f1         0.567    0.0001   0.0215   0.0234   0.8430   -
f1         1.133    0.0022   0.0863   0.0997   0.7553   -
f1         1.700    0.0075   0.1512   0.1862   0.6771   Marker2
```

The first column contains the family name and the columns headed '`location`' and '`marker`' are as before and the '`NPL`' contains the $\bar{Z}_i$ of section 5.1 in the Allegro report. The '`LOD`' column has the family contribution to the combined LOD, '`Zlr`' is obtained from the LOD as in the combined file and '`info`' is the per-family information measure described in the Allegro report.

Note that the LOD score in the family results is multiplied with the sign of dhat in the overall results. This multiplication has the effect, that if the LOD in the family result file is positive, then the family contributed evidence for linkage, and if if the LOD is negative, then the family contributed evidence against linkage. The combined LOD score is then the sum of the absolute values of the per-family LOD scores.

As said before the haplotyping results are in four files, `haplo.out`, `ihaplo.out`, `founder.out` and `inher.out`. The file `inher.out` can be useful for identifying locations of crossovers, in particular double crossovers. For comparison of haplotypes between families the most valuable files are `haplo.out` and `ihaplo.out`. The latter of these is:

```
                                          1 2
                                          r r
                                          e e
                                          k k
                                          r r
                                          a a
                                          M M

f1      mother   0        0        2 0  1 2
f1      mother   0        0        2 0  0 0
f1      father   0        0        1 0  1 1
f1      father   0        0        1 0  2 1
f1      son      father   mother   1 2  1 1
f1      son      father   mother   1 2  1 2
f1      daughter father   mother   2 2  2 1
f1      daughter father   mother   2 2  1 2
```

The files `xover.out` and `fxover.out` contain the combined and familial results of the observed crossover calculations. The combined file is:

```
location distance xoverrate nxover  nmei marker
  0.000    1.700    3.961    0.16    4   Marker1
Total:     1.700    3.961    0.16    4   Marker2
```

The column '`distance`' is the distance between the marker on the corresponding line and the marker on the next line cM (the 'input map' of the Allegro report). The next column, '`xoverrate`' is the calculated 'observed crossover rate' between these markers times 100 (to make it commensurable with the distance), nxover is the 'observed' number of crossovers and nmei is the number of observable meioses.

The file `uninformative.out` contains all uninformative marker-family pairs. In our example it is empty.

2 Reference

This part of the manual contains an alphabetic list of keywords that may be placed in the Allegro option file. The list is not quite complete, there are a few options that are not documented here and the interested user is referred to the file 'options.h', which is distributed as part the Allegro source code. The format of the Allegro input files is described under the '`PREFILE`' and '`DATFILE`' options, and a description of output is given with the description of the keywords that cause output files to be written (or by referring to the example in part 1). For a description of the format of the option file, and instructions on how to run Allegro, see the beginning of part 1 of this manual. In the following, optional arguments are within square brackets, and italic typeface is used to show generic arguments, which should be replaced by an actual value. The notation '`on/off`' means that one of the words '`on`' or '`off`' should be used.

`CROSSOVERRATE [`*combined-file* `[`*per-family-file*`]]`

Allegro calculates an 'observed map' or 'observed crossover rate' as described in section 5.3 of the Allegro report. The output is written to two files, one with combined results, and one with per-family results; the default file names are `xover.out` and `fxover.out`. This option may be useful for finding errors in the data, for example genotype errors or marker order problems. An example of the combined output may be found in section 1.7 along with a description of its contents. The per-family output file has an additional family id column, but is otherwise very similar to the combined file.

`DATFILE` *file-name*

This option causes the information in the Linkage style datfile with the specified name to be read in and used in subsequent analyses or simulation.

The format of Linkage style datfiles is a little cryptic and much of the information there is ignored by Allegro. Recall the example datfile of section 1.2:

```
3 0 0 5
0 0.0 0.0 0
2 3
1 2
0.9 0.1
1
0.09 0.90 0.90
3 5 # Marker1
0.10 0.20 0.30 0.20 0.10
3 3 # Marker2
0.10 0.30 0.60
0 0
1.7
1 0.10000 0.450000
```

The first number in the first line is the number of loci of interest. One of these loci should always be a disease locus, so this number should be one greater than the number of marker loci. The third number in the first line specifies whether a disease locus is sex-linked or not. 0 means not sex-linked, 1 means sex-linked. The third line specifies the order of the markers on the chromosome being analyzed.

The fourth line should always read 1 2, because Allegro will assume the disease locus has only two alleles. The fifth line specifies the frequencies of the disease locus alleles, first the wild type frequency and then the mutation frequency. The sixth line specifies the number of liability classes. The seventh line lists the penetrances, the first penetrance is given no mutated allele, the second given one, and the third given two. If a disease is sex-linked then a separate line with the two male penetrances must be given before a line with the three female penetrances. If more than one liability class is specified then a separate line with three penetrances must be given for each liability class. See also the '`PREFILE`' option for a description of how to put liability classes into prefiles. It is possible to override the penetrances given here in the MODEL option; see the description of this below.

Next comes a list of marker loci. Each marker takes up two lines. The first line should begin with a 3, followed by the number of alleles at the marker, followed by a '#' and finally the name of the marker. The second line lists the allele frequencies for the marker.

The second to last line contains the distances between the markers as they are listed in the third line. The '`UNIT`' option can be used to set the unit of the distances. The last line is ignored by Allegro.

`ENTROPY on/off`

Turning this option on (off is the default) causes the entropy information measure to be calculated and written to the combined output file from each allele sharing model (see the MODEL option). This option exists for compatibility with Genehunter. The entropy information is termed I_E in the Allegro report.

HAPLOTYPE [*haplotype-file* [*imputed-haplotype-file* [*founder-file* [*inheritance-vector-file*]]]]

Causes Allegro to reconstruct haplotypes. As explained in section 5.2 in the Allegro report, the haplotype reconstruction is found by first finding the most likely inheritance vector path (or one of them, if there are more than one that are equally likely), using the Viterby algorithm. A specification of the inheritance vector at each locus, constitutes an inheritance vector path. All the founder alleles are then numbered (from 1 to $2f$, with f the number of founders), and this numbering is propagated down the pedigree, in accordance with the determined inheritance vector path. An individual which is genotyped at a marker will now fix two founder alleles at the marker (possibly in unspecified order). The set of all individuals which are genotyped at the marker will thus fix a subset of the founder alleles to an allele assignment, or to one of several possible allele assignments (because genotype orders are unspecified). Note that in absence of genotype errors, there is at least one possble assignment. Using the population allele frequencies the most probable of the possible allele assignments is now chosen, and the corresponding alleles are propagated down the pedigree. This process will therefore order the genotypes of all the people who are genotyped at the marker, and impute ordered genotypes for some of the people who are not genotyped.

The created haplotype file contains the ordered genotypes of the genotyped people and the imputed haplotype file contains the imputed ordered genotypes. The inheritance vector file contains the determined most likely inheritance vector path and the founder file contains the corresponding propagated numbering of the founder alleles.

The default file names are `haplo.out`, `ihaplo.out`, `founder.out` and `inher.out`. Sections 1.1 and 1.7, contain a simple example of haplotyping and some of the produced output; see also the discussion in section 1.4.

MAXMEMORY *n*

Sets the maximum amount of CPU memory that Allegro is allowed to allocate to *n* Megabytes. Allegro estimates the memory it requires, and if that estimate exceeds *n* it tries to save memory by recalculating or writing to disk instead of storing in memory. If this option is missing then 2048 is assumed. This option is useful for preventing Allegro from eating up all the computer's memory. See also `SWAPDIRNAME`. `SWAPDIRNAME`.

MAXSTEPLENGTH *x*

Causes all statistics and distributions (LOD scores there among) to be calculated at least every *x* cM. For example, if *x* is 0.6 and there are markers at 1 cM and 2.5 cM, then calculations will be carried out at 1, 1.5, 2 and 2.5 cM. See also `STEPS` and `STEPFILE`.

MAXUNLOOP *n*

This parameter sets the level of loop unrolling in the fast Fourier transforms of the HMM step. Its effect is described in some detail in section 3.1 in the Allegro report, which reports on run time experimentation indicating that the optimal value for *n* is 2 on a Pentium computer, 3 on a Sun computer, and 4 on a Dec Alpha computer. The speed of Allegro is not very sensitive to the value of this parameter, but the user may nevertheless wish to perform his own timing experimentation to determine the optimal value on the platform he is using.

MODEL `mpt/spt lin/exp` *scoring-function weighting-scheme* [*combined-file* [*per-family-file*]]

This option causes an allele sharing model analysis to be carried out using the specified scoring function and family weighting scheme. If '`mpt`' is specified then full multipoint analysis is performed, but with '`spt`' single point analysis is carried out (see section 1.3). With '`lin`' the linear model of Kong and Cox (1997) is used and '`exp`' causes their exponential model to be used. There are six supported scoring functions, specified as '`pairs`', '`all`', '`homoz`', '`mnallele`', '`robdom`' and '`ps:mm/mf/ff`' corresponding to S_{pairs}, S_{all}, S_{homoz}, S_{mnallele} and S_{robdom} respectively. These different scoring functions are

discussed in section 1.3 and in section 5.1 in the Allegro reort. The Allegro report recommends the exponential model, and quotes McPeek (1999) in recommending 'S_{pairs} as a compromise choice for a scoring function that performs well over all disease models'. The '`ps:mm/mf/ff`' allows the user to give meiosis specificy weights as described in Karason et al. (2002): $S_{\text{ps}} = mm * t_{mm} + mf * t_{mf} + ff * t_{ff}$, where t_{mm} is 1 if a pair shares IBD, both getting the same allele through their mothers, and 0 otherwise. t_{mf} is the same but one getting the allele through its father, and t_{ff} is 1 if both get the same allele through their fathers.

Allegro offers three different ways of specifying the family weighting scheme. If '`equal`' is specified then all families are weighted equally. If '`power:`*n*' is specified then families are weigthed according to the standard deviation of the score function under the null hypothesis of no linkage to the power *n*. The third possibility is to give the name of a file that the weights of the families will be read from. This should be a two column file, the first column containing names of families to be analyzed and the second column containing the weights of the families. The '`power:`' weighting scheme is discussed at the end of section 5.1 in the Allegro report, which (a little cautiously) recommends *n* = 0.5.

The default names of the output files are constructed from the arguments specified, for example `exppairs.1.mpt` (constructed from '`MODEL mpt exp pairs power:1`') and `linrobdom.spt` (from '`MODEL spt lin robdom equal`'). These are the names of files with combined results, per-family files have the same names prefixed with an `f`. Section 1.7 gives examples of output from allele sharing analyses and describes the meaning of individual columns. See also the options `NPLEXACTP` and `ENTROPY` which cause columns to be added to the files.

Note that any number of MODEL lines may be placed in the option file, thus enabling several different analyses to be carried out in the same Allegro run.

`MODEL mpt/spt par [freq:`*freq* `pen:`*p0/p1/p2*`] [het] [`*combined-file* `[`*per-family-file*`]]`

Causes a classical parametric analysis to be performed for an autosome. The parameter *freq* is the disease allele frequency and *p0*, *p1* and *p2* are the penetrances with no, one and two disease alleles respectively. These arguments cannot be given if liability classes are being used. If '`freq`' and '`pen`' are omitted the values in the datfile are used. If '`het`' is specified then a heterogeneity parameter is estimated.

The default names of the parametric output files are `param.mpt` and `fparam.mpt` for multipoint analysis, and `param.spt` and `fparam.spt` for single point analysis. As with allele sharing models, it is possible to carry out several different parametric analyses in a single Allegro run by placing several different MODEL lines in the option file (with different frequencies and/or penetrances), but then it is necessary to explicitly name the output files for each model differently. Otherwise the output from later mentioned models will overwrite output from previously mentioned models.

Section 1.7 gives an example of output from parametric analysis and describes the meaning of individual columns.

`MODEL mpt/spt par X [freq:`*frq* `pen:`*p0/p1/p2/pm0/pm1*`] [het] [`*combined-file* `[`*fam-file*`]]`

Perform parametric analysis on a sex-linked chromosome. The penetrances *p0*, *p1* and *p2* are for females, and *pm0* and *pm1* for males. Otherwise the description above of autosomal parametric analysis applies.

`NPLEXACTP on/off`

Turn on `NPLEXACTP` to have Allegro calculate Genehunter's 'exact p-values' associated with the NPL scores and write them to the output files of each allele sharing MODEL. Under complete information these p-values are exact, otherwise they are an approximation. The p-values are placed in a column headed '`nplexactp`', between the '`Zlr`' and '`info`' columns. The default is '`off`'.

`PAIRWISEIBD mpt/spt` *whichpairs*

Calculate pairwise IBD sharing probabilities, either single-point or multipoint. The possible values for *whichpairs* are `all`, `genotyped`, `affected` and `informative`. Two files are created for each

`PAIRWISEIBD` line, one gives the prior probabilities of each pair's sharing, and the other gives the corresponding posterior probabilities (see section 1.5). The default names of the files are `prior.mpt`, `prior.spt`, `posterior.mpt` and `posterior.spt`. Each line in the prior file contains the family, the individuals in a pair and the probability that they share 0, 1 and 2 alleles (e.g. `f1 p1 p2 0.25 0.50 0.25`), and the posterior file contains in addition the location and the marker name (e.g. `f1 6.670 p1 p2 0.25 0.50 0.25 D3S450`).

`PREFILE` *file-name*

Pedigree and genotype data will be read from a Linkage style prefile. Following is the example of a simple prefile from section 1.2:

```
f1  father   0       0      1 0  0 0  0 0
f1  mother   0       0      2 0  0 0  0 0
f1  daughter father  mother 2 2  1 2  1 2
f1  son      father  mother 1 2  1 1  1 2
```

Usually Linkage style prefiles have the following columns:

1. Family name
2. Individual identifier
3. Father's name, 0 if unknown
4. Mother's name, 0 if unknown
5. Sex, 1 for male and 2 for female
6. Affection status, 1 for unaffected, 2 for affected and 0 for unknown affection status
7. First allele of first marker, 0 if unknown
8. Second allele of first marker, 0 if unknown

The alleles of the second marker are in columns 9 and 10, and so on for as many markers as are specified in the accompanying datfile. If the X-chromosome is being analyzed then males are homozygous, but genotypes for every marker should still occupy two columns.

If liability classes are used then an extra column is added after column 6 (the affection status column). This column specifies the liability class of each person and should be between 1 and the number of liability classes specified in the datfile. Note that every person must have a liability class specified, even individuals with unknown affection status, and that 0's are not allowed. If more than one family should be analysed, description of all the families are placed in the same prefile, one after the other, with no intervening separation (no blank lines). The order of individuals within families is immaterial.

`SEXSPECIFIC on/off`

Different marker maps are used for males and females. An extra line of marker distances should be given in the DATFILE. The first line is the marker distances in males and the second is the marker distances in females.

`SIMULATE [dloc:`*dloc*`] [npre:`*npre*`] [rep:`*rep*`] [err:`*err*`] [yield:`*yield*`] [het:`*het*`]`

Instructs Allegro to simulate multipoint data given a single locus disease model, pedigree structures and phenotypes (who should be affected). Information on the disease model and the locations and allele frequencies of markers is given by the datfile. The pedigree structures, and information on who should be affected and who should be genotyped is given by the prefile. Only the first 8 columns of the prefile are used. The first six columns are the same as usually, and the seventh and eighth column are genotype flag columns, they should contain two nonzeros for those individuals who will have genotype information simulated, and two 0's for those who will not. Note that simulation cannot be performed in the same run as any type of analysis (i.e. with any MODEL lines in the option file). The simulation is discussed shortly in section 5.2 of the Allegro report.

The results of the simulation are reported in prefiles with family structure and phenotypes in the first 6 columns and the generated genotypes in columns after column 6, their number being determined by the

number of markers in the datfile. The generated prefiles will be named after the input prefile, but suffixed with numbers, as indicated by the example below.

The parameter *dloc* gives the location of the disease locus, in cM, with respect to the first marker to be simulated. If '`dloc:`' is omitted then multipoint data will be simulated based on an unlinked disease locus.

The *npre* parameter specifies the number of prefiles to be generated, and *rep* specifies how many times the family pattern given in the input prefile will be repeated in each generated prefile. The default value for both of these parameters is one. For example, if *npre* = 3, *rep* = 10, and the input prefile contains one sib pair and one cousin pair, then three prefiles will be generated, each containing 10 sib pairs and 10 cousin pairs.

The parameter *err* gives the error rate, i.e. the probability that a genotype in the generated prefiles is not the one that was simulated, but instead a random genotype (distributed as specified in the datfile). The *yield* parameter gives the yield of genotyping, i.e. the probability that a genotype that has been simulated for a person will be put into the generated prefile, and *het* gives a simple way of simulating heterogeneity. Each family has a probability *het* of having its data simulated from the null instead of from the disease model given in the datfile. The default error rate is 0, the default yield is 1, and the default value of *het* is 0.

An example option line to instruct Allegro to do simulation is:

```
SIMULATE dloc:50.53 npre:100 rep:50 err:0.01 yield:0.9 het:0.7
```

This will simulate genotypes of families with the same family structure as in the input prefile, at the markers given in the input datfile, given the phenotypes (affection status) of the prefile and the parametric disease model of the datfile. The seventh column in the prefile determines the individuals receiving simulated genotypes. The disease locus is set at 50.3 cM from the first marker, each simulated genotype has a 1% probability of being wrong, the probability of a simulated genotype being put in to the prefile is 90% (otherwise a 0 is put in its place), and each family has a 70% chance of being completely unconnected to the disease locus. The simulation is repeated 100 x 50 times, and the results are reported in 100 output prefiles, each containing the family structure of the input prefile repeated 50 times, together with the simulated genotypes. If the input prefile is named `xx.pre`, the created prefiles are named `xx.pre.001`, `xx.pre.002`, ..., `xx.pre.100`.

`STEPFILE` *file-name*

All statistics and distributions are calculated at the locations specified in step-file. The step-file contains a list of locations (in cM). See also `STEPS`.

`STEPS` *n*

All statistics and distributions (among them LOD scores) are calculated at $n - 1$ positions between consecutive pairs of markers. For example if *n* is 2 and there are markers at 1 and 2.5 cM then calculations will be performed at 1, 1.75 and 2.5 cM. Only one of the keywords `STEPS`, `STEPFILE` and `MAXSTEPLENGTH` can be specified. If none is specified then '`STEPS 1`' is assumed.

`SWAPDIRNAME` *directory-name*

Name a directory for the intermediate files created to save on CPU memory due to the limit set by '`MAXMEMORY`'.

`UNINFORMATIVE` [*file-name*]

Causes all marker-family pairs that are uninformative to be written to a file, by default named `uninformative.out`. For example, if marker Marker1 is uninformative for family f1, then the line '`Marker1 f1`' will be written to the file.

`UNIT centimorgan/recombination`

Sets the unit of the marker distances in the datfile. If this option is not specified and all the distances are less than 0.5 then the unit is assumed to be recombination fraction, otherwise cM.

References

Gudbjartsson DF, Jonasson K, Frigge M, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nature Genetics 25:12–13

Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsdottir A (2005) Allegro version 2. Nature Genetics. 37: 1015–1016

Karason A, Gudjonsson JE, Upmanyu R, Antonsdottir AA, Hauksson VB, Runasdottir EH, Jonsson HH, Gudbjartsson DF, Frigge ML, Kong A, Stefansson K, Valdimarsson H, Gulcher JR (2002) A susceptibility gene for psoriatic arthritis maps to chromosome 16q: evidence for imprinting. Am J Hum Genet 72(1):125–31.

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Kruglyak L, Lander ES (1998) Faster multipoint linkage analysis using Fourier transforms. J Comput Biol 5:1–7

McPeek MS (1999) Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet Epidemiol 16:225-249

Sobel E, Lange K, O'Connel JR, Weeks DE (1996) Haplotyping algorithms. In: Speed T, Waterman MS (eds) Genetic mapping and DNA sequencing. The IMA volumes in mathemtics and its applications 81. Springer-Verlag, New York

Terwilliger JD, Ott J (1994) Analysis of human genetic linkage, Johns Hopkins University Press, Baltimore and London

Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118–127