PEDIGREE ANALYSIS PACKAGE

Sandra J. Hasstedt

Version 5.0

May, 2002

Department of Human Genetics
University of Utah
Salt Lake City

I. INTRODUCTION

The Pedigree Analysis Package (PAP) comprises a set of FORTRAN 90 programs for computing likelihoods [Elston & Stewart 1971] or simulating phenotypes on pedigree members using genetic models. PAP performs one of eight options selected at runtime: **(1)** compute the likelihood, lod score, or chi-square statistic of specified parameter values, **(2)** compute the probability of each genotype for pedigree members, **(3)** simulate phenotypes for output into files, **(4)** perform a search for the maximum likelihood, lod score, or chi-square statistic over specified parameters, **(5)** maximize the likelihood, lod score, or chi-square statistic over specified parameters (with or without standard errors), **(6)** compute the standard errors of parameters for known estimates, **(7)** simulate phenotypes and estimate parameter values, **(8)** compute a grid of likelihoods, lod scores, or chi-square statistics over one or two parameters. Likelihood maximization uses either GEMINI [Lalouel 1979], for which the source is provided, or NPSOL [Gill et al 1986], for which the source must be purchased.

PAP may be used for segregation analysis, variance components analysis, linkage analysis, measured genotype analysis, transmission disequilibrium testing, or genetic model fitting. The genetic model may contain any number of loci and alleles. Model flexibility follows from the user selecting frequency, transmission, discrete major locus, quantitative major locus, and within genotype subroutines from a library of each, thereby specifying the assumptions and parameterization for an analysis or simulation. Similar data flexibility in PAP allows the pedigrees to be any size or structure and contain multiple ancestral branches and inbreeding or exchange loops; traits may be discrete or quantitative. Phenotypes may be simulated assuming any model available in PAP. Marker genotypes may be simulated for any number of linkage markers.

Using PAP incurs as its only obligation the requirement to appropriately reference each use. This manual should always be referenced. In addition, published approximations or models should be referenced upon using the corresponding subroutine. Check the description of the selected subroutine in Appendix C to identify any reference.

This manual contains practical advice on performing likelihood analysis in addition to technical details of the usage of PAP. The chapters progress from basic to more complex information. All users should read chapters II through V; Chapters VI and VII contain information of less general interest; the Appendix contains reference information not meant to be read sequentially.

Instructions for obtaining the PAP source code can be obtained by sending an e-mail message to **sandy@genetics.utah.edu**. Any questions or problems should be sent to the same address. Information about obtaining NPSOL is available at **www.sbsi-sol-optimize.com**.

II. PREPARING THE INPUT FILES

Three required and three optional files provide the sample and population information to PAP. The required files include *trip.dat*, which defines the pedigree structure, *header.dat*, which describes the variables, and *phen.dat*, which contains the phenotypes. The optional files include *order.dat*, which specifies the computational order, *ascer.dat*, which designates the individuals for ascertainment correction, and *popln.dat*, which provides population information on genetic markers and diseases. Program *preped* (section III.1) combines *trip.dat, order.dat, header.dat, phen.dat,* and *ascer.dat* into *papin.dat* (section B.6) for use by *descstat* (section III.2), *prepap* (section III.3), *simrk* (section III.4), *simul* (section III.5), and *papdr* (section III.6). The following sections describe the contents of *trip.dat, order.dat, header.dat, phen.dat, ascer.dat*, and *popln.dat*; sections B.1-5 detail their formats.

II.1. Pedigree Structure: *trip.dat*

File *trip.dat* defines the pedigree structure through identification of the father and mother of each sample member. The file contains four columns: pedigree number, father's ID number, mother's ID number, and offspring's ID number. Only the offspring's ID number can equal zero (to indicate a childless couple). Pedigree members without parents in the sample enter only as parents. Unstudied individuals must be included when needed to connect studied members of the pedigree.

Section B.1 details the format of *trip.dat*. The file must be sorted before use.

II.2. Computation Order: *order.dat*

In the absence of optional file *order.dat*, program *preped* (section III.1) determines the order in which the nuclear families within the pedigree enter into the likelihood computation. In pedigrees without loops, the order affects computation speed little. However, when a pedigree contains loops, the order can greatly affect computational speed. Supplying file *order.dat* overrides the automatic order determination in *preped*, allowing the user to define an order for more rapid computation.

As each nuclear family enters the likelihood computation, probabilities must be stored on any member of this or a previous nuclear family who occurs again in a later nuclear family. These individuals form the cutset. Larger cutsets require more computation time. A pedigree without loops need never have a cutset larger than 1. A pedigree with loops has at least one cutset of size 2 or larger. Program *preped* prints out the maximum cutset size for each pedigree. You can use *order.dat* to order the families to minimize the maximum cutset size, thereby speeding the computation.

File *order.dat* contains three columns: pedigree number, husband's ID number, and wife's ID number. The file must be ordered by pedigree number, but need not include all the pedigrees in file *trip.dat* (section II.1).

Section B.2 details the format of *order.dat*.

II.3. Marker/Trait Descriptions: *header.dat*

File *header.dat* describes the variables included in *phen.dat* (section II.4). The required information includes names and types for each variable, locations in *popln.dat* (section II.5) for discrete traits and markers, and the number of alleles for markers. Optional records in *header.dat* specify a missing value code (default -9999), a phenotype simulation code (default -99999), a value to subtract (default 0), a value to divide by (default 1), and power for the transformation (default 1) for each variable. The phenotype simulation code allows you to specify which phenotypes to simulate. By specifying a value to divide by, you can scale a quantitative phenotype to put all the parameters in the same range, thereby improving the maximization performance (see section V.4). By specifying to subtract the mean and divide by the standard deviation, you can standardize the variable. By specifying a power for transformation, you can transform quantitative phenotypes using the function $y = r/P[(x/r+1)^P-1]$ [Maclean et al 1976], where x represents the phenotype, P represents the power, and r = 6.

Section B.3 details the format of *header.dat*.

II.4. Phenotype Data: *phen.dat, ascer.dat*

File *phen.dat* contains an entry for each studied pedigree member recording phenotypes of all the traits and markers collected. You do not need to include an entry in *phen.dat* for unstudied individuals included in *trip.dat* (section II.1) in order to connect studied individuals.

Optional file *ascer.dat* contains an entry for each proband or potential proband. For ascertainment correction by the method of Thompson & Cannings [1980], *ascer.dat* contains phenotypes of all traits and markers measured on probands before deciding to study each pedigree. For example, if an early heart attack initiated study of a pedigree, but blood pressure was subsequently measured, the proband's entry in *ascer.dat* records a heart attack in the appropriate column, but contains a missing value in the blood pressure column; in contrast, the proband's record in *phen.dat* includes both heart attack status and blood pressure level. For ascertainment correction by the ascertainment-assumption-free method [Ewens & Shute 1986, Shute & Ewens 1988a], *ascer.dat* contains phenotypes of potential probands. See section VI.4 for more information on ascertainment correction.

For each type of variable entered in *phen.dat* or *ascer.dat*, the form of the phenotype follows:

(1) For gender, the phenotype equals 1 for male and 2 for female. In *header.dat* (section II.3), IVARTP equals 1 and IVARDF equals 0.

(2) For a disease, the phenotype equals 1 for normal and 2 for affected. In *header.dat* (sections II.3), IVARTP equals 1 and IVARDF equals either 0 or the location of prevalence or incidence figures in *popln.dat* (section II.5).

(3) For a category designation, such as an environmental dichotomy, phenotypes 1 and 2 may be defined by the user. In *header.dat* (section II.3), IVARTP equals 1 and IVARDF equals 0.

(4) For disease severity, the phenotype may be coded with any integers. In *header.dat* (section II.3), IVARTP equals 1 and IVARDF equals the location of prevalence figures for each level of disease severity in *popln.dat* (section II.5).

(5) For a quantitative trait, the phenotype equals a number assumed to have a decimal on the right if not included. In *header.dat* (section II.3), IVARTP equals 2 and IVARDF equals 0.

(6) For a shared environmental effect, the phenotype equals an arbitrary but small, positive integer. You assign all members sharing a defined effect (for example, the same household) the same number. You assign the missing value code to anyone who does not share the effect with anyone else in the sample. The same number can and should be reused in a different pedigree. In *header.dat* (section II.3), IVARTP equals 2 and IVARDF equals 0.

(7) For a codominant marker (also see **(8)**), the phenotype equals the two alleles in two contiguous columns. For an X-linked marker in males, the second column contains 0. In file *header.dat* (section II.3), NCOL will be larger than NDATA, IVARTP equals 5 for an autosomal marker or 6 for an X-linked marker, and IVARDF equals the location of the marker allele frequencies in *popln.dat* (section II.5) or (-) the number of marker alleles.

(8) For a codominant marker (also see **(7)**), the phenotype equals $I (I - 1)/2 + J$ for a genotype at an autosomal locus or a female genotype at an X-linked locus and equals $N (N + 1)/2 + I$ for a male genotype at an X-linked locus, where I and J represent the alleles, $J \leq I$, and N equals the number of alleles. In *header.dat* (section II.3), IVARTP equals 3 for an autosomal marker or 4 for an X-linked marker and IVARDF equals the location of the marker allele frequencies in *popln.dat* (section II.5) or (-) the number of marker alleles.

(9) For a non-codominant marker, the phenotype equals the number assigned in *popln.dat* (section II.5). In *header.dat* (section II.3), IVARTP equals 3 for an autosomal marker or 4 for an X-linked marker and IVARDF equals the location of information about the marker in *popln.dat* (section II.5).

(10) To indicate an unknown phenotype, the phenotype equals -9999 or the missing value code specified for the variable in *header.dat* (section II.3).

(11) To indicate that the phenotype should be simulated, the phenotype equals -99999 or the simulated phenotype code specified for the variable in *header.dat* (section II.3). Except when phenotypes are simulated (programs *simrk* and *simul* and options 3 and 6 of program *papdr*), this code is treated as a missing value. The simulated phenotype code needs to be used only when phenotypes for a trait or marker are to be retained for some individuals, simulated for others, and considered missing values for yet others. Otherwise, you can choose to simulate all phenotypes, phenotypes without missing values, phenotypes with missing values, phenotypes without missing values, or phenotypes without missing values for another trait or marker. See section VI.11 for more information about simulation.

Section B.4 details the format of *phen.dat* and *ascer.dat*.

II.5. Population Information: *popln.dat*

File *popln.dat* contains population information about genetic markers and diseases. The information about a genetic marker can include allele frequencies only or allele frequencies and, for non-codominant markers, genotype/phenotype relationships. The information about a disease can include incidence or prevalence figures, specific for age groups or disease severity.

Section B.5 details the format of *popln.dat*.

III. USING THE PROGRAMS

PAP comprises seven programs with test data that double as examples; Appendix A lists all the distributed files. Sections III.1 through III.7 list, for each of the seven programs, the constituent subroutines and the input and output files. If a copy of any output file exists, the program overwrites it.

Each program can be linked by using the command "make" followed by the name of the program. For example, the executable for *preped* is produced by the command "make preped". The executable for *papdr* using GEMINI and incorporating subroutines *papfqhw, paptcms, dmlpr0, qmlpr0, papen, papcr* is produced by the command "make papdrg a=hw b=ms c=0 d=0". For more information, use the command "make".

III.1. Data Preparation Program: *preped*

Source:	*preped, phenin, popin, pselec, pedgen, pedot, filerr, daterr, inmin, yesno*
Input:	console, *trip.dat* (section II.1), *order.dat* (optional, section II.2), *header.dat* (section II.3), *phen.dat* (section II.4), *ascer.dat* (optional, section II.4)
Output:	console, *papin.dat* (section B.6)

Program *preped* combines the phenotype data and pedigree structure information into a single file for input to *descstat* (section III.2), *prepap* (section III.3), *simrk* (section III.4), *simul* (section III.5), and *papdr* (section III.6). Program *preped* must be executed whenever any of the input files change.

Program *preped* asks six questions: **(1)** Ascertainment correction? Respond with "y" if you intend to make an ascertainment correction and have prepared file *ascer.dat* (section II.4). Respond with "n" otherwise. **(2)** Assortative mating? Respond with "y" if you intend to assume assortative mating in your genetic model. Respond with "n" otherwise. **(3)** Minimum pedigree size? Respond with "1" unless you would like to eliminate any pedigrees in your sample in which fewer than some number were measured. **(4)** Eliminate parents without offspring? Respond with "y" to eliminate measured parents without measured children. Respond with "n" otherwise. **(5)** Add gender for parents with missing values? Respond with "y" if you plan to use an X-linked or gender-specific model; a request to input the variable representing gender follows. Respond with "n" otherwise. **(6)** Add category designation if unstudied? Respond with "y" if you plan to use a category-specific model; a request to input the variable representing the category follows. Respond with "n" otherwise. When using an X-linked or category-specific model, even unstudied individuals must have gender or category specified; positive responses to questions **(5)** and **(6)** insert them for unstudied individuals.

If *header.dat* so specifies, *preped* will subtract a specified value from each phenotype, divide each phenotype by a specified value, and/or transform each phenotype using the function y

$= r/P[(x/r+1)^P-1]$ [Maclean et al 1976], where x represents the phenotype, P represents the power and r = 6.

Program *preped* includes in *papin.dat* only those unstudied individuals needed to connect studied individuals. Therefore, to limit your analysis to adults (for example) you need only delete the records for youths from *phen.dat*; *trip.dat* need not be altered. File *papin.dat* does not include any unnecessary individuals, which would slow the computation. Upon detecting an error in the pedigree structure or determining that the storage is insufficient, *preped* outputs an error message to the console and terminates execution.

III.2. Descriptive Statistics Program: *descstat*

Source: *descstat, pedin, pselec, filerr, daterr, moderr, rnmin, inmin, yesno*
Input: console, *papin.dat* (produced by *preped*, section III.1)
Output: console

Program *descstat* outputs the number, mean, standard deviation, skew, kurtosis, minimum, and maximum of any variable in *papin.dat*. Ranges on any of the variables may be specified. Program *descstat* allows you to check that your sample size is correct and check that scaling, standardization, or transformation have had the desired effect. Upon detecting an error or determining that the storage is insufficient, *descstat* outputs an error message to the console and terminates execution.

III.3. Analysis Preparation Program: *prepap*

Source: *prepap, modin, popin, papgn, sbrlib, modnew, pselec, parin, parnew, genot, modot, filerr, daterr, moderr, rnmin, inmin, yesno*
Input: console, *papin.dat* (produced by *preped*, section III.1), *popln.dat* (optional, section II.5), *freq.dat* (provided), *tran.dat* (provided), *dmlp.dat* (provided), *qmlp.dat* (provided), *wgen.dat* (provided), *model.dat* (optional)
Output: console, *model.dat* (section B.7)

Program *prepap* writes *model.dat* for use by *simrk* (section III.4), *simul* (section III.5), and *papdr* (section III.6). Program *prepap* determines the genetic model and the associations between loci and variables either interactively or by reading *model.dat*; then *prepap* interactively requests parameter values and maximization or grid designations and writes *model.dat*. Files *freq.dat, tran.dat, dmlp.dat, qmlp.dat,* and *wgen.dat* inform *prepap* about the available frequency, transmission, discrete major locus, quantitative major locus, and within genotype subroutines; *papin.dat* and *popln.dat* inform *prepap* about the data. File *model.dat* optionally informs *prepap* about the genetic model if the user is changing only the parameter values. Please note that the input *model.dat* will be overwritten by *prepap*. If insufficient storage has been allocated, *prepap* outputs an error message to the console and terminates execution.

Six default models may be selected in place of user input of the model information. These models include: **(1)** univariate segregation analysis (discrete or quantitative trait), **(2)** bivariate segregation analysis (two discrete traits, two quantitative traits or one discrete and one quantitative trait), **(3)** measured genotype analysis (one marker and either one discrete or one quantitative trait), **(4)** 2-point linkage analysis (two traits/markers), **(5)** expected lod score estimation (one discrete or one quantitative trait), **(6)** multi-point linkage analysis (at least three traits/markers). Options **(1)** through **(3)** use subroutine *dmlprsv* and options **(3)** through **(6)** use subroutine *dmlprpn* for a discrete trait; all options use *qmlprmv* for a quantitative trait (see sections C.3 and C.4).

III.4. Simulation Program: *simrk*

Source:	*simrk, mrkin, popin, pedin, papsmg, drandi, filerr, daterr, moderr, simerr, inmin, yesno*
Input:	console, *papin.dat* (produced by *preped*, section III.1), *model.dat* (optional; produced by *prepap*, section III.3), *popln.dat* (optional; section II.5)
Output:	console, *model.dat, pap.log, header.sim, phen.1, phen.2,..., phen.n* for n replicates

Program *simrk* simulates codominant marker genotypes using the chromosome-based method of Terwilliger et al [1993]. File *papin.dat* specifies the fixed pedigree structure and pattern of missing values; file *model.dat* defines the genetic model, which may be produced either by executing *prepap* or through interactive input to *simrk*; file *popln.dat* optionally specifies marker allele frequencies. The simulated phenotypes are output in the format of *phen.dat* (section B.4) in files *phen.1, phen.2,..., phen.n* for n replicates. The appropriate *header.dat* is also output. File pap.log lists, by replicate, the haplotypes simulated for each pedigree member, missing values included.

Marker genotypes may be simulated for either autosomal or X-linked inheritance. Although *model.dat* may be produced using either option 6 or 7 of *prepap*, interactive input using *simrk* is preferable because of the large required size of *prepap* for this purpose. Distances between markers may be input to *simrk* as either cMs or recombination probabilities, but can be input to *prepap* only as recombination probabilities. Program *simrk* requests the length of the chromosome and location of the first marker and uses the Sturt [1976] map function.

In order to simulate any phenotype other than a codominant marker, you must use program *simul* (section III.5). If you fix rather than simulate any phenotype for any pedigree member, you must instead use program *papdr* (section III.6) option 3.

See section VI.11 for more information about simulation.

III.5. Simulation Program: *simul*

Source: *simul, datin, modin, popin, papgn, sbrlib, pedin, simph, papsf, papsg, papcsg0, papfq, paptc, dmlpr, qmlpr, papwg, domdis, thresh, devia, integ, drandi, filerr, daterr, moderr, simerr, inmin, yesno*

Input: console, *papin.dat* (produced by *preped*, section III.1), *model.dat* (produced by *prepap*, section III.3), *popln.dat* (optional; section II.5), *freq.dat, tran.dat, dmlp.dat, qmlp.dat, wgen.dat* (supplied with source code)

Output: console, *pap.log, header.sim, phen.1, phen.2,..., phen.n* for n replicates

Program *simul* simulates phenotypes for output to data files. File *papin.dat* specifies the fixed pedigree structure; file *model.dat* defines the genetic model and parameter values. The simulated phenotypes are output in the format of *phen.dat* (section B.4) in files *phen.1, phen.2,..., phen.n* for n replicates. The assigned genotypes for each individual and each replicate are output in file *pap.log*. The corresponding data descriptions are output as *header.dat* (section B.3).

When linking *simul* you select from libraries of subroutines named *papfq, paptc, dmlpr, qmlpr*, and *papwg* with 1-3 additional characters in each name. Line 2 of *model.dat* lists the subroutines you selected when executing *prepap*.

If you fix rather than simulate any phenotype for any individual, you must instead use program *papdr* (section III.6) option 3.

See section VI.11 for more information about simulation.

III.6. Analysis Program: *papdr*

Source(G): *papdr, datin, modin, popin, papgn, sbrlib, pedin, simph, papmv, alloff, repeat, papgp, papsf, papsr, papse, papes, papgd, pappr, pappg, papsg, papcsg, paptr, paplk, pappn, pappl, pappd, papnf, papcl, assort, modot, pappo, papro, filerr, daterr, moderr, paperr, domdis, thresh, devia, integ, drandi, papsm, inmin, yesno, papmxg, step, stepfn, update*

Source(N): *papdr, datin, modin, popin, papgn, sbrlib, pedin, simph, papmv, alloff, repeat, papgp, papsf, papsr, papse, papes, papgd, pappr, pappg, papsg, papcsg, paptr, paplk, pappn, pappl, pappd, papnf, papcl, assort, modot, pappo, papro, filerr, daterr, moderr, paperr, domdis, thresh, devia, integ, papsm, inmin, yesno, objfun, confun, blas1* (NPSOL), *blas2* (NPSOL), *chsubs* (NPSOL), *cmsub*s (NPSOL), *f06sub*s (NPSOL), *lssubs* (NPSOL), *mcsubs* (NPSOL), *npsubs* (NPSOL), *opsubs (*NPSOL), *qrsubs* (NPSOL), *srsubs* (NPSOL)

Input: console, *popln.dat* (section II.5), *papin.dat* (produced by *preped*, section III.1), *model.dat* (produced by *prepap*, section III.3), *freq.dat, tran.dat, dmlp.dat, qmlp.dat, wgen.dat* (supplied with source code)

Output: console, *pap.out*, *pap.log* (option 3, 7), *model.dat* (option 4, 5, 6, 8), *header.sim* (option 3), *phen.1*, *phen.2*, ... (option 3)

Program *papdr* performs one of eight execution options (see section IV.1) and outputs the parameter values and results to the console and to *pap.out*. If a simulation option (3 or 7) is selected, output from each replicate is recorded in *pap.log*. If option 3 is selected, the simulated phenotype data are output into *header.dat* and *phen.1, phen.2, ..., phen.n* for n replicates. If options 4, 5, 6, or 8 is selected, the best parameter values are output in *model.dat*. Note that the input *model.dat* will be overwritten.

The source code included upon linking *papdr* differs depending on whether maximization will be performed with GEMINI (designated (G) above) or NPSOL (designated (N) above). The source for GEMINI is included with the PAP source code. The source for NPSOL is available from **www.sbsi-sol-optimize.com**.

When linking *papdr* you select from libraries of subroutines named *papfq, paptc, dmlpr, qmlpr, papwg, papen,* and *papcr* with 1-3 additional characters in each name. Line 2 of *model.dat* lists the subroutines you selected for your application.

III.7. Estimation Program: *gpe*

Source: *gpe, phenin, popin, pselec, rnmin, yesno, inmin, filerr, daterr*
Input: console, *prob.dat* (sorted *pap.out* from *papdr*, option 2), *header.dat* (section II.3), *phen.dat* (section II.4)
Output: console

Program *gpe* estimates genotype frequencies, means, and standard deviations using genotype probability estimators [Hasstedt & Moll 1989]. Estimates may be obtained for specified ranges of other variables contained in *phen.dat*. The estimation procedure uses the genotype probabilities to partially assign an individual to a particular genotype.

File *prob.dat* contains the sorted *pap.out* produced by option 2 of *papdr* (section III.6). File *phen.dat* need not be the file for which *prob.dat* was computed, but also must be sorted. File *header.dat* corresponds to *phen.dat*.

IV. DEFINING THE MODEL

Program *prepap* (section III.3) is used to define the model to be used when computing likelihoods or simulating phenotypes using *simul* (section III.5) or *papdr* (section III.6). PAP can assume a variety of genetic models related to any number of markers and traits. Different parameterizations are available to define the model. This section describes the features that are available in PAP; models can be defined by combining these features.

Section IV.1 describes the execution options for program *papdr*. Sections IV.2-9 describe the parameterizations and section IV.10 describes the genotype assignment options available in *papdr* and also program *simul*. See section D for examples of each feature.

IV.1. Execution Options

Program *papdr* (section III.6) performs eight functions: **(1)** compute the likelihood of a set of parameter values, **(2)** compute genotype probabilities, **(3)** simulate phenotypes according to a genetic model and output them to files, **(4)** search for the maximum likelihood one parameter at a time, **(5)** maximize the likelihood with or without standard error calculation, **(6)** compute standard errors of known parameter estimates, **(7)** simulate phenotypes according to a genetic model and maximize the likelihood to estimate parameter values, **(8)** compute likelihoods of a grid of parameter values. Except for execution option 7, *model.dat* is rewritten each time a better likelihood is computed. If performing a linkage analysis and ascertainment correction has not been specified, *papdr* offers the option of lod score rather than likelihood computation. If performing the transmission disequilibrium test and ascertainment correction has not been specified, *papdr* offers the option of chi-squared statistic rather than likelihood computation.

See section D.1 for examples demonstrating the execution options. See sections IV.1.1 through IV.1.8 for descriptions of the execution options.

IV.1.1. Single likelihood computation

Execution option 1 computes the likelihood of the specified parameter values. File *pap.out* lists by pedigree, as well as for the complete sample, a count of total and measured members and common and natural (multiplied by -2) logarithms of the likelihoods before and after ascertainment correction. If performing a linkage analysis and ascertainment correction has not been specified, lod scores rather than likelihoods may be computed. If performing the transmission disequilibrium test and ascertainment correction has not been specified, chi-squared statistics rather than likelihoods may be computed.

I use execution option 1 to: **(1)** check for the correct sample size in my data file, **(2)** test for genotype inconsistencies in the markers, **(3)** compute lod scores by pedigree, **(4)** compare the

likelihoods of two different models within pedigrees, **(5)** obtain a likelihood for comparison to a known value (hand computed or using other subroutines).

See section D.1.1 for an example demonstrating execution option 1.

IV.1.2. Genotype probabilities

Execution option 2 computes genotype probabilities for either designated individuals or all sample members. The genotype probabilities computed pertain to the genetic model and parameter values contained in *model.dat*. File *pap.out* includes only the ID numbers and the corresponding probabilities. For multi-locus genotypes, either single-locus or multi-locus probabilities may be output. For pedigree members untyped for a marker, *pap.out* will indicate possible incorrect probabilities due to genotype combining (section VI.3). File *pap.out* as output from option 2, sorted, and renamed *prob.dat*, serves as input to program *gpe* (section III.7).

I use execution option 2 to: **(1)** identify possible gene-carriers for a gene inferred from segregation analysis, **(2)** produce the probabilities to estimate parameters using GPEs [Hasstedt & Moll 1989] with program *gpe* (section III.7), **(3)** explore counter-intuitive results by checking the genotype assignment of selected pedigree members.

See section D.1.2 for an example demonstrating execution option 2.

IV.1.3. Simulation and output to files

Execution option 3 simulates phenotypes reflecting the genetic model and parameter values contained in *model.dat* and outputs files *header.sim, phen.1, phen.2,..., phen.n* for n replicates. Pedigree structure is fixed. Phenotypes for some pedigree members may be fixed or designated as missing. For unconditional simulation (no phenotypes fixed) programs *simrk* (section III.4) or *simul* (section III.5) may be used instead. The assigned genotypes for each individual and each replicate are output in file *pap.log*.

I use execution option 3 to: **(1)** generate distinct but similar data for each student in a linkage class, **(2)** produce replicates with identical inheritance to estimate power for segregation analysis or linkage analysis, **(3)** produce data with known inheritance in order to evaluate analysis methods.

See section D.1.3 for an example demonstrating execution option 3.

IV.1.4. Likelihood search over parameters

Execution option 4 searches for the maximum likelihood, one parameter at a time, while fixing the values of all other parameters. Since parameters are not usually independent, this approach will not attain the maximum likelihood unless executed repeatedly. File *model.dat* is

rewritten each time a better likelihood is computed. If performing a linkage analysis and ascertainment correction has not been specified, lod scores rather than likelihoods may be computed. If performing the transmission disequilibrium test and ascertainment correction has not been specified, chi-squared statistics rather than likelihoods may be computed.

I use execution option 4 when I am have difficulty maximizing, using execution option 5 (section IV.1.5), because of: **(1)** the parameters being in the vicinity of their boundaries; **(2)** the parameters being in the vicinity of values that produce zero likelihoods; **(3)** a flat likelihood surface.

See section D.1.4 for an example demonstrating execution option 4.

IV.1.5. Maximization of parameters

Execution option 5 maximizes the likelihood of specified parameters using either GEMINI [Lalouel 1979] or NPSOL [Gill et al 1986]. You may maximize any number of parameters simultaneously and may restrict two or more parameters to the same value. File *model.dat* is rewritten each time a better likelihood is computed. If performing a linkage analysis and ascertainment correction has not been specified, lod scores rather than likelihoods may be computed. If performing the transmission disequilibrium test and ascertainment correction has not been specified, chi-squared statistics rather than likelihoods may be computed.

Maximization, the primary execution option selected, simultaneously produces maximum likelihood estimates of the parameters and maximized likelihoods for statistical testing; you may also compute standard errors. However, you will save computer time by reserving standard error computation (using execution option 6) for a restricted set of models determined from a set of exploratory runs executed without standard errors. Section V.4 discusses the problems encountered when maximizing the likelihood.

See section D.1.5 for an example demonstrating execution option 5.

IV.1.6. Standard errors of parameters

Execution option 6 approximates standard errors of parameters by computing numerical second derivatives. You must provide maximum likelihood estimates; file *model.dat* output from execution option 5 serves as input. Standard errors will not be computed for a parameter estimated at either a lower or an upper bound. Execution will terminate prematurely upon finding a higher likelihood when computing derivatives; *papdr* outputs a new version of *model.dat* containing the parameter values with a higher likelihood.

I use execution option 6 to: **(1)** compute standard errors for the models for which I plan to publish parameter estimates, **(2)** verify or refute a local maximum when option 5 terminates with an equivocal code.

See section D.1.6 for an example demonstrating execution option 6.

IV.1.7. Simulation and estimation

Execution option 7 simulates phenotypes reflecting the genetic model and parameter values in *model.dat*, estimates the specified parameters in the simulated data, and outputs the average across the replicates of the estimates and the log likelihood or lod score. File *pap.log* contains the parameter estimates and log likelihood of each replicate. If performing a linkage analysis and ascertainment correction has not been specified, lod scores rather than likelihoods can be computed. If performing the transmission disequilibrium test and ascertainment correction has not been specified, chi-squared statistics rather than likelihoods can be computed.

I use execution option 7 to: **(1)** evaluate the accuracy of the mixed model approximation on my data through its ability to return estimates close to the simulated values, **(2)** test goodness-of-fit by comparing the maximum likelihood on my data to a maximum likelihood distribution produced by simulation, **(3)** determine if a set of disease pedigrees contains sufficient information to detect linkage, **(4)** estimate the number of pedigrees of a fixed structure to collect for a linkage study, **(5)** identify which pedigrees with partial marker typing are worth more typing.

See section D.1.7 for an example demonstrating execution option 7.

IV.1.8. Grid on a parameter

Execution option 8 divides user-specified ranges of one or two parameters into 5 equal intervals and computes the likelihood of each value. Program *papdr* outputs both common and natural (multiplied by -2) logarithms of the likelihood. For values of frequencies or variance components that exceed a sum of 1, *papdr* outputs 0. File *model.dat* is rewritten each time a better likelihood is computed. If performing a linkage analysis and ascertainment correction has not been specified, lod scores rather than likelihoods may be computed. If performing the transmission disequilibrium test and ascertainment correction has not been specified, chi-squared statistics rather than likelihoods may be computed.

I use execution option 8 to: **(1)** provide detail of the likelihood surface when I suspect a flat surface, **(2)** search for a higher likelihood when a parameter maximizes on the boundary (see section V.4.2), **(3)** test for linkage using a grid on the recombination probability.

See section D.1.8 for an example demonstrating execution option 8.

IV.2. Transmission Parameters

Subroutine *paptcms* (section C.2.7) assumes Mendelian segregation. Alternatively, subroutine *paptctp* (section C.2.9) allows you to estimate the allele transmission probabilities in order to test for major locus inheritance. This test compares the likelihood of the general model with estimated allele transmission probabilities to the likelihood of its submodel, Mendelian segregation; a similar comparison to the likelihood of another submodel, environmental nontransmission, tests an alternative hypothesis. Rejecting environmental nontransmission while failing to reject Mendelian segregation supports major locus inheritance.

Subroutine *paptctp* defines τ_1, τ_2, τ_3, the allele transmission probabilities, as the probability that a parent of genotype 1, 2, or 3, respectively, transmits allele 1 to an offspring. Subroutine *paptctp* then applies the Hardy-Weinberg law to compute offspring genotype transmission probabilities for each pair of parental genotypes. For this reason, the frequency subroutine *(papfq*, section C.1) selected should also assume Hardy-Weinberg equilibrium and compute the genotype frequencies from the allele frequency p.

Both Mendelian segregation and environmental nontransmission maintain a constant allele frequency across generations. The same restriction on the general model obeys the equation $p = p^2\tau_1 + 2pq\tau_2 + q^2\tau_3$. Subroutine *paptcet* (section C.2.6) includes only τ_1 and τ_3 as parameters; τ_2 is computed to conform to the constraint.

Subroutines *paptcms*, *paptctp*, and *paptcet* apply to any genotype-assignment code (section IV.10). For codes with sex-specific genotypes, the allele transmission probabilities depend on the sex of both parent and offspring [Demenais & Elston 1981]. Subroutine *paptcet* restricts the genetic model to 1 locus with 2 alleles; *paptcms* makes no restrictions. Subroutine *paptctp* allows any number of loci, but restricts the final locus in the model to 2 alleles and allows estimation of transmission for that locus, assuming Mendelian transmission for all previous loci.

See section D.2.1 for examples demonstrating the use of *paptctp*.

IV.2.1. Mendelian segregation

Mendelian transmission probabilities of $\tau_1 = 1$, $\tau_2 = ½$, and $\tau_3 = 0$ restrict a parent of genotype 1 (AA) to transmit exclusively allele A, a parent of genotype 2 (Aa) to transmit allele A with probability ½, and a parent of genotype 3 (aa) to never transmit allele A.

See section D.2.1.1 for an example demonstrating Mendelian segregation using *paptctp*.

IV.2.2. Environmental nontransmission

Environmental nontransmission results from making the transmission probabilities independent of the parental genotypes. That is, the transmission probabilities are all the same regardless of the parental or offspring genotypes. The allele frequency may be set equal to the common transmission probability to ensure equilibrium across generations. Otherwise, a different frequency may be estimated for founders than for nonfounders. This is unreasonable in multi-generation pedigrees where founders occur in all generations and at all ages. On the other hand, a frequency difference between founders and nonfounders may accurately represent nuclear families with all founders (parents) older than all nonfounders (children).

Application of the Hardy-Weinberg law in *paptctp* forces a relationship between the genotype transmission probabilities. When dominance reduces the model to two distinguishable phenotypes, the two proportions are unrestricted. However, when three phenotypes can be distinguished, their occurrence in Hardy-Weinberg proportions places a genetic constraint on a supposedly environmental model.

You can extend the environmental model to any number of alleles and loci by using *paptce* (section C.2.5), which equates the genotype transmission probabilities to the genotype frequencies. You can eliminate the Hardy-Weinberg assumption by using *paptce* with *papfqg* (section C.1.3). You can perform commingling analysis of a quantitative trait by using the environmental model with a transformation (section IV.7).

See section D.2.1.2 for an example demonstrating environmental nontransmission using *paptctp*.

IV.2.3. General transmission

Both Mendelian segregation and environmental nontransmission constitute submodels of a general transmission model with each allele transmission probability estimated to a value between 0 and 1. However, τ_1 and τ_3 often estimate to 1 and 0, respectively. Estimation on the boundary complicates determining the degrees of freedom in testing the Mendelian and environmental models.

Estimating p, τ_1, τ_2, and τ_3 ignores the assumption of equilibrium across generations made in the Mendelian and environmental nontransmission models. Alternatively, you can estimate p, τ_1 and τ_3 and compute τ_2 from p, τ_1 and τ_3 using *paptcet* (section C.2.6) [Demenais & Elston 1981].

See section D.2.1.3 for an example demonstrating general transmission using *paptctp*.

IV.3. Transmission Disequlibrium Test Parameters

The transmission disequilibrium test [Spielman & Ewens 1996] can be performed using subroutine *paptctdt* (section C.2.8) specifying a marker as the trait to be analyzed. The frequency subroutine *(papfq*, section C.1) selected must assume Hardy-Weinberg equilibrium and compute genotype frequencies from allele frequencies. Genotype-assignment code 5 (section IV.10) must be specified when running *prepap* (section III.3) designating as the category the disease of interest. Parameters of the model include transmission probabilities to both affected and unaffected offspring; the transmission to affected offspring may be estimated while fixing at 0.5 the transmission to unaffected offspring. The transmission of any number of marker alleles may be estimated simultaneously up to one fewer than the total number of alleles. The transmission parameters for different alleles may be equated, thereby reducing the number of degrees of freedom, either by setting them to maximize together in prepap, or leaving multiple alleles unmaximized. The results obtained when equating alleles may differ slightly from the results obtained upon reassigning alleles due to different genotype inference for ungenotyped individuals. The transmission to unaffected offspring can also be estimated to test for distorted segregation of the marker; however, note that all unstudied pedigree member are classified in the unaffected category along with all truly unaffected pedigree members.

The transmission parameter and chi-square statistic may be overestimated when unstudied parents have few offspring. Parental genotype inference assumes the values of the transmission parameters and can differ substantially depending on the values. One way to test whether this is occurring is to set the transmission parameters to 0.5, use option 7, simulate marker genotypes for individuals with missing values, then maximize over the transmission parameters. Using this approach, parental genotypes are inferred assuming transmission of 0.5 rather than the value used in the likelihood calculation.

See section D.2.2 for an example demonstrating general transmission using *paptctp*.

IV.4. Recombination Parameters

The recombination probability θ becomes a parameter in multi-locus models. Parameter θ may be gender-specific or equal for men and women. For more than two loci, you specify the order of the loci and θ_i represents the recombination probability between locus i and locus i + 1. When estimating θ, you may fix or simultaneously estimate other parameters. In linkage analysis, you test if θ differs significantly from ½. For the hypothesis that two traits result from the same locus, you test if θ differs significantly from 0.0.

To assume free recombination in a multi-locus model, use *paptcms* (section C.2.7). To assume linkage and autosomal inheritance, use *paptcal* (section C.2.1). For linkage with other genotype-assignment codes (section IV.10), consult the list of subroutines in section C.2. None of the linkage subroutines restrict the number of loci or alleles.

See section D.3 for examples demonstrating recombination.

IV.5. Frequency Parameters

The probability of a specified genotype for a founder equals the corresponding genotype frequency; this assumes that founders in the pedigrees constitute a random sample of the population. Since the frequencies sum to 1, a constraint reduces the number of parameters; the final frequency (for each locus or total) cannot be estimated.

To assume Hardy-Weinberg equilibrium and linkage equilibrium for the genotype frequencies, use *papfqhw* (section C.1.5). Alternatively, to assume or test for linkage disequilibrium or deviations from Hardy-Weinberg equilibrium, use another versions of *papfq* (section C.1).

See section D.4 for examples demonstrating the frequency parameters.

IV.5.1. Allele frequencies

Subroutine *papfqd* (section C.1.2) uses linkage disequilibrium D to compute genotype frequencies. D equals the excess of the haplotype frequency over the product of allele frequencies (locus 1, allele 1 and locus 2, allele 1). Therefore, the range of D depends on the allele frequencies; 0 represents equilibrium. Subroutine *papfqd* restricts the genetic model to two loci with 2 alleles each. You might select this parameterization over *papfqc* (section IV.5.2) in order to fix the allele frequencies.

See section D.4.1 for examples demonstrating the use of *papfqd*.

IV.5.2. Conditional allele frequencies

Subroutine *papfqc* (section C.1.1) uses conditional allele frequencies in computing the genotype frequencies. Any locus in the model can be conditional on another locus; the locus with conditional frequencies must have 2 alleles. Other loci may have any number of alleles. Equality of the conditional allele frequencies equates to linkage equilibrium; significant differences indicate association between the loci. You might select this parameterization over *papfqd* (section IV.5.1) for diseases associated with a marker locus. For example, estimate the frequency of the allele for diabetes mellitus (locus 1) conditional on HLA-DR3, conditional on HLA-DR4, and conditional on HLA-DRX (locus 2).

See section D.4.2 for examples demonstrating the use of *papfqc*.

IV.5.3. Haplotype frequencies

Subroutine *papfqh* (section C.1.4) uses haplotype frequencies in computing the genotype frequencies assuming Hardy-Weinberg equilibrium. This parameterization removes any constraints on the numbers of loci or alleles.

See section D.4.3 for examples demonstrating the use of *papfqh*.

IV.5.4. Genotype frequencies

Subroutine *papfqg* (section C.1.3) uses genotype frequencies directly, eliminating all assumptions about relationships between the frequencies.

See section D.4.4 for examples demonstrating the use of *papfqg*.

IV.6. Discrete Trait Parameters

For a discrete trait, the genetic model assumes an underlying continuous liability scale distributed as a normal density within each genotype. Disease occurs when liability exceeds a threshold determined such that the integral to the right of the threshold within each genotype equals the corresponding affection probability. The model may include any number of loci; for models with more than one locus, additivity on the liability scale is assumed.

All the subroutines for discrete data except *dmlprsv* (section IV.6.5) apply only to dichotomous (unaffected/affected) phenotypes. All the subroutines for discrete data except *dmlprpn* (section IV.6.1) restrict the genetic model to two alleles with the second the disease allele. Subroutines *dmlprin*, *dmlprinc*, *dmlprpr* and *dmlprsv* require that the population incidence or prevalence figures be entered into *popln.dat* (section II.5).

See section D.5 for examples demonstrating the discrete trait parameters. See sections IV.6.1 through IV.6.5 for descriptions of the discrete trait parameters.

IV.6.1. Affection probability

Subroutine *dmlprpn* (section C.3.3) uses as parameters the affection probability for each genotype. An advantage of this parameterization is that the genetic model can include any number of alleles. A disadvantage of this parameterization is that disease prevalence cannot be restricted to the same value for different models. You can restrict the penetrances to increase across genotypes by editing file *model.dat* (section B.7) to specify linear constraints on the parameters.

Subroutines *dmlprpr* (section IV.6.2) or *dmlprin* (section IV.6.3) are alternatives if the disease is age-dependent. Subroutine *dmlprsv* (section IV.6.4) can be used to restrict the prevalence.

See section D.5.1 for an example demonstrating the use of *dmlprpn*.

IV.6.2. Prevalence

Subroutine *dmlprpr* (section C.3.4) restricts the affection probabilities to correspond to age-specific (and possibly gender-specific) population prevalence figures (given in *popln.dat*). The subroutine allows only 2 alleles per locus, but any number of loci. The prevalence figures determine a series of points T_i ordered inversely on the liability scale such that the sum over the genotypes of the integrals from T_i to ∞ equals the corresponding prevalence. Therefore, the model assumes that earlier onset corresponds to higher liability. For an affected individual with age at examination in interval i, the affection probability for a genotype equals the integral from T_i to ∞ of the normal density for that genotype. For an unaffected individual with age at examination in interval i, the affection probability for a genotype equals the integral from $-\infty$ to T_{i-1} of the normal density for that genotype.

The disease allele must be the second of the two alleles. For any individual with disease status assigned, age at examination must be available and entered as a separate column in *phen.dat* (section II.4).

The major locus effect is parameterized as dominance d and displacement t. If we represent the mean liability for the three genotypes at one locus as μ_1, μ_2, and μ_3 and the variance within genotypes as σ^2 with the total mean and variance restricted to equal 0 and 1, respectively, $d = (\mu_2 - \mu_1)/(\mu_3 - \mu_1)$ and $t = (\mu_3 - \mu_1)/\sigma$. Note that t differs from the definition of displacement in Morton & MacLean [1974]. For multi-locus models, d and t become locus-specific and t adds across loci. The output in *pap.out* includes the affection probability for each age interval as computed for each genotype.

For a gender-specific model, use category-specific autosomal inheritance (genotype-assignment code = 5) and include prevalence figures for both males and females in *popln.dat*. Dominance and displacement are gender-specific.

Subroutine *dmlprin* (section IV.6.3) is an alternative if onset age is available. Subroutines *dmlprpn* (section IV.6.1) or *dmlprsv* (section IV.6.4) can be used if the disease is not age-dependent.

See section D.5.2 for an example demonstrating the use of *dmlprpr*.

IV.6.3. Incidence

Subroutine *dmlprin* (section C.3.1) restricts the affection probabilities to correspond to age-specific (and possibly gender-specific) population incidence figures (given in *popln.dat*). The subroutine applies only to 2-allele inheritance. The incidence figures determine a series of points T_i ordered inversely on the liability scale such that the sum over the three genotypes of the integrals from T_i to T_{i-1} equals the corresponding incidence. Therefore, the model assumes that earlier onset corresponds to higher liability. For an affected individual with onset age in interval i, the affection probability for a genotype equals the integral from T_i to T_{i-1} of the normal density for that genotype. For an affected individual with unknown onset age but age at examination in interval i, the affection probability for a genotype equals the integral from T_i to ∞ of the normal density for that genotype. For an unaffected individual with age at examination in interval i, the affection probability for a genotype equals the integral from $-\infty$ to T_{i-1} of the normal density for that genotype.

The disease allele is the second of the two alleles. Onset age or age at examination must be available for any individual assigned disease status as separate columns in *phen.dat* (section II.4).

The major locus effect is parameterized as dominance d and displacement t. If we represent the mean liability for the three genotypes at one locus as μ_1, μ_2, and μ_3 and the variance within genotypes as σ^2 with the total mean and variance restricted to equal 0 and 1, respectively, $d = (\mu_2 - \mu_1)/(\mu_3 - \mu_1)$ and $t = (\mu_3 - \mu_1)/\sigma$. Note that t differs from the definition of displacement in Morton & MacLean [1974]. For multi-locus models, d and t become locus-specific and t adds across loci. The output in *pap.out* includes the affection probability for each age interval as computed for each genotype.

For a gender-specific model, use category-specific autosomal inheritance (genotype-assignment code = 5) and include incidence figures for both males and females in *popln.dat*. Dominance and displacement are gender-specific.

Subroutine *dmlprpr* (section IV.6.2) is an alternative if onset age is not available. Subroutine *dmlprpn* (section IV.6.1) or *dmlprsv* (section IV.6.4) can be used if the disease is not age-dependent. For a disease for which incidence varies with some measured characteristic, such as level of obesity or smoking status, use subroutine *dmlprinc* instead; specify the category for each individual as an integer in file *phen.dat* and repeat the incidence for all age groups for the number of categories in file *popln.dat*.

See section D.5.3 for an example demonstrating the use of *dmlprin*.

IV.6.4. Severity Classes

Subroutine *dmlprsv* (section C.3.4) restricts the affection probabilities to correspond to severity-specific (and possibly gender-specific) population prevalence figures (given in *popln.dat*). The subroutine applies only to 2-allele inheritance. The prevalence figures determine a series of points T_i, i increasing across the liability scale, such that the sum over the three genotypes of the integrals from T_{i-1} to T_i equals the corresponding prevalence. Therefore, the model assumes that greater severity corresponds to higher liability. For an individual affected at severity i, the affection probability for a genotype equals the integral from T_{i-1} to T_i of the normal density for that genotype. For an unaffected individual, the affection probability for a genotype equals the integral from $-\infty$ to T_1 of the normal density for that genotype.

The disease allele is the second of the two alleles. Severity may be coded in any manner. However, the severity categories in *popln.dat* should be ordered with unaffected first, then disease categories through increasing levels of severity. The corresponding ranges will associate the appropriate severity codes with each category. For an unaffected/affected dichotomy, *popln.dat* contains two entries, for unaffected and affected.

The major locus effect is parameterized as dominance d and displacement t. If we represent the mean liability for the three genotypes at one locus as μ_1, μ_2, and μ_3 and the variance within genotypes as σ^2 with the total mean and variance restricted to equal 0 and 1, respectively, $d = (\mu_2 - \mu_1)/(\mu_3 - \mu_1)$ and $t = (\mu_3 - \mu_1)/\sigma$. Note that t differs from the definition of displacement in Morton & MacLean [1974]. For multi-locus models, d and t become locus-specific and t adds across loci. The output in *pap.out* includes the affection probability for each age interval as computed for each genotype.

For a gender-specific model, use category-specific autosomal inheritance (genotype-assignment code = 5) and include prevalence figures for both males and females in *popln.dat*. Dominance and displacement are gender-specific.

Subroutines *dmlprpr* (section IV.6.2) and *dmlprin* (section IV.6.3) are alternatives for age-dependent diseases, but allow only an unaffected/affected dichotomy.

See section D.5.4 for an example demonstrating the use of *dmlprsv*.

IV.7. Quantitative Trait Parameters

For a quantitative trait, the phenotypes are assumed to distribute normally within genotypes. The penetrance equals the height of the normal density (including $1/\sqrt{2\pi}$) as the phenotype. You can scale, standardize, or transform phenotypes by specification in file *header.dat* (see section II.3).

The effects on the genotype means of up to ten covariates may be estimated for each quantitative trait when using subroutines *qmlprmv* or *qmlprdd*. This extension allows for genotype-specific age or measured environmental effects. Only linear effects are estimated; for cross or squared terms, include the product or square as a variable in file *phen.dat* and treat as another covariate

Quantitative phenotypes may be transformed using the function $y = r/P[(x/r+1)^P-1]$ [Maclean et al 1976], where x represents the phenotype, P represents the power and r = 6. Set P = 1 for no transformation. Estimating P within a 1-genotype model allows you to determine the transformation to correspond to a single normal density. Estimating P within the environmental model (section IV.2.2) allows you to test for a mixture of distributions. If you then insert the estimate of P into *header.dat* (section II.3) and run *preped* (section III.1), further analyses will consider the transformed phenotypes. Estimating P within a genetic model allows transformation of the phenotypes to obtain the best fit to each model.

Only subroutines *qmlprdd* (section IV.7.2) and *qmlprddp* (section) restrict the number of alleles to two and the within-genotype standard deviations to equality.

See section D.6 for examples demonstrating the quantitative trait parameters. See sections IV.7.1 through IV.7.3 for descriptions of the quantitative trait parameters.

IV.7.1. Means/Standard Deviations

Subroutine *qmlprmv* (section C.4.3) parameterizes the model as means μ_i, standard deviations σ_i, and slope s_{ji} for covariate j for each genotype i. The number of loci and alleles included in the genetic model are unrestricted.

See section D.6.1 for examples demonstrating the use of *qmlprmv*.

IV.7.2. Dominance/Displacement

Subroutine *qmlprdd* (section C.4.1) restricts the genetic model to two alleles. The parameters comprise the total mean μ_T, total standard deviation σ_T, dominance d, displacement t for the mean, mean slope s_j for covariate j, and dominance d_{sj}, displacement t_{sj} for covariate j. If p represents the frequency of allele 1, q = 1 - p, μ_i and s_{ij} represent the mean and slope of covariate j, respectively, for genotype i, i = 1, 2, 3, and σ represents the within-genotype standard deviation, then $\mu_T = p^2\mu_1 + 2pq\mu_2 + q^2\mu_3$, $s_j = p^2s_{1j} + 2pqs_{2j} + q^2s_{3j}$, $\sigma_T^2 = p^2[\mu_1^2 + \Sigma s_{1j}^2] + 2pq[\mu_2^2 + \Sigma s_{2j}^2] + q^2[\mu_3^2 + \Sigma s_{3j}^2] + \sigma^2$, $d = (\mu_2 - \mu_1)/(\mu_3 - \mu_1)$, $t = (\mu_3 - \mu_1)/\sigma$, $d_{sj} = (s_{2j} - s_{1j})/(s_{3j} - s_{1j})$, $t_{sj} = (s_{3j} - s_{1j})/\sigma$. The reverse equations equal: $\mu_1 = 2pqdt - q^2t$, $\mu_2 = \mu_1 + dt$, $\mu_3 = \mu_1 + t$, $s_{1j} = 2pqd_{sj}t - q^2t_{sj}$, $s_{2j} = s_{1j} + d_{sj}t_{sj}$, $s_3 = s_{1j} + t_{sj}$. Note that t differs from the definition of displacement in Morton & MacLean [1974]. For multi-locus models, d, t, d_{sj}, and t_{sj} are locus-specific and t and t_{sj} add across loci.

See section D.6.2 for examples demonstrating the use of *qmlprdd*.

IV.7.3. Means/Standard Deviations/Threshold

Subroutine *qmlprmvt* (section C.4.4) parameterizes the model as means for each genotype μ_i, standard deviations for each genotype σ_i, and threshold T. Parameter T specifies the lower limit of the trait for individuals whose phenotypes are impossible to obtain because medication or the disease process has altered the level. For example, medicated hypertensives might be included in an analysis of blood pressure by specifying that their levels exceed the diagnostic threshold. Each would be designated a missing value for the quantitative trait in *phen.dat*, and as affected for a corresponding disease trait. The number of loci and alleles included in the genetic model are unrestricted.

See section D.6.3 for examples demonstrating the use of *qmlprmvt*.

IV.7.4. Dominance/Displacement/Proportion

Subroutine *qmlprddp* (section C.4.2) restricts the genetic model to two alleles. The parameters comprise the total mean μ_T, total standard deviation σ_T, dominance d, displacement t, and proportion P. If p represents the frequency of allele 1, q = 1 - p, μ_i represent the genotype means, i = 1, 2, 3, and σ represent the within-genotype standard deviation, then $\mu_T = p^2\mu_1 + 2pq\mu_2 + q^2\mu_3$, $\sigma_T^2 = p^2\mu_1^2 + 2pq\mu_2^2 + q^2\mu_3^2 + \sigma^2$, $d = (\mu_2 - \mu_1)/(\mu_3 - \mu_1)$, and $t = (\mu_3 - \mu_1)/\sigma$. Note that t differs from the definition of displacement in Morton & MacLean [1974]. For multi-locus models, d and t become locus-specific and t adds across loci. Parameter P specifies the proportion of individuals in the population whose phenotypes are impossible to obtain because medication or the disease process has altered the level. For example, medicated hypertensives might be included in an analysis of blood pressure by specifying that their levels exceed the 95^{th} percentile. Each would be designated a missing value for the quantitative trait in *phen.dat*, and as affected for a corresponding disease trait.

See section D.6.4 for examples demonstrating the use of *qmlprddp*.

IV.8. Within-Genotype Parameters

The variation within the normal densities for each genotype may be attributed to polygenes or environmental factors shared by pedigree members, which contribute to the correlation between pedigree members, or to individual-specific environmental factors. No correlations are possible between members of different pedigrees.

Subroutine *papwgml* (section C.5.3) assumes all variation within genotypes is due to individual-specific environmental factors. There are no associated parameters.

Subroutine *papwgvc* (section C.5.4) attributes within-genotype variation to variance components: heritability h^2 and shared environmental factors c_i^2. The number of shared environmental components is unrestricted; each corresponds to a column in *phen.dat* (section II.4) designating individuals who share an environmental effect. Parameters h^2 and c_i^2 are proportions of the within-genotype variance. Note that h^2 differs from the definition of heritability in Morton & MacLean [1974].

Subroutine *papwgam* (section C.5.1) adds assortative mating [Hasstedt 1995] to the variance components model. The added parameter comprises a correlation a^2, which is constant across the range of the phenotype and contributes to a within-genotype correlation between spouses and a modification of the major locus genotype frequencies. When including assortative mating, the genetic model is restricted to autosomal inheritance of one locus with two alleles, the genotype means or penetrances must be non-decreasing across the genotypes, and the alleles at multiple loci must be linkage equilibrium. You can restrict the genotype means or penetrances to increase across genotypes by editing file *model.dat* (section B.7) to specify linear constraints on the parameters. Program *preped* (section III.1) generates a different input file when the model includes assortative mating.

Subroutine *papwgfc* (section C.5.2) parameterizes within-genotype correlations according to genetic relationships. The gender-specific familial correlations comprise husband-wife ρ_{hw}, mother-daughter ρ_{md}, mother-son ρ_{ms}, father-daughter ρ_{fd}, father-son ρ_{fs}, sister-sister ρ_{ss}, sister-brother ρ_{sb}, and brother-brother ρ_{bb}. It is possible to restrict $\rho_{md} = \rho_{ms} = \rho_{fd} = \rho_{fs}$ and $\rho_{ss} = \rho_{sb} = \rho_{bb}$. Outside the nuclear family, zero correlation is assumed for all relative pairs. Gender must be included in *phen.dat* (section II.4) for any individual assigned a trait phenotype.

Subroutine *papwgml* is used with *papend/papcr* for discrete traits, *papenq/papcr* for quantitative traits, or *papendq/papcr* for models including both discrete and quantitative traits. To compute the likelihood exactly, subroutines *papwgvc*, *papwgam*, and *papwgfc* are used with *papende/papcrde* for discrete traits, *papenqe/papcrqe* for quantitative traits, or *papendqe/papcrdqe* for models including both discrete and quantitative traits. Exact computation of the likelihood when the model includes correlations between pedigree members requires summing over the probabilities of all combinations of genotypes for pedigree members; exact calculation requires too much computer time for pedigrees sizes exceeding about ten members. Alternatively, *papenda/papcrda* for discrete traits, *papenqa/papcrqa* for quantitative traits, or *papendqa/papcrdqa* for models including both discrete and quantitative traits approximate the likelihood [Hasstedt 1993].

See section D.7 for examples demonstrating the within-genotype parameters.

IV.9. Multivariate Parameters

One marker and any number of traits may be associated with each locus in the genetic model; a trait, but not a marker, may be associated with multiple loci. You specify the locus-variable associations through responses to queries from *prepap* (section III.3).

Associating a marker and trait with the same locus produces measured genotype analysis [Boerwinkle et al 1986]. For example, you can estimate means for total serum cholesterol level within genotypes for apolipoprotein E defined by electrophoresis. By associating total serum cholesterol with a second locus as well, you can simultaneously account for the LDL receptor defect.

Using the within-genotype subroutines *papwgml* or *papwgfc* (section IV.8), the multivariate parameter, for each pair of traits, comprises the correlation between the traits ρ. Using the within-genotype subroutine *papwgvc* (section IV.8), each correlation is partitioned into two: the genetic correlation ρ_g reflects the effect of the same polygenes on both traits; the environmental correlation ρ_e derives from shared environmental factors. For more than two traits, correlations between all possible pairs may be estimated.

Estimates of ρ_g and ρ_e often differ greatly. For example, one may be positive and the other negative. It is unlikely that such estimates reflect biological reality, but rather result from a violation of the assumptions of the model. In such cases ρ_g and ρ_e can be held to the same value for maximization. File *model.dat* must be edited to make this change.

See section D.9 for examples demonstrating the multivariate parameters.

IV.10. Genotype-Assignment Codes

PAP allows seven genotype-assignment codes. The options include: **(1)** autosomal, **(2)** X-linked, **(3)** parent-specific autosomal, **(4)** parent-specific X-linked, **(5)** category-specific autosomal, **(6)** autosomal/X-linked mixed, **(7)** autosomal/X-linked admixture.

For each option, you may specify any number of loci and any number of alleles per locus. Program *prepap* (section III.3) presents the order of the genotypes when requesting parameter values. Alternatively, section VI.2 describes the order of the genotypes.

See section D.10 for examples demonstrating genotype-assignment codes. See sections IV.10.1 through IV.10.7 for descriptions of the genotype assignment codes.

IV.10.1. Autosomal inheritance

This is the standard model for autosomal inheritance.

See section D.10.1 for an example demonstrating autosomal inheritance.

IV.10.2. X-linked inheritance

This is the standard model for X-linked inheritance.

See section D.10.2 for an example demonstrating X-linked inheritance.

IV.10.3. Parent-specific autosomal

This extension of autosomal inheritance distinguishes between heterozygotes according to the parental origin of each allele. This distinction allows the penetrance or transmission probabilities to differ between the two types of heterozygotes. For example, parent-specific autosomal genotype assignments allow you to independently estimate parameters describing age of onset in maternally- and paternally-transmitted disease.

See section D.10.3 for examples demonstrating parent-specific autosomal inheritance.

IV.10.4. Parent-specific X-linked

This extension of X-linked inheritance distinguishes between heterozygotes according to the parental origin of each allele. This distinction allows the penetrance or transmission probabilities to differ between the two types of heterozygotes.

See section D.10.4 for examples demonstrating parent-specific X-linked inheritance.

IV.10.5. Category-specific autosomal

This extension of autosomal inheritance allows you to classify individuals in two categories and independently specify the penetrance or transmission probabilities for all genotypes. Possible dichotomies for separate parameter specification include males and females, nonsmokers and smokers, unmedicated individuals and individuals taking medication. Corresponding to the assignments of genotypes for females (category 2) before males (category 1), category 2 always precedes category 1 in the genotype order.

The genotype order for category-specific autosomal inheritance equals the order for the autosomal/X-linked mixed model (section IV.10.6).

See section D.10.5 for examples demonstrating category-specific autosomal inheritance.

IV.10.6. Autosomal/X-linked mixed model

The autosomal/X-linked mixed model [Hasstedt & Skolnick 1984] encompasses both autosomal and X-linked inheritance. Therefore, successively comparing the likelihoods of the autosomal and X-linked models to the likelihood of the general model tests the two modes of inheritance. Rejection of both modes of inheritance suggests an intermediate form of transmission that is not interpretable as genetic. Alternatively, the general form of the autosomal/X-linked admixture model (section IV.10.7) more realistically assumes that alleles with each mode of inheritance occur in the sample. However, the admixture model requires more parameters than the mixed model and may not be not realistic for a single pedigree.

The autosomal/X-linked mixed model includes three genotypes for males. When the parameters correspond to X-linkage, the frequency of the third genotype equals 0. You must define the alleles to specify phenotypic equivalence between the second and third genotypes; a dominant model has the normal allele first; a recessive model has the disease allele first; a codominant model is nonsensical.

Transmission probabilities from a father with genotype 2 distinguish between autosomal and X-linked inheritance. If both the son and daughter allele transmission probabilities equal ½, you obtain autosomal inheritance. If the son allele transmission probability equals 1 and the daughter allele transmission probability equals 0, you obtain X-linked inheritance. All the other transmission probabilities assume their Mendelian values.

Frequency subroutine *papfqax* assumes Hardy-Weinberg equilibrium for females and generational equilibrium for males. The frequencies in males of the three genotypes equal:

(1) $p^2 \tau/[q + (p - q)\tau]$,
(2) $pq/[q + (p - q)\tau]$,
(3) $q^2 (1-\tau)/[q + (p - q)\tau]$,

where τ represents the son's allele transmission probability. If $\tau = ½$, you obtain autosomal frequencies; if $\tau = 1$, you obtain X-linked frequencies. Subroutine *papfqax* restricts the model to 1 locus with 2 alleles.

See section D.10.6 for examples demonstrating autosomal/X-linked mixed model.

IV.10.7. Autosomal/X-linked admixture

The autosomal/X-linked admixture model encompasses both autosomal and X-linked inheritance. Therefore, successively comparing the likelihoods of the autosomal and X-linked models to the likelihood of the general model tests the two modes of inheritance. Rejection of both submodels supports heterogeneity in the mode of inheritance. Alternatively, the

autosomal/X-linked mixed model (section IV.10.6) requires fewer parameters for testing the alternative modes of inheritance, but has an unrealistic general model.

An autosomal/X-linked admixture model requires a minimum of two loci, one autosomal and one X-linked. The submodels, autosomal or X-linked inheritance, restrict locus 1 or 2, respectively, to 1 allele. Additional loci may represent markers linked to the autosomal or X-linked form.

See section D.10.7 for examples demonstrating autosomal/X-linked admixture.

V. PERFORMING THE ANALYSIS

You perform an analysis by defining a set of models and maximizing the likelihood of each using *papdr*. For each model, you:

(1) Execute *prepap* to write *model.dat*, specifying the model and parameter values and indicating the parameters to maximize.

(2) Link *papdr*, incorporating the selected versions of *papfq, paptc, dmlpr, qmlpr, papwg, papen, papcr* (listed in *model.dat*).

(3) Execute *papdr*, selecting option 5.

(4) Examine your results in *pap.out*.

Some investigators feel the testing sequence should proceed from complex to simple models. However, when maximizing likelihoods, following the reverse sequence allows you to use parameter estimates from one analysis as initial values in the next analysis.

V.1. Preparing for the Analysis

The following procedure details how you prepare the data files to initiate an analysis. In addition, when you modify the data files, you should repeat step **(3)**.

(1) Assign a unique ID number to each pedigree member and input *trip.dat* (section II.1), *header.dat* (section II.3), and *phen.dat* (section II.4). File *ascer.dat* (section II.4) is also needed if you are making an ascertainment correction.

(2) Enter frequency and phenotype information about all markers and incidence or prevalence figures about diseases as needed in *popln.dat* (section II.5).

(3) Execute *preped* (section III.1) to combine *trip.dat* (section II.1) and *phen.dat* (section II.4) into *papin.dat* for input to *papdr*. If *preped* terminates due to exceeding the array dimensions:

(a) Increase the parameter values in the include file.
(b) Compile all implicated source routines and relink *preped*.
(c) Execute *preped*.
(d) Repeat until *preped* terminates normally.

If *preped* terminates because of an error:

(a) Correct *trip.dat (section II.1)*,
(b) Execute *preped*.
(c) Repeat until *preped* terminates normally.

V.2. Checking for Errors

Errors occur easily at any stage of an analysis. Perform preliminary tests to assure correct data files before beginning an analysis; repeat the tests upon modifying the data files. Verify the correctness of each result.

V.2.1. Performing preliminary tests

The following tests identify some errors in the data files. You should not proceed with the analysis until you have completed these tests successfully.

(1) For each quantitative trait, use *descstat* to check that the sample size, mean, and variance are correct.

(2) For each discrete trait, use *descstat* to check for the correct counts by affection status.

(3) For each marker, use *descstat* to check for the correct count, use *papdr* to check for offspring inconsistent with their parents.

V.2.2. Verifying analysis results

You should not accept analysis results without critically examining them for correctness. Suggested tests follow:

(1) Compare the count which is output to the console by *papdr* to the sample size.

(2) Examine the model information output to the console by *papdr* to assure that the model was correctly specified and that parameter equivalences were correctly defined.

(3) Check the termination code in *pap.out* to assure that a maximum has been obtained.

(4) For a quantitative trait, compare the proportions and means in the two components of the recessive and dominant models. If they differ greatly, assign initial values to correspond to the estimates of the other and repeat the maximization of each model.

(5) Confirm that submodels have lower likelihoods than general models.

(6) Compare the estimates to the results of other studies and to your expectations. For example, question a high frequency when a pedigree was ascertained through a rare trait.

V.3. Applying Constraints to Parameter Values

When you estimate multiple allele frequencies at a locus, or multiple variance components for a trait, *prepap* (section III.3) outputs linear constraints in *model.dat* (section B.7) to constrain the sum of the frequencies or variance components to less than one. You can constrain other parameters by editing *model.dat* to add the constraints. For example, to ensure that penetrances or genotypic means increase across genotypes, you need two constraints. The first specifies boundaries of 0 and 1 and gives coefficients of –1 for the penetrance of genotype 1 and +1 for the penetrance of genotype 2. The second again specifies boundaries of 0 and 1 and gives coefficients of –1 for the penetrance of genotype 2 and +1 for the penetrance of genotype 3. In both cases, the coefficients of all other parameters equal 0. These constraints specify that $0 < f_2 - f_1$ and $0 < f_3 - f_2$, where f_i designates the penetrance of genotype i.

See Section B.7 for the format of the linear constraints in *model.dat*.

V.4. Maximizing the Likelihoods

Maximization comprises one of the most difficult aspects of an analysis. Both multiple local maxima and boundary maxima complicate the maximization procedure.

Both GEMINI and NPSOL perform better on parameters of similar magnitude. Since some parameters are probabilities ranging from 0 to 1, you might scale your quantitative traits to have a similar mean and standard deviation. You can standardize or scale your traits through specification in *header.dat* (section II.3).

V.4.1. Finding the global maximum

Both GEMINI and NPSOL find only local maxima, usually in the region of the initial values of the parameters. However, the parameter space may contain multiple maxima, some inside and others outside the region of interest. One option is to start the maximization from a number of different initial values; upon selecting execution option 5, respond positively when asked about selecting random initial values; the boundaries on the parameters given in *model.dat* must be reasonable values for the parameters.

Alternatively, a negative response to any of the following questions indicates failure to obtain the appropriate maximum.

(1) Do the estimates make sense and conform to other information about the trait?
(2) Is the likelihood higher than the likelihood of all submodels?
(3) Do the estimates for the recessive and dominant model represent similar proportions and means for each distribution?
(4) Do heterozygotes have an intermediate mean for the codominant model?

V.4.2. Determining when a maximum occurs on the boundary

Both NPSOL and the latest version of GEMINI check for boundary values of the parameters as part of the maximization procedure. Nevertheless, either may mistakenly conclude that the likelihood maximizes with the parameter on the boundary. In particular, this mistake may occur when attaining the true maximum requires that two boundary parameters change simultaneously.

Some parameters, such as means and standard deviations, seldom present boundary problems. Their parameter space either encompasses the complete range (means) or the maximum necessarily occurs away from the boundary (standard deviations). But other parameters, in particular transmission probabilities and affection probabilities, frequently have a boundary maximum. To complicate matters further, frequencies or variance components sum to one, requiring that the upper bound of the second and following values depend on the previous values. For a variance components model, difficulty locating an interior maximum may indicate violation of the assumptions of the model.

Some approaches to exploring the boundary region for an interior maximum when parameter φ maximized on the bound follow:

(1) Fix φ to the bound or slightly away and estimate the other parameters.
(2) Fix the other parameters to their estimated values and grid φ over a narrow range to search for a higher likelihood.
(3) If the likelihood is higher when φ is away from the bound, restart the maximization with the parameters at those values.
(4) If the likelihood is highest when φ is on the bound, conclude that the maximum occurs with φ on the bound.

V.5. Interpreting the Results

The magnitude of an isolated likelihood is meaningless; likelihoods derive meaning through comparison with other likelihoods. When computed for discrete data, the likelihood equals a probability, therefore a number less than 1. Therefore, the logarithm of the likelihood is negative, and a higher likelihood is a negative number of smaller magnitude. When computed for quantitative data, the penetrance equals the height of a normal density, which is not a probability. Generally, the likelihood ranges from below 1 for larger standard deviations to above 1 for very small standard deviations. When the likelihood exceeds 1, the logarithm of the likelihood is positive, and a higher likelihood is a positive number with larger magnitude.

To test a hypothesis, you compare the maximized likelihood of a submodel to the maximized likelihood of a general model. You form a submodel by restricting one or more parameters estimated in the general model. You may restrict a parameter by setting it to a particular value, such as setting the heterozygote allele transmission probability to ½. Or you may

restrict a parameter by setting it equal to another parameter, such as equating two genotype means to specify a dominance relationship. A submodel must always have a lower likelihood than the general model. If not, the maximum likelihood has not been obtained for the general model.

Investigators usually test hypotheses using a chi-square test. However, the application to pedigrees may violate the assumptions [Cannings et al 1980] necessary for negative two multiplied by the natural logarithm of the likelihood ratio to approximate a chi-square distribution. The degrees of freedom for the test equals the difference in the number of parameters estimated in obtaining the two likelihoods. Self & Liang [1987] address the properties of the chi-square distribution when a parameter maximizes on the boundary.

VI. UNDERSTANDING THE PACKAGE

The following sections describe the functioning of different aspects of PAP and in conjunction a selection of subroutines. For other descriptions, see section VII for *papfq, paptc, dmlpr, qmlpr, papwg*, section E for *filerr, daterr, moderr, simerr*, and *paperr*, and the NPSOL manual for *objfun* and *confun*. Appendix A briefly describes all the subroutines.

VI.1. Computation Order

Program *preped* (section III.1) calls subroutine *pedgen* to assign an order for likelihood calculation to pedigree members and outputs the phenotypes in this order in *papin.dat* (section B.6). The output excludes any individuals included in *trip.dat* but missing from *phen.dat* who are not needed to connect the pedigree.

The final family for the computational sequence is selected as the family in which both parents are founders; if both parents are founders in more than one family, the program selects the one with the most descendants. Generation of the sequence proceeds in reverse by selecting the nuclear family in which the last offspring is a parent, then the last offspring of that family, and so on, returning to the previous offspring whenever a line of descent has been completed.

The computational order generated by *pedgen* may not yield the fastest computation for pedigrees that contain loops. The user can override the default order by specifying a preferred order in file *order.dat* (section II.2).

VI.2. Genotype Order

The general rules determining genotype order are: female genotypes precede male genotypes, X-linked loci precede autosomal loci, haplotypes are formed first and then genotypes from the haplotypes (within sets of autosomal or X-linked loci), and a heterozygote for which the maternal allele has the lower number precedes the reverse heterozygote.

In the examples given, upper and lower case letters represent alleles 1 and 2, respectively, at each locus. The first locus is A, the second is B, the third is C.

Haplotypes are incremented through the final locus first, then the next to last, etc. Therefore the ordering for two loci is: AB, Ab, aB, ab. The ordering for three loci is: ABC, ABc, AbC, Abc, aBC, aBc, abC, abc.

Genotypes are formed from haplotypes across a lower triangular matrix of haplotype pairs. For two loci, there are 10 genotypes, as follows:

	AB	Ab	aB	ab
AB	1			
Ab	2	3		
aB	4	5	6	
ab	7	8	9	10

For an X-linked model (option 2), male genotypes 11 through 14 equal AB, Ab, aB, ab. For the category-specific model (option 5) or autosomal/X-linked mixed model (option 6), category 1 (male) genotypes 11 through 20 are defined as for category 2 (females).

For autosomal/X-linked admixture (option 7), the autosomal and the X-linked genotypes are formed separately; the final genotype is the product of the two sets.

For parent-specific autosomal inheritance (option 3), there are 16 genotypes defined as:

	AB	Ab	aB	ab
AB	1	2	3	4
Ab	5	6	7	8
aB	9	10	11	12
ab	13	14	15	16

Genotypes 2 and 5 differ in that b came from the father for genotype 2 and from the mother for genotype 5.

For parent-specific X-linked inheritance (option 4), there are 4 additional genotypes for males, genotype 17 through 20.

Subroutine *papgn* stores 1-locus genotypes in array IGENLC(IG, LOC) and genotypes for each trait in array IGTRAT(IG, ITRAIT) where IG represents the multi-locus genotype, LOC represents the locus, and ITRAIT represents the trait. In addition, *papgn* fills array IALLEL with the 2 alleles at each locus corresponding to each genotype. These arrays facilitate penetrance assignments.

VI.3. Genotype Combining

Untyped pedigree members slow the likelihood computation on marker data since all genotypes must be considered for these individuals. However, two facts allowed me to speed the computation: **(1)** the possible marker genotypes for an untyped individual are limited by the types on descendants; **(2)** when a parental allele does not occur in any descendant, its exact identity does not affect the likelihood. These two facts were translated into a faster algorithm by: **(1)** determining beforehand the set of possible genotypes for each untyped individual and restricting the likelihood computation to that set; **(2)** combining into a single allele, the set of alleles which comprise the candidates for an unobserved allele, further restricting the set of possible genotypes.

Subroutine *pappn* (section VI.8) stores in PENTR zero for each impossible or eliminated genotype and, for each founder, stores the summed frequency for all combined genotypes for each founder with a missing value.

Subroutine *papmv* eliminates impossible genotypes for each untyped individual by examining the alleles in parents and offspring. If only one genotype is possible, *papmv* stores the genotype as the phenotype. If more than one genotype is possible, *papmv* checks if the possible genotypes include all the alleles, which indicates that genotypes can be combined. If genotypes cannot be combined, *papmv* assigns a phenotype corresponding to indicators in ICOMBN designating which genotypes are possible. If genotypes can be combined, *papmv* call *alloff* to designate the alleles that can be combined, and stores indicators in ICOMBN.

Subroutine *alloff* identifies alleles which occur in descendants and were possibly transmitted by the individual (ALD) and all alleles that occur in the individuals descent group (ALP).

Subroutine *repeat* makes *alloff* recursive by calling *alloff* with an offspring the designated individual.

Please note that the genotype probabilities (*papdr* execution option 2) computed for an untyped individual may not be exactly correct. The probability of the combined genotype is computed, and the frequencies are used to distribute that probability among the possible genotypes.

VI.4. Ascertainment Correction

Families for genetic studies are usually ascertained through one or more family members presenting with a particular phenotype in order to enrich the sample for the trait of interest. Consequently, the data are not a random sample of the population and correction for the method of ascertainment in the analysis is necessary in order to obtain parameter estimates appropriate for the population. However, it should be noted that it may be impossible to obtain population parameter estimates from an ascertained sample [Burton et al 2001]. Of the available methods of ascertainment correction, the classical method applies only to nuclear families and the method of Thompson & Cannings [1980] assumes single selection (ascertainment probability is very small). The ascertainment-assumption-free (AAF) method [Ewens & Shute 1986] suffers neither shortcoming, but requires the identification of potential probands; correcting for additional probands will not bias the parameter estimates, but may eliminate much of the information in the pedigree.

The standard theory of ascertainment correction applies to nuclear families ascertained through affected offspring [Elandt-Johnson 1971]. If π represents the ascertainment probability, $\pi \cong 0$ corresponds to single selection where the probability of ascertaining a family increases in proportion to the number of affected offspring and $\pi = 1$ corresponds to complete selection where

the probability of ascertaining a family is independent of the number of affected offspring. Applying the standard method requires the assumption of independent sampling of offspring and requires either knowing or estimating π.

As an alternative for unknown π, Ewens & Shute [1986] and Shute & Ewens [1988a] proposed an ascertainment-assumption-free (AAF) method of ascertainment correction, which conditions the likelihood of the sample on the subset of the data relevant to ascertainment. Ewens & Green [1988] extended the AAF method to quantitative phenotypes, but the threshold must be known. The AAF method produces unbiased parameter estimates, but larger standard errors than are produced by the appropriate standard ascertainment correction. However, application of the inappropriate ascertainment correction produces biased parameter estimates.

The Thompson & Cannings [1980] (TC) method assumes single selection and conditions the sample likelihood on the probability of the observation that prompted study of the family. For a nuclear family ascertained through an affected offspring, the ascertainment correction using the TC method equals the probability of one affected child with the parents' phenotypes unknown. In contrast, the ascertainment correction using AAF method equals the probability of the observed numbers of affected and unaffected children with the parents' phenotypes unknown.

Both the TC and AAF [Shute & Ewens 1988b] methods extend to pedigrees. The TC correction again equals the probability of the proband. To apply the AAF method, potential probands must be determined. If only one nuclear family was available for ascertainment, the AAF correction would be the same as for nuclear families. If offspring in more than one family are available, the conservative correction equals the probability of the observed numbers of affected and unaffected offspring in all the related sibships with the connecting adults classified as unknown.

The ascertainment correction in PAP divides the pedigree likelihood by the likelihood on the ascertainment subset. Either the TC or AAF correction may be used by appropriate selection of the individuals and phenotypes entered into *ascer.dat* (section II.4).

VI.5. Use of Logarithms

Calculation in logarithms within PAP eliminates the need for scaling to avoid underflow. Therefore, PAP performs each multiplication as an addition and performs each addition in subroutine *papsm*. As the logarithm of 0 ($-\infty$), a machine-dependent constant ZMIN equals a large negative number; when testing for 0, PAP uses ZMAX (10 times ZMIN). The genotype frequencies, genotype transmission probabilities, and penetrance probabilities are stored as natural logarithms when computed.

Subroutine *papsm* returns the logarithm of the sum of two numbers e^A and e^B. If the difference A - B exceeds the number of significant natural logarithms digits (SIGE), *papsm*

returns the larger of A and B as the sum. Otherwise, *papsm* returns B + log (e^{A-B} + 1.0) as the sum.

VI.6. Likelihood Calculation

PAP uses double precision for all computation.

PAP computes likelihoods in one of two ways. When linked with one of the exact variance components mixed model subroutines (*papende/papcrde*, *papenqe/papcrqe*, *papendqe/papcrdqe*), *pappd* sums over all sets of genotypes for the pedigree members. When linked with any other penetrance subroutine, *papnf* and *papcl* compute the likelihood one nuclear family at a time in the order specified in *papin.dat*.

Subroutines *pappd* and *papnf* cycle through all possible combinations of genotypes for a set of pedigree members. For *pappd* the set includes the complete pedigree; for *papnf* the set includes only nuclear family members who require joint likelihoods (are in the cutset, have a joint likelihood stored on them, or are parents in this nuclear family). When *papnf* completes all combinations, it stores the values computed for cutset members to be retrieved when required for another nuclear family in the pedigree.

Subroutine *papnf* calls *papcl* for each genotype combination to store genotype transmission probabilities for any offspring in the family who require joint likelihoods and to compute likelihoods for any others. Since offspring in a nuclear family are independent, conditional on parental genotypes, *papcl* computes the likelihood for each offspring as a sum over all genotypes, and the likelihood of all offspring as the product of these terms. PRLST (previous likelihood storage) stores log likelihoods when computed; storage takes place in subroutines *papnf* during computation. IGENLK (genotype likelihood) stores the corresponding cutset member's genotype or code number for a cutset larger than one. IDIREC (previous likelihood directory) contains pointers into PRLST for the stored likelihoods. PRLST is a one-dimensional array. The first NGENTL locations in PRLST contain genotype frequencies (stored as logarithms) for use equivalently to stored likelihoods for founders. The actual stored likelihoods start in location NGENTL + 1. Each set of individuals requires at most $NGENTL^{NCS}$ likelihoods where NGENTL represents the number of genotypes and NCS represents the number in the cutset; fewer spaces are required if some likelihoods equal zero. For a cutset of size two, the code number equals (IGEN1 - 1) * NGENTL + IGEN2 where IGEN1 and IGEN2 represent the genotypes of the first and second cutset members, respectively. For a cutset of size 3, the code number equals ((IGEN1 - 1) * NGENTL + IGEN2 - 1) * NGENTL + IGEN3 where IGEN1, IGEN2, and IGEN3 represent the genotypes of the first, second, and third cutset members, respectively. This code number is stored in the same location in IGENLK as the corresponding log likelihood is stored in PRLST.

Subroutine *pappl* computes the likelihood of a pedigree by calling *pappd* for exact computation or looping through the nuclear families and calling *papnf*.

VI.7. Genotype Probabilities

Genotype probabilities are computed by looping through all genotypes, and for each genotype computing the likelihood after setting the specified individual's penetrance probabilities to zero for all except the designated genotype. Probabilities may be computed for single locus or multi-locus genotypes.

Subroutine *pappr* calls *pappl* (section VI.6) to compute each of the NG likelihoods, where NG is the number of genotypes for which probabilities are computed. When all are computed, *pappr* computes the probability of each genotype by dividing each likelihood by the sum across all genotypes.

VI.8. Penetrance Probabilities

The penetrance probability for individual INO and genotype IGEN is stored in PENTR(INO, IGEN) as a natural logarithm.

Subroutine *pappn* first stores the genotype frequencies for founders and zero (logarithm of one) for nonfounders in PENTR. Second, for X-linked or category-specific autosomal inheritance, *pappn* stores ZMIN (logarithm of zero) for genotypes of the other gender or category. Third, if the model includes markers, *pappn* stores ZMIN (logarithm of zero) for impossible genotypes and sums the frequencies of combined genotypes (section VI.3) in founders. Fourth, for discrete traits, *pappn* stores ZMIN when the appropriate penetrance equals zero, or, for quantitative traits, *pappn* transforms and computes the deviation. Finally, *pappn* calls *papwg* (section VII.5) and *papen* to compute penetrances for the traits in the model. Subroutine *papcr* is called by *pappd*, *papnf*, and *papcl* (section VI.6) to correct the penetrance conditional on genotypes.

Subroutine *papen* comes in ten forms: one for markers: *papen*; three for major locus inheritance: *papend*, *papenq*, *papendq*; three for exact mixed model likelihood computation: *papende*, *papenqe*, *papendqe*; and three for approximate mixed model likelihood computation: *papenda*, *papenqa*, *papendqa*. Subroutines *papend*, *papende*, and *papenda* are for discrete traits; subroutines *papenq*, *papenqe*, and *papenqa* are for quantitative traits; and subroutines *papendq*, *papendqe*, and *papendqa* are used when both discrete and quantitative traits are included in the model. The major locus and approximate mixed version of *papen* perform the conditioning step (section VI.12). The approximate mixed version of *papen* computes the standard deviation correction for cutset members.

Subroutine *papcr* comes in seven forms: *papcr* is used with *papen*, *papend*, *papenq*, and *papendq*; *papcrde*, *papcrqe*, *papcrdqe* are used with *papende*, *papenqe*, and *papendqe*, respectively; *papcrda*, *papcrqa*, *papcrdqa* are used with *papenda*, *papenqa*, and *papendqa*, respectively. The exact mixed model versions of *papcr* perform the conditioning step (section

VI.12) and compute the penetrance. The approximate mixed model version of *papcr* corrects the conditioning step for genotypes of family members.

VI.9. Transmission Disequilibrium Test

Let τ_j represent the probability that a parent heterozygous for marker allele j and any other allele transmits allele j to an affected offspring for j = 1,...,m alleles. Under the null hypothesis of no linkage between the disease and marker, $\tau_j = 0.5$ for all j. If q_j represents the population frequency of allele j with

$$\sum_{j=1}^{m} q_j = 1$$

and assuming Hardy-Weinberg equilibrium, then parameters τ_j are subject to the constraint:

$$\sum_{j=1}^{m} q_j(1-q_j)\tau_j = \sum_{j=1}^{m} q_j(1-q_j)(1-\tau_j) \qquad (1)$$

Equation (1) effectively constrains to equality the numbers of transmitted and non-transmitted alleles. For m = 2, equation (1) specifies that $\tau_1 = 1 - \tau_2$.

Although the τ_j, the transmission probabilities for parental-allele/offspring-allele pairs, are the relevant parameters of the model, in order to compute the likelihood we need to specify the transmission probability for each parental-genotype/offspring-allele pair. That is, we need t_{jk} = probability that heterozygote jk transmits allele j, where $t_{kj} = 1 - t_{jk}$, such that

$$\tau_j = \sum_{k=1}^{m} q_k t_{jk}. \qquad (2)$$

For m = 2, $t_{12} = \tau_1$. For m > 2, the t_{jk} are underdetermined by the τ_j; one solution to equation (2) is

$$t_{jk} = \frac{1}{2} + \frac{(\tau_j - \tau_k)(1-q_j)(1-q_k)}{\sum_{i=1}^{m} q_i(1-q_i)}. \qquad (3)$$

Then the transmission probability tr(im|ij,mn), the probability that parents with genotypes ij and mn produce an offspring with genotype im, equals:

$$\text{tr(im|ij,mn)} = \begin{cases} 2t_{ij}t_{mn} & \text{if } i = n,\ j = m,\ i \neq j \\ t_{ij}t_{mn} & \text{if } i \neq j,\ m \neq n \\ t_{ij} & \text{if } i \neq j,\ m = n \\ t_{mn} & \text{if } i = j,\ m \neq n \\ 1 & \text{if } i=j,\ m=n \end{cases}$$

where t_{jk} is defined as in equation (3) for affected offspring and t_{jk} = ½ for unaffected offspring.

The transmission probability comprises one component of the likelihood of a genetic model [Elston and Stewart, 1971]; the complete likelihood of a genetic model with G genotypes on a pedigree of N individuals is proportional to

$$\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\cdots\sum_{g_N=1}^{G}\prod_{n=1}^{N} pn(x_n \mid g_n)\prod_{i=1}^{F} fr(g_i)\prod_{j=F+1}^{N} tr(g_j \mid g_{jf}, g_{jm}) \qquad (4)$$

where the F founders are assumed, without loss of generality, to be ordered first, g now designates the genotype and jf and jm designate the parents of j. Dropping the subscript which designates the individual for ease of notation, the penetrance probability pn(x|g), the probability that phenotype x is expressed, conditional on genotype g, in this case equals:

$$\text{pn(x|g)} = \begin{cases} 0 \text{ if genotype g is not consistent with marker phenotype x} \\ 1 \text{ if genotype g is consistent with marker phenotype x} \\ 1 \text{ if x = unknown} \end{cases}$$

The genotype frequencies, fr(g), the probability of the random occurrence in the population of genotype g, in this case equals the population frequency of marker genotype g. Hardy-Weinberg equilibrium is assumed.

VI.10. Marker Simulation

Program *simrk* uses the chromosome-based simulation method of Terwilliger et al [1993] to simulate marker genotypes. Pedigree structure is fixed. The distances between markers may be entered either in cM or recombination probabilities. The Sturt [1976] map function is assumed. This model assumes at least one crossover per chromosome.

Program *simrk* requests the length of the chromosome and location of the first marker. It loops through the pedigrees and call *papsmg* to simulate genotypes for each pedigree.

Subroutine *papsmg* first assigns genotypes to founders. For each locus, a random number is selected twice in order to assign an allele to each chromosome according to the allele frequencies. Upon completion of the assignment of alleles along each chromosome in founders, alleles are assigned to non-founders, ordering the assignments such that alleles in both parents have already been assigned. For each chromosome, the number of crossovers is randomly selected according to a Poisson distribution with parameter the chromosome length. Then crossover locations are randomly selected according to a uniform distribution. This determines the alleles along each chromosome from parental chromosomes.

VI.11. Phenotype Simulation

All simulation options assume a fixed pedigree structure. For each trait or marker included in the model, each pedigree member can be assigned a simulated phenotype, can retain an existing phenotype, or can be assigned as missing. If any phenotypes for a trait or marker are to be retained and others are missing, the simulation code must be used in *phen.dat* (section II.4) to designate which phenotypes to simulate. Otherwise, the simulation code need not be used. The options, for each trait or marker, for designating phenotypes to be simulated include: **(1)** all pedigree members, **(2)** all except pedigree members assigned a missing value, **(3)** all except pedigree members assigned a missing value for a different trait or marker, **(4)** all pedigree members assigned the simulation code. If you use simulation to determine the number of nuclear families needed to detect linkage to a recessive trait, choose option **(1)** assuming all family members are studied. If you use simulation to obtain parameter estimates or a likelihood distribution to compare to the results on an existing data set, choose option **(2),** which would simulate a phenotype for each studied pedigree member. If you use simulation to estimate expected lod scores, choose option **(3)** and simulate a marker phenotype for each pedigree member assigned known disease status. If you use simulation to decide whether to type additional pedigree members, choose option **(4)** and use the known marker phenotypes for typed individuals and the simulation code for untyped individuals with known disease status. If you want to simulate phenotypes ascertained through a proband, choose option **(4)** and assign the disease designation to the proband and use the simulation code for other pedigree members.

Genotypes may be simulated conditionally or unconditionally. Conditional genotype simulation is necessary if any phenotypes are retained, as for expected lod score analysis where the marker is simulated but disease phenotypes are retained, or in disease pedigrees if one individual (a proband) is designated affected and phenotypes of other family members are simulated.

For unconditional simulation, a random number is generated for each founder and a multi-locus genotype is assigned in accord with the population genotype frequencies. Then a random number is generated for each nonfounder and a multi-locus genotype is assigned in accord with the transmission probabilities specific for the genotypes of the parental pair.

For conditional simulation, a random number is generated, genotype probabilities are computed, and a multi-locus genotype is assigned in accord with the genotype probabilities for each individual.

To assign a marker phenotype, the one-locus genotype is extracted from the multi-locus genotype. If available, the phenotype/genotype relationships are obtained from *popln.dat* and the phenotype corresponding to the simulated genotype is stored. Otherwise, the genotype itself is stored as the phenotype.

To assign a trait phenotype, a normal deviate is randomly selected for each pedigree member, these deviations are reverse conditioned (section VI.13) using the conditioned correlation matrix, and an appropriate phenotype is assigned by *dmlpr* (discrete trait) or *qmlpr* (quantitative trait).

To simulate phenotypes in pedigrees ascertained through a specified proband, you fix the proband's phenotype in *ascer.dat* and *phen.dat* (section II.4).

Subroutine *simph* designates traits to be simulated in ISIMT(INO, ITRAIT) and the genotypes to be simulated in ISIMG(INO, IMARKR) for individual INO. Subroutine *simph* also determines if the simulation must be conditional.

Subroutine *papsg* performs either unconditional or conditional simulation of genotypes, then assigns phenotypes.

VI.12. Conditioning Process

For a disease, we assume affection occurs upon exceeding a specified threshold T on a normally-distributed liability curve. Consequently, the likelihood of a polygenic model equals an N-variate normal integral, where N represents the number of individuals and integral i ranges from T to ∞ for i affected or from -∞ to T for i unaffected. Pearson [1903], Mendell and Elston [1974], and Rice et al. [1979] approximated the N-variate normal integral with the product of N univariate normal integrals. That is, to approximate the N-variate integral, repeat the following steps 1 to 3 for i = 1, 2,..., N: (1) Approximate the univariate normal integral for individual i; (2) Condition means, variances, and correlations for individuals i + 1 through N on the truncated normal distribution of individual i; (3) Assume the remaining (N-i)-variate density distributes normally. Approximation occurs at steps 1 and 3.

Now consider instead a polychotomous phenotype. The likelihood again equals an N-variate normal integral, but now integral i ranges from S_i to T_i, $-\infty \leq S_i \leq T_i \leq \infty$. The doubly truncated normal density for individual i has mean

(1)

$$\frac{\phi(S_i) - \phi(T_i)}{\Phi(T_i) - \Phi(S_i)} = a_i$$

and variance

(2)

$$V_i = 1 + \frac{S_i\phi(S_i) - T_i\phi(T_i)}{\Phi(T_i) - \Phi(S_i)} - a_i^2 = 1 - v_i^2$$

where φ(.) represents the normal density and Φ(.) represents the normal distribution. Let r_{ij} represent the correlation between variates i and j, which may reflect the variance components model or familial correlations (section IV.8). Conditioning on individual i modifies the integral range for individual j to

(3)

$$S_i' = \frac{S_i - r_{ij}a_i}{\sqrt{1 - r_{ij}^2 v_i^2}}$$

and (4)

$$T_i' = \frac{T_i - r_{ij}a_i}{\sqrt{1 - r_{ij}^2 v_i^2}}$$

and modifies the variance for individual j to

(5)

$$V_i' = \frac{V_i}{1 - r_{ij}^2 v_i^2}$$

and modifies the correlation between individuals j and k to

(6)

$$r_{jk}' = \frac{r_{jk} - r_{ij}r_{ik}v_i^2}{\sqrt{1 - r_{ij}^2 v_i^2}\sqrt{1 - r_{ik}^2 v_i^2}}$$

where the prime indicates a conditioning step which, for ease of notation, will not be specified explicitly. Therefore, to approximate the N-variate integral, repeat steps 1 through 2 for i = 1, 2,...,N: (1) Approximate the probability for individual i as $\Phi(T_i) - \Phi(S_i)$; (2) Compute a_j', v_j', S_j', T_j', r_{jk}' for j = i + 1, ..., N, k = j + 1, ..., N using equations 1-6.

For a dichotomous trait, S_i = threshold and $T_i = \infty$ for i affected and $S_i = -\infty$ and T_i = threshold for i unaffected. For a standardized quantitative phenotype x_i, $S_i = T_i = x_i$; then $a_i = x_i$, $v_i = 1$, and $\varphi(x_i)$ substitutes for $\Phi(T_i) - \Phi(S_i)$ in the likelihood computation. For a multivariate model the variates may be different traits; each correlation depends on the identity of the pair of variates; the traits may be a mixture to data types.

Subroutines *papend*, *papenq*, *papendq*, *papenda*, *papenqa*, *papendqa*, *papcrde*, *papcrqe*, and *papcrdqe* (section VI.8) perform the conditioning step.

VI.13. Reverse Conditioning

The conditioning process converts a multivariate normal density into the product of univariate normal densities. Therefore, multivariate normal phenotypes with the desired correlation structure can be simulated by first simulating independent normal deviates and applying the conditioning process in reverse.

The correlations needed for the reverse conditioning process are obtained by first applying the forward conditioning process (section VI.12) used for a quantitative trait. Conditioning modifies the correlation between individuals j and k to

$$r_{jk}' = \frac{r_{jk} - r_{ij} r_{ik}}{\sqrt{1 - r_{ij}^2}\sqrt{1 - r_{ik}^2}}$$

After conditioning on all previous individuals, correlation r_{jk}' contain the effects on the variance and correlation induced by the correlation structure assumed for the variates. To simulate multivariate normally distributed data, now simulate independent normal deviates and reverse the conditioning process to produce data that reflect the desired correlation structure. A random normally distributed deviate x_i will be selected for each variate. Then starting with variate N, reverse conditioning on variate i modifies the deviation for variate j to

$$x_j = x_j' \sqrt{1 - r_{ij}^2} + r_{ij} x_i'$$

The phenotypes x_j produced from this process will be multivariate normally distributed with the specified correlation structure. Each deviate can be used in conjunction with the simulated genotype and the parameters of the model to produce a quantitative phenotype, an affection status, or a disease severity code.

VI.14. Transformation of Quantitative Phenotypes

Quantitative phenotypes may be transformed using the function $y = r/P[(x/r+1)^P-1]$ [Maclean et al 1976], where x represents the phenotype, P represents the power and r = 6. Set P = 1 for no transformation. For $P \neq 1$, the variable must be standardized. Estimating P within a 1-genotype model allows you to determine the transformation to correspond to a single normal density. Estimating P within the environmental model (section IV.2.2) allows you to test for a mixture of distributions. If you then insert the estimate of P into *header.dat* (section II.3) and run *preped* (section III.1), further analyses will consider the transformed phenotypes. Estimating P within a genetic model allows transformation of the phenotypes to obtain the best fit to each model.

VII. WRITING CUSTOM SUBROUTINES

Appendix C lists the frequency *(papfq*), transmission *(paptc*), major locus discrete *(dmlpr*), major locus quantitative *(qmlpr*), and within genotype *(papwg*) subroutines included in PAP. If you want to modify the parameterization or add parameters, you can write a subroutine within specified constraints. I recommend starting with a similar subroutine and making the required modifications; reparameterization usually requires only a few lines of code added at the beginning. For example, to estimate the variance instead of the standard deviation in *qmlprmv* replace VARWG = PNPAR(2, 1, ITRAIT) ** 2 with VARWG = PNPAR(2, 1, ITRAIT).

Carefully test any modified subroutine. If the modification represents a reparameterization, compare likelihoods using the modified subroutine to likelihoods using the original subroutine for corresponding parameter values. If the modification adds parameters, perform two sets of tests. First, restrict the added parameters of the modified subroutine and compare the likelihood to a corresponding likelihood using the original subroutine. Second, simplify the data and, for a nontrivial value of the added parameter, compare a hand-calculated likelihood to the likelihood using the modified subroutine.

VII.1. Frequency: *papfq*

Subroutine *papfq* stores the genotype frequencies as logarithms in the first NGENTL elements of array PRLST. If necessary, *papfq* first computes the genotype frequencies and ensures a sum of 1. Array FREQ contains parameter values for computing the genotype frequencies.

After producing a modified subroutine, name it *papfq* with 1-3 added letters and add the relevant information to *freq.dat* (section B.8). Assign the number associated with the subroutine in *freq.dat* in a DATA statement in a BLOCK DATA SUBROUTINE appended to *papfq*.

See subroutine *papfqg* for an example of the minimum structure of *papfq*. The parameters in FREQ are genotype frequencies and *papfqg* stores them in array PRLST after ensuring a sum of 1. A data statement assigns the number 3 to *papfqg* corresponding to the assignment made in *freq.dat*.

VII.2. Transmission: *paptc*

Subroutine *paptc*, for a specified pair of parental genotypes, stores each nonzero genotype transmission probability as a logarithm in array TRPRB, stores the corresponding genotype in array IGENTP, and stores the number of nonzero values in NGENTP. Arrays TRNDAU, TRNSON, RCMFTH, and RCMMTH may contain parameter values for computing the genotype transmission probabilities.

After producing a modified subroutine, name it *paptc* with 1-3 added letters and add the relevant information to *tran.dat* (section B.9). Assign the number associated with the subroutine in *tran.dat* in a DATA statement in a BLOCK DATA SUBROUTINE appended to *paptc*.

See subroutine *paptce* for an example of the minimum structure of *paptc*. Subroutine *paptce* ignores the parental genotypes that enter as arguments and assigns the frequencies to TRPRB. A data statement assigns the number 5 to *paptce* corresponding to the assignment made in *tran.dat*.

VII.3. Major Locus Discrete: *dmlpr*

Subroutine *dmlpr* performs three functions, numbered 1, 2, 4. **(1)** Assign each member INO of the sample to a category for trait IT in array ICATEG(INO, IT). **(2)** Store the genotype- and category-specific affection probability (in PEN and as a logarithm in PRBPHN) along with the corresponding deviations for the lower (DEVL) and upper (DEVU) thresholds. **(4)** Assign an appropriate simulated phenotype using the simulated genotype and simulated deviation.

Name the modified subroutine *dmlpr* with 1-3 added letters and add the relevant information to *dmlp.dat* (section B.10). Assign the number associated with the subroutine in *dmlp.dat* in a DATA statement in a BLOCK DATA SUBROUTINE appended to *dmlpr*.

See subroutine *dmlprpn* for an example of the minimum structure of *dmlpr*. Subroutine *dmlprpn* **(1)** assigns affection status to ICATEG, **(2)** stores the penetrance probability input as PARMD as PEN and PRBPHN, finding the corresponding deviations using function *devia*, **(4)** assigns phenotype 1 to individuals whose simulated deviation from their genotype mean falls below the threshold and assigns phenotype 2 to individuals whose simulated deviation from their genotype mean falls above the threshold.

VII.4. Major Locus Quantitative: *qmlpr*

Subroutine *qmlpr* performs four functions. **(1)** Assign each measured member INO of the sample as category -1 for trait IT in array ICATEG(INO, IT). **(2)** Store the genotype- and category-specific means (in GMEAN) and variance (in VARWG). **(3)** Assign an appropriate simulated phenotype using the simulated genotype and simulated deviation.

Name the modified subroutine *qmlpr* with 1-3 added letters and add the relevant information to *qmlp.dat* (section B.8). Assign the number associated with the subroutine in *qmlp.dat* in a DATA statement in a BLOCK DATA SUBROUTINE appended to *qmlpr*.

See subroutine *qmlprmv* for an example of the minimum structure of *qmlpr*. Subroutine *qmlprmv* **(1)** assigns to ICATEG, -1 if measured and 0 if unmeasured, **(2)** stores the means and standard deviations which were input as PARMQ as GMEAN and VARWG, **(3)** applies the

mean, variance, covariates and transformation parameters for the simulated genotype to the simulated deviation to produce a simulated phenotype.

VII.5. Within Genotype: *papwg*

Subroutine *papwg* has two purposes. First, transfers parameter from PARWG to the correlation array RHO. Second, it assigns a correlation to each pair in the pedigree in array RWG.

Name the modified subroutine *papwg* with 1-3 added letters and add the relevant information to *wgen.dat* (section B.9). Assign the number associated with the subroutine in *wgen.dat* in a DATA statement in a BLOCK DATA SUBROUTINE appended to papwg.

See subroutine papwgml for an example of the minimum structure of papwg. It stores values from PARWG in RHO as the correlations between traits. Then it generates the pedigree correlation matrix using those correlations.

VIII. ACKNOWLEDGMENTS

Version 1 of PAP resulted from experience at the University of Utah in implementing versions of the pedigree analysis programs written by Elizabeth Thompson [1976, 1977]. A separate version evolved for each problem encountered (linkage, cancer penetrance functions, calculation of genotype probabilities). PAP represented an attempt to include many applications in one program as well as to minimize user errors. Version 1 derived directly from programs for zero-loop and arbitrarily complex pedigrees by Thompson [1976, 1977] and a zero-loop version for linkage by Kravitz [1979]. The terminology followed Cannings et al [1978] and PAP computed joint likelihoods as in Cannings et al [1976b].

Many people participated in the development of Version 1. Mark Skolnick originated the idea and promoted its implementation. Tim Bishop originally incorporated GEMINI into the likelihood calculation structure and wrote the initial genotype probability subroutine. Jean-Marc Lalouel assisted with the development of the original polygenic model and major locus/polygenic mixed models that used polychotomies. Kerry Kravitz, Chris Cannings, Elizabeth Thompson, Jon Hill, Dorit Carmelli and others contributed to discussions on scaling, efficient memory utilization and other problems.

Version 2 of PAP resulted from the need to make the source code more portable to facilitate its installation by other users. Peter Cartwright assisted in modifying the source code.

Version 3 of PAP responded to my own and other users' desires for a larger repertoire of genetic models. Feedback from many PAP users (Pat Moll, Patti Kramer, Candace Kammerer, and Jean MacCluer among others) provided invaluable suggestions which made Version 3 more versatile and easier to use. In particular, Pat Moll continually contributed advice and encouragement; she also computed likelihoods of the variance components models using her program for testing the corresponding models in PAP. Peter Cartwright again assisted with technical aspects of the distribution of PAP.

Version 4 of PAP incorporated the changes and additions that accumulated over the intervening four and a half years since Version 3 was released. Andy Marks assisted with technical aspects of the distribution of PAP.

Version 5 of PAP incorporates the changes and additions made since Version 4 was released.

Versions 1 and 2 were supported by NIH grants T32 AM-07115, CA-16573, GM-27192 and HL-21088. Versions 3, 4, and 5 were supported by NIH grant HD-17463.

IX. REFERENCES

Boerwinkle E, Chakraborty R, Sing CF (1986) The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 50:181-194.

Bovill EG, Hasstedt SJ, Callas PW, Valliere JE, Scott BT, Bauer KA, Long GL (2000) The G20210A prothrombin polymorphism is not associated with increased thromboembolic risk in a large protein C deficient family. Thromb Haemost 83:366-370.

Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC (2001) Ascertainment adjustment: Where does it take us? Am J Hum Genet 67:1505-1514.

Cannings C, Thompson EA, Skolnick M (1976b) The recursive derivation of likelihoods on complex pedigrees. Adv Appl Prob 8:622-625.

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Adv Appl Prob 10:26-61.

Cannings C, Thompson EA, Skolnick MH (1980) Pedigree analysis of complex models. In Current Developments in Anthropological Genetics, Vol. 1: Theory and Methods, Mielke JH, Crawford MH (eds). Pp 251-298.

Demenais FM, Elston RC (1981) A general transmission probability model for pedigree data. Hum Hered 31:93-99.

Elandt-Johnson RC (1971) Probability models and statistical methods in genetics. Wiley, New York

Elbein SC, Hoffman MD, Yount PA, Teng K, Fuller D, Barrett K, Leppert MF, Hasstedt SJ (1999) A genome-wide search for type 2 diabetes susceptibility genes in Utah caucasians. Diabetes 48:1175-1182.

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523-542.

Ewens WJ, Shute NCE (1986) A resolution of the ascertainment sampling problem. I. Theory. Theor Popul Biol 30:388-412

Ewens WJ, Green RM (1988) A resolution of the ascertainment sampling problem. IV. Continuous phenotypes. Genet Epidemiol 5:433-444

Gill PE, Murray W, Saunders MA, Wright MH (1986) NPSOL: A Fortran package for nonlinear programming. Technical Report SOL 86-2, Stanford University, Stanford, CA.

Hasstedt SJ (1982) A mixed model likelihood approximation for large pedigrees. Comp Biomed Res 15:295-307.

Hasstedt SJ, Skolnick M (1984) A general autosomal/X-linked model. Genet Epidemiol 1:21-36.

Hasstedt SJ, Atkin CL, San Juan AC Jr (1986) Genetic heterogeneity among kindreds with Alport syndrome. Am J Hum Genet 38:940-953.

Hasstedt SJ, Moll PP (1989) Estimation of genetic model parameters: Variables correlated with a quantitative phenotype exhibiting major locus inheritance. Genet Epidemiol 6:319-332.

Hasstedt SJ (1991) A variance components/major locus likelihood approximation on quantitative data. Genet Epidemiol 8:113-125.

Hasstedt SJ (1993) A variance components/major locus likelihood approximation for quantitative, polychotomous, and multivariate data. Genet Epidemiol 10:145-158.

Hasstedt SJ (1995) Phenotypic assortative mating in segregation analysis. Genet Epidemiol 12:109-127.

Hasstedt SJ, Leppert M, Filloux F, van de Wetering BJM, McMahon WM (1995) Intermediate inheritance of Tourette syndrome assuming assortative mating. Am J Hum Genet 57:682-689.

Hasstedt SJ, Hoffman M, Leppert MF, Elbein SC (1997) Recessive inheritance of obesity in familial NIDDM and lack of linkage to nine candidate genes. Am J Hum Genet 61:668-677.

Hasstedt SJ, Bovill EG, Callas PW, Long GL (1998) An unknown genetic defect increases venous thrombosis risk, through interaction with protein C deficiency. Am J Hum Genet 63:569-576.

Kravitz KD (1979) A computer program for pedigree analysis of quantitative variables. Dept of Medical Biophysics and Computing, University of Utah, Technical Report No. 9.

Lalouel JM (1979) GEMINI -- a computer program for optimization of a nonlinear function. Dept of Medical Biophysics and Computing, University of Utah, Technical Report No. 14.

MacLean CJ, Morton NE, Elston RC, Yee S (1976) Skewness in commingled distributions. Biometrics 32:695-699.

Morton NE, MacLean CJ (1974) Analysis of family resemblance. III. Complex segregation of quantitative traits. Am J Hum Genet 26:489-503.

Mendell NR, Elston RC (1974) Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. Biometrics 30:41-57.

Pearson K (1903) On the influence of natural selection on the variability and correlation of organs. Philos Trans R Soc Lond [A] 200:1-66.

Rice J, Reich T, Cloninger CR (1979) An approximation to the multivariate normal integral: Its application to multifactorial qualitative traits. Biometrics 35:451-459.

Self SG, Liang KY (1987) Large sample properties of the maximum likelihood estimator and the likelihood ratio test on the boundary of the parameter space. J Am Stat Assoc 82:605-610.

Shute NCE, Ewens WJ (1988a) A resolution of the ascertainment sampling problem. II. Generalization and numerical results. Am J Hum Genet 43:374-386

Shute NCE, Ewens WJ (1988b) A resolution of the ascertainment sampling problem. II. Pedigrees. Am J Hum Genet 43:387-395

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983-989

Sturt E (1976) A mapping function for human chromosomes. Ann Hum Genet 40:147-163

Terwilliger JD, Speer M, Ott J (1993) Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 10:217-224

Thompson EA (1976) Peeling programs for zero-loop pedigrees. Dept of Medical Biophysics and Computing, University of Utah, Technical Report No. 14.

Thompson EA (1977) Peeling programs for pedigrees of arbitrary complexity. Dept of Medical Biophysics and Computing, University of Utah, Technical Report No. 6.

Thompson EA, Cannings C (1980) Sampling schemes and ascertainment. In CF Sing, M Skolnick (eds) The genetic analysis of common diseases: Applications to Predictive Factors in Coronary Heart Disease. Alan Liss, New York.

A. FILE LISTS

A.1. Command Files

Makefile	Compiles and links all programs
test.sh	Runs test data using GEMINI
pap.sh	Called by test.sh to execute *papdr*

A.2. Include Files

c1.inc	Parameters and commons used by *preped*
c1234567.inc	Parameters and commons used by *preped, descstat, prepap, simrk, simul, papdr, gpe*
c12456.inc	Parameters and commons used by *preped, descstat, simrk, simul, papdr*
c124567.inc	Parameters and commons used by *preped, descstat, simrk, simul, papdr, gpe*
c134567.inc	Parameters and commons used by *preped, prepap, simrk, simul, papdr, gpe*
c1356.inc	Parameters and commons used by *preped, prepap, simul, papdr*
c1456.inc	Parameters and commons used by *preped, simrk, simul, papdr*
c1567.inc	Parameters and commons used by *preped, simul, papdr, gpe*
c16.inc	Parameters and commons used by *preped, papdr*
c17.inc	Parameters and commons used by *preped, gpe*
c23456.inc	Parameters and commons used by *descstat, prepap, simrk, simul, papdr*
c2456.inc	Parameters and commons used by *descstat, simrk, simul, papdr*
c3.inc	Parameters and commons used by *prepap*
c3456.inc	Parameters and commons used by *prepap, simrk, simul, papdr*
c356.inc	Parameters and commons used by *prepap, simul, papdr*
c4.inc	Parameters and commons used by *simrk*
c56.inc	Parameters and commons used by *simul, papdr*
c56t.inc	Parameters and commons used by *simul, papdr* for traits
c6.inc	Parameters and commons used by *papdr*
c67.inc	Parameters and commons used by *papdr, gpe*
c6t.inc	Parameters and commons used by *papdr* for traits
const.inc	Constants used in *simul, papdr*
gemini.inc	Parameters and commons used by *gemini*
i1.inc	Include files for *preped*
i1237.inc	Include files for *preped, descstat, prepap, gpe*
i134567.inc	Include files for *preped, prepap, simrk, simul, papdr, gpe*
i17.inc	Include files for *preped, gpe*
i2.inc	Include files for *descstat*
i2456.inc	Include files for *descstat, simrk, simul, papdr*
i3.inc	Include files for *prepap*
i356.inc	Include files for *prepap, simul, papdr*
i36.inc	Include files for *prepap, papdr*
i4.inc	Include files for *simrk*
i5.inc	Include files for *simul*
i56.inc	Include files for *simul, papdr*
i56d.inc	Include files for *simul, papdr* for discrete traits
i56t.inc	Include files for *simul, papdr* for traits
i6.inc	Include files for *papdr*
i6t.inc	Include files for *papdr* for traits
i7.inc	Include files for *gpe*
papch.inc	I/O logical unit numbers

A.3. Source Files

alloff.f	Determines alleles in descendants of an individual untyped for a marker
assort.f	Iterates to find equilibrium genotype and mating frequencies
confun.f	Defines the nonlinear constraints for NPSOL
daterr.f	Error messages for *preped, descstat, simul, papdr, gpe*
datin.f	Reads in data files for *simul* and *papdr*
descstat.f	Computes descriptive statistics from *papin.dat*
devia.f	Locates the normal deviation corresponding to a specified probability
diff.f	Computes the derivative of a parameter
dmlpr0.f	Dummy subroutine used when no discrete traits are included
dmlprin.f	Converts incidence figures to normal deviates and assigns age categories
dmlprinc.f	Converts incidence figures to normal deviates and assigns age categories allowing 2nd category class
dmlprpn.f	Converts penetrances to normal deviates
dmlprpr.f	Converts prevalence figures to normal deviates and assigns categories
dmlprsv.f	Converts prevalence for severity designations to normal deviates
domdis.f	Translates dominance and displacement into genotype means and a common standard deviation
drandi.f	Random number generator
filerr.f	Prints error messages for *preped, descstat, prepap, simul, papdr, gpe*
genot.f	Outputs the alleles comprising a particular genotype
gpe.f	Estimates frequencies, genotype means using the genotype probabilities
inmin.f	Interactive input of an integer, checking bounds
integ.f	Computes the normal integral below a specified point
moderr.f	Prints error messages for *descstat, prepap, simul, papdr*
modin.f	Reads model information and parameter values from *model.dat*
modnew.f	Queries the user for input for *model.dat*
modot.f	Writes model information and parameter values to *model.dat*
mrkin.f	Reads model information for *simrk*
objfun.f	Returns log likelihood values to NPSOL
papcl.f	Calculates the likelihood of children in a nuclear family
papcr.f	Dummy subroutine used when no within genotype correlations
papcrda.f	For discrete traits, corrects approximation for genotype
papcrde.f	Penetrance for discrete traits and the exact mixed model
papcrdqa.f	For discrete or quantitative traits, corrects approximation for genotype
papcrdqe.f	Penetrance for discrete or quantitative traits and the exact mixed model
papcrqa.f	For quantitative traits, corrects approximation for genotype
papcrqe.f	Penetrance for quantitative traits and the exact mixed model
papcsg.f	Simulates genotypes conditionally and assigns phenotypes
papcsg0.f	Dummy subroutine used in *simul*
papdr.f	Driver of the analysis program
papen.f	Dummy subroutine used when no traits are included
papend.f	Penetrance for discrete traits and major locus model
papenda.f	Penetrance for discrete traits and the approximate mixed model
papende.f	Penetrance for discrete traits and exact mixed model
papendq.f	Penetrance for discrete/quantitative traits and major locus model
papendqa.f	Penetrance for discrete or quantitative traits and the approximate mixed model
papendqe.f	Penetrance for discrete/quantitative traits and exact mixed model
papenq.f	Penetrance for quantitative traits and major locus model
papenqa.f	Penetrance for quantitative traits and the approximate mixed model
papenqe.f	Penetrance for quantitative traits and exact mixed model
paperr.f	Prints error message for *simul* and *papdr*
papes.f	Simulates phenotypes and estimates parameters
papfqc.f	Genotype frequencies computed from conditional allele frequencies at some loci and unconditional allele frequencies at other loci
papfqd.f	Genotype frequencies computed for two autosomal loci allowing for linkage disequilibrium
papfqg.f	Genotype frequencies stored

papfqh.f	Genotype frequencies computed from haplotype frequencies assuming Hardy-Weinberg equilibrium
papfqhw.f	Genotype frequencies computed from allele frequencies assuming Hardy-Weinberg equilibrium and linkage equilibrium
papfqxa.f	Genotype frequencies computed for the autosomal/X-linked model
papgd.f	Computes likelihoods for a grid of parameter values
papgn.f	Stores genotype and alleles by locus and trait for the model
papgp.f	Genotype probability routine
paplk.f	Computes the likelihood on a pedigree or the sample
papmv.f	Determines possible genotypes and indistinguishable genotypes for each individual untyped for a marker
papmxg.f	Maximizes the likelihood using GEMINI
papmxn.f	Maximizes the likelihood using NPSOL
papnf.f	Computes the likelihood of one nuclear family
pappd.f	Computes the likelihood of a pedigree considering all genotype sets
pappg.f	Determines possible genotypes for pedigree members
pappl.f	Computes the likelihood of one pedigree
pappn.f	Computes the penetrance for each individual in a pedigree
pappo.f	Outputs the parameter values to the screen and *pap.out*
pappr.f	Computes the probability of each genotype for an individual
papro.f	Outputs the results from *papdr*
papse.f	Computes standard errors
papsf.f	Calls *papsg* to simulate phenotypes and writes the phenotypes in files
papsg.f	Simulates genotypes and assigns phenotypes
papsm.f	Adds two numbers, each stored as logs
papsmg.f	Simulates marker genotypes for one pedigree
papsr.f	Searches one parameter at a time for the best likelihood
paptcal.f	Computes transmission probabilities for autosomal linkage
paptcapl.f	Computes transmission probabilities for parent-specific autosomal linkage
paptcasl.f	Computes transmission probabilities for category-specific linkage
paptcaxl.f	Computes transmission probabilities for autosomal/X-linked admixture
paptce.f	Stores genotype frequencies as transmission probabilities
paptcet.f	Computes transmission probabilities from parameters assuming generational equilibrium
paptcms.f	Transmission probabilities for Mendelian segregation and no linkage
paptctdt.f	Transmission probabilities for the transmission disequilibrium test
paptctp.f	Computes transmission probabilities from parameters
paptcxl.f	Computes transmission probabilities for X-linked inheritance
paptcxpl.f	Computes transmission probabilities for parent-specific X-linked inheritance
paptr.f	Transfers parameter values from the maximization array into the appropriate names
papwg.f	Dummy subroutine used when no traits are included
papwgam.f	Produces the correlation matrix from variance components assuming assortative mating
papwgfc.f	Produces the correlation matrix from specified familial correlations
papwgml.f	Produces the correlation matrix assuming no within-genotype correlation
papwgvc.f	Produces the correlation matrix from variance components
parin.f	Queries the user for input for *model.dat*
parnew.f	Reads in parameter values and boundaries
pedgen.f	Assembles father, mother, offspring triples into pedigrees
pedin.f	Reads *papin.dat*
pedot.f	Outputs *papin.dat*
phenin.f	Reads *phen.dat*
popin.f	Reads *popln.dat*
prepap.f	Produces *model.dat* through interactive input
preped.f	Combines *trip.dat, order.dat, header.dat, phen.dat, ascer.dat* into *papin.dat*
pselec.f	Lists variables for the user's selection
qmlpr0.f	Dummy subroutine used when no quantitative traits are included
qmlprdd.f	Converts dominance and displacement to genotype means and variances

qmlprddp.f	Converts dominance and displacement to genotype means and variances and proportion for affecteds
qmlprmv.f	Stores the genotype means and standard deviation
qmlprmvt.f	Stores the genotype means and standard deviation and computes a deviation corresponding to the threshold
repeat.f	Calls *alloff* to make it recursive
rnmin.f	Interactive input of a double precision number, checking bounds
sbrlib.f	Reads the subroutine library files
setbnd.f	Determines the parameters being maximized
simerr.f	Errors from *simrk*, *simul*, and *papdr*
simph.f	Determines traits/markers and phenotypes to be simulated
simrk.f	Driver for chromosome-based genotype simulation program
simul.f	Driver for simulation program
step.f	Performs line search for GEMINI
stepfn.f	Computes likelihood at new values as part of search and maximization options
thresh.f	Locates threshold corresponding to prevalence for multiple genotypes
update.f	Updates matrix for GEMINI
yesno.f	Interactive input of yes/no response

A.4. Input Files

dmlp.dat	Major locus for discrete traits subroutine library
freq.dat	Frequency subroutine library
qmlp.dat	Major locus for quantitative traits subroutine library
tran.dat	Transmission subroutine library
wgen.dat	Within genotype subroutine library

A.5. Test Data Files

ascer.tst	Phenotype data for probands
header.tst1	Variable names and characteristics
header.tst2	Variable names and characteristics
header.tst3	Variable names and characteristics
header.tst4	Variable names and characteristics
pap1.in	Interactive input for option 1
pap2.in	Interactive input for option 2
pap3.in	Interactive input for option 3
pap4.in	Interactive input for option 4
pap5.in	Interactive input for option 5
pap6.in	Interactive input for option 6
pap7.in	Interactive input for option 7
pap8.in	Interactive input for option 8
phen.tst	Phenotype data
popln.tst	Population marker information
preped.in	Interactive input for *preped*
preped2.in	Interactive input for *preped*
trip.tst	Pedigree structure

A.6. Parameter Files

m_1_1	*papfqhw*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_1_2	*papfqhw*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 2, responses y, y
m_1_3_1	*papfqhw*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 3, responses 2, 1, 1
m_1_3_2	*papfqhw*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_1_4	*papfqhw*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 4, response n

m_1_5	*papfqhw*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 5, responses n, n, n
m_1_6	*papfqhw*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 6, response n
m_1_7	*papfqhw*, *paptcal*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 7, responses 2, 10, n, 1
m_1_8	*papfqhw*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 8, response n
m_2_1_1	*papfqhw*, *paptctp*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_2_1_2	*papfqhw*, *paptctp*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_2_1_3	*papfqhw*, *paptctp*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_2_2_1	*papfqhw*, *paptctdt*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_2_2_2	*papfqhw*, *paptctdt*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_3_1	*papfqhw*, *paptcal*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_3_2	*papfqhw*, *paptcal*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_4_1_1	*papfqd*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_4_1_2	*papfqd*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_4_2_1	*papfqc*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_4_2_2	*papfqc*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_4_3_1	*papfqh*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_4_3_2	*papfqh*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_4_4_1	*papfqg*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_4_4_2	*papfqg*, *paptcms*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, option 1
m_5_1	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_5_2	*papfqhw*, *paptcms*, *dmlprpr*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_5_3	*papfqhw*, *paptcms*, *dmlprin*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_5_4	*papfqhw*, *paptcms*, *dmlprsv*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_6_1	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgml*, *papenq*, *papcr*, option 1
m_6_2	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprdd*, *papwgml*, *papenq*, *papcr*, option 1
m_6_3	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmvt*, *papwgml*, *papendq*, *papcr*, option 1
m_6_4	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprddp*, *papwgml*, *papendq*, *papcr*, option 1
m_7_1_1	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgvc*, *papenqe*, *papcrqe*, option 1
m_7_1_2	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgvc*, *papenqe*, *papcrqe*, option 1
m_7_2_1	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgvc*, *papenqa*, *papcrqa*, option 1
m_7_2_2	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgvc*, *papenqa*, *papcrqa*, option 1
m_7_3_1	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgfc*, *papenqa*, *papcrqa*, option 1
m_7_3_2	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgfc*, *papenqa*, *papcrqa*, option 1
m_8_1	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwaml*, *papenq*, *papcr*, option 1
m_8_2	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgam*, *papenq*, *papcr*, option 1
m_9_1_1	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgml*, *papenq*, *papcr*, option 1
m_9_1_2	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprmv*, *papwgml*, *papenq*, *papcr*, option 1
m_9_1_3	*papfqhw*, *paptcms*, *dmlpr0*, *qmlprdd*, *papwgml*, *papenq*, *papcr*, option 1
m_9_2_1	*papfqhw*, *paptcms*, *dmlprpn*, *qmlprmv*, *papwgml*, *papendq*, *papcr*, option 1
m_9_2_2	*papfqhw*, *paptcms*, *dmlprpn*, *qmlprmv*, *papwgml*, *papendq*, *papcr*, option 1
m_10_1	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_2	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_3_1	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_3_2	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_4_1	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_4_2	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_5_1	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_5_2	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_6_1	*papfqxa*, *paptctp*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_6_2	*papfqxa*, *paptctp*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_6_3	*papfqxa*, *paptctp*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_7_1	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_7_2	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1
m_10_7_3	*papfqhw*, *paptcms*, *dmlprpn*, *qmlpr0*, *papwgml*, *papend*, *papcr*, option 1

A.7. Output Files

papout.tst Results of running the test dat

NAME	DESCRIPTION	COLUMN	TYPE

B. FILE FORMATS

All data files are limited to 80 character lengths. Repeating variables that exceed 80 characters extend to the following line. All integers must be right justified within the field. Real variables may include a decimal or whole numbers may be right justified. Double precision variables have seven-digit precision.

B.1. Pedigree Structure: *trip.dat*

(see section II.1)

Line repeated for each non-founder pedigree member.

LPEDGR	Pedigree number	1-8	Integer
JFATH	Father's ID number	9-16	Integer
JMOTH	Mother's ID number	17-24	Integer
JCHLD	Pedigree member's ID number	25-32	Integer

B.2. Computation Order: *order.dat*

(see section II.2)

Line repeated for each pair of parents.

JPED	Pedigree number	1-8	Integer
IFAMEM(1)	Husband's ID number	9-16	Integer
IFAMEM(2)	Wife's ID number	17-24	Integer

B.3. Trait/Marker Descriptions: *header.dat*

(see section II.3)

Required first line(s).

NDATA	Number of variables in *phen.dat*	1-4	Integer
NCOL	Number of columns in *phen.dat*	5-8	Integer
NAME(I)	Name of variable I, I = 1, NDATA	9-16,17-24, etc	Char

NAME	DESCRIPTION	COLUMN	TYPE
Required line(s).			
IVARTP(I)	Type of variable I, I = 1, NDATA = 1, discrete = 2, quantitative = 3, autosomal marker = 4, X-linked marker = 5, autosomal codominant marker specified as alleles = 6, X-linked codominant marker specified as alleles	9-12,17-20,etc	Integer
IVARDF(I)	Definition of variable I, I = 1, NDATA = 0, if IVARTP(I) = 1 and prevalence/incidence rates will not be used or if IVARTP(I) = 2 = location in *popln.dat* of prevalence/incidence rates if IVARTP(I) = 1 = (-) number of alleles, if IVARTP(I) = 3, 4, 5, or 6 and if no frequency/phenotype information in *popln.dat* = location in *popln.dat* of frequency/phenotype information, if IVARTP(I) = 3, 4, 5, or 6	13-16,21-24,etc	Integer
Optional line(s). MISVL(I) = -9999 if not included.			
-1	Code to indicate line of missing values	1-8	Integer
MISVL(I)	Missing value code for variable I, I = 1, NDATA	9-16,17-24,etc	Real
Optional line(s). PHSIM(I) = -99999 if not included.			
-2	Code to indicate line of simulation indicators	1-8	Integer
PHSIM(I)	Simulation indicator for variable I, I = 1, NDATA	9-16,17-24,etc	Real
Optional line(s). MEAN(I) = 0 if not included.			
-3	Code to indicate line of values to subtract	1-8	Integer
MEAN(I)	Value to subtract from variable I, I = 1, NDATA	9-16,17-24,etc	Real
Optional line(s). SD(I) = 1 if not included.			
-4	Code to indicate line of values to divide by	1-8	Integer
SD(I)	Value by which to divide variable I, I = 1, NDATA	9-16,17-24,etc	Real
Optional line(s). PWR(I) = 1 if not included.			
-5	Code to indicate line of power PWR(I) for the transformation function $y = r/P[(x/r+1)^P-1]$ [Maclean et al 1976], where x represents the phenotype, p represents the power and r = 6.	1-8	Integer

B.4. Phenotype Data: *phen.dat* and *ascer.dat*

(see section II.4)

Line repeated for each measured individual.

ID	ID number corresponding to *trip.dat*	1-8	Integer
PHEN	Individual's phenotype for column I, I = 1, NCOL	9-16,17-24,etc	Real

NAME	DESCRIPTION	COLUMN	TYPE

B.5. Population Information: *popln.dat*

(see section II.5)

File popln.dat can contain sets of lines of two types: for traits and for markers.

TRAIT SET: Initial line.

NCATEG	Number of age groups or severity categories	1-4	Integer

TRAIT SET: Line repeated NCATEG times.

IRANG1(I)	Lower limit of age group or severity category	1-4	Integer
IRANG2(I)	Upper limit of age group or severity category	5-8	Integer
RATE(1, I)	Incidence/prevalence in males for a sex-specific model in both genders combined for a non-sex-specific model		
RATE(2, I)	Incidence/prevalence in females for a sex-specific model unused if for a non-sex-specific model	17-24	Real

MARKER SET: Initial line.

NLINES	Number of additional lines in the set	1-4	
CONPN	Phenotype/genotype correspondent indicator = T, phenotype/genotype correspondences will follow = F, codominant inheritance	5-6	Logical
XLINK	Inheritance indicator = T, X-linked inheritance = F, autosomal inheritance	7-8	Logical

MARKER SET: Second line.

NALL	Number of marker alleles	1-3	Integer
FR(I)	Frequency of allele I, I = 1, NALL	4-18,19-33,etc	Real

MARKER SET: Included if CONPN = T.

NGNT	Number of marker genotypes	1-4	Integer
NMKPHN	Number of marker phenotypes	5-8	Integer

MARKER SET: Repeated NMKPHN time if CONPN = T.

IP(I)	Phenotype/genotype correspondent for marker genotype I, I = 1, NGNT = 1, genotype possible for phenotype = 0, genotype not possible for phenotype	1-2,3-4,etc	Integer

NAME	DESCRIPTION	COLUMN	TYPE

B.6. Pedigree and Phenotype Data: *papin.dat*

(written by *preped* (section III.1))

First line.

NDATA	Number of variables	1-8	Integer
IASCR	Ascertainment correction indicator = T, ascertainment correction = F, no ascertainment correction	9-10	Logical
AMATE	Ascertainment correction indicator = T, ascertainment correction = F, no ascertainment correction	11-12	Logical

Second line(s) (taken from header.dat).

NAMEV(I)	Name of variable I, I = 1, NDATA	1-8,9-16,etc	Char

Following line(s) (taken from header.dat).

IVARTP(I)	Type of variable I, I = 1, NDATA	1-4,9-12,etc	Integer
IVARDF(I)	Definition of variable I, I = 1, NDATA	5-8,13-16,etc	Integer

Repeated for each pedigree.

NUMPED	Pedigree number (taken from *trip.dat*)	1-8	Integer
NINDPD	Number of individuals in the pedigree	9-16	Integer
NFAMPD	Number of nuclear families in the pedigree	17-24	Integer
IDNMBR(I)	ID of individual assigned pedigree location I, I=1,NINDPD	25-32,49-56,etc	Integer
LOCFTH(I)	Pedigree location of father of I, I = 1, NINDPD	33-40,57-64,etc	Integer
LOCMTH(I)	Pedigree location of mother of I, I = 1, NINDPD	41-48,65-72,etc	Integer

Repeated NFAMPD times.

IFATH	Family location of father	1-8	Integer
IMOTH	Family location of mother	9-16	Integer
NCHJT	Number of offspring requiring joint likelihoods	17-24	Integer
ICH(I)	Family location of offspring, I = 1, NCHJT	25-32,+	Integer
NCS	Number in the cutset	33-40	Integer
NMEMJT	Number requiring joint likelihoods	41-48	Integer
NFAMEM	Number in the family	49-56	Integer
LOCID(I)	Pedigree location for family location I, I = 1, NFAMEM	57-64,65-72,etc	Integer

Repeated NINDPD times.

PH(I)	Phenotype for variable I, I = 1, NDATA	1-13,14-26,etc	Double

NAME	DESCRIPTION	COLUMN	TYPE

B.7. Model and Parameter Information: *model.dat*

(written by prepap (section III.3))

First line.

NAMEMD	Name to identify the file	2-11	Char
IFRSBR	Number of frequency subroutine selected	12-14	Integer
ITRSBR	Number of transmission subroutine selected	15-17	Integer
IMDSBR	Number of major locus discrete subroutine selected	18-20	Integer
IMQSBR	Number of major locus quantitative subroutine selected	21-23	Integer
IWGSBR	Number of within genotype subroutine selected	24-26	Integer
IPNSBR	Number of penetrance subroutine assigned	27-29	Integer
IVTPMD	Code designating quantitative/discrete = 0, neither discrete nor quantitative = 1, discrete only = 2, quantitative only = 3, both discrete and quantitative	30-32	Integer

Second line.

FRNAME	Name of frequency subroutine selected	2-11	Char
TRNAME	Name of transmission subroutine selected	13-22	Char
MDNAME	Name of major locus discrete subroutine selected	24-33	Char
MQNAME	Name of major locus quantitative subroutine selected	35-44	Char
WGNAME	Name of within genotype subroutine selected	46-55	Char
PNNAME	Name of penetrance subroutine assigned	57-66	Char
CRNAME	Name of correlated penetrance subroutine assigned	68-77	Char

Third line.

IXACOD	Genotype-assignment code	1-3	Integer
NLOCI	Number of loci	4-6	Integer
NLOCIX	Number of X-linked loci	7-9	Integer
NALL(I)	Number of alleles, I = 1, NLOCI	10-12,13-15,etc	Integer

Fourth line.

NVAR	Number of variables in the model	1-3	Integer
NMARKR	Number of markers in the model	4-6	Integer
NTRAIT	Number of traits in the model	7-9	Integer
NCNTP(1)	Number of types of concomitants for discrete traits	10-12	Integer
NCN(I,1)	Number of concomitants of type I, I = 1,NCNTP(1)	13-15,+	Integer
NCNTP(2)	Number of types of concomitants for quantitative traits	16-18	Integer
NCN(I,2)	Number of concomitants of type, I = 1, NCNTP(2)	19-21,+	Integer
NCNTP(3)	Number of types of concomitants for within genotypes	22-24	Integer
NCN(I,3)	Number of concomitants of type, I = 1, NCNTP(3)	25-27,+	Integer
LOCVAR(I)	Location of variable in *phen.dat*, I = 1, NVAR	28-30,31-33,etc	Integer
IVARTP(0)	Type of simulated variable	31-33	Integer
IVARDF(0)	(-) Number of alleles for simulated marker	34-36	Integer

Line included if NMARKR > 0.

KMARKR(I)	Position in LOCVAR of marker I, I = 1, NMARKR	1-3,7-9,etc	Integer
LMARKR(I)	Locus of marker I, I = 1, NMARKR	4-6,10-12,etc	Integer

NAME	DESCRIPTION	COLUMN	TYPE
Line included if NTRAIT > 0.			
KTRAIT(I)	Position in LOCVAR of trait I, I = 1, NTRAIT	1-3,10-12,etc	Integer
LTRAIT(I)	First locus of trait I, I = 1, NTRAIT	4-6,13-15,etc	Integer
JTRAIT(I)	Last locus of trait I, I = 1, NTRAIT	7-9,16-18,etc	Integer
Line included if NTRAIT > 0.			
KCNML(I,J)	Major locus concomitant locations, I = 1, NCNTP(L), J = 1, NTRAIT	1-3,+	Integer
KCNWG(I)	Within genotype concomitant locations, I = 1, NCNTP(3)	4-6,+	Integer
Repeated for each parameter type for each discrete trait.			
NPR	Number of major locus discrete parameters	1-3	Integer
PARMD(I)	Values of major locus discrete parameters, I = 1, NPR	4-18,19-33,etc	Double
Repeated for each parameter type for each quantitative trait.			
NPR	Number of major locus quantitative parameters	1-3	Integer
PARMQ(I)	Values of major locus quantitative parameters, I = 1, NPR	4-18,19-33,etc	Double
Repeated for each parameter type for each trait.			
NPR	Number of within genotype parameters	1-3	Integer
PARWG(I)	Values of within genotype parameters, I = 1, NPR	4-18,19-33,etc	Double
Determined by the frequency subroutine selected.			
NFREQS	Number of sets of frequencies	1-3	Integer
LOCOND(I)	(If IFRTYP = 4) Locus to condition the frequency on, I = 1, NLOCI	4-6,7-9,etc	Integer
Line repeated NFREQS times.			
NFREQ	Number of frequencies (negative indicates in *popln.dat*)	1-3	Integer
FREQ(I)	Frequency of classification I, I = 1, NFREQ	4-18,19-33,etc	Double
If INDRCM ≠ 0.			
N	Number of probabilities = NLOCI - 1	1-3	Integer
RCMMTH(I)	Female probability of recombination between locus I-1 and locus I, I = 2, NLOCI	4-18,19-33,etc	Real
If INDRCM ≠ 0 and NLOCI > NLOCIX + 1.			
N	Number of probabilities = NLOCI - NLOCIX - 1	1-3	Integer
RCMFTH	Male probability of recombination between locus I-1 and locus I, I = NLOCIX + 1, NLOCI	4-18,19-33,etc	Real

NAME	DESCRIPTION	COLUMN	TYPE
If INDTRP ≠0.			
N	Number of probabilities = NGENTL	1-3	Integer
TRNDAU(I)	Probability allele transmitted to daughter if parent genotype I, I = 1, NGENTL	4-18,19-33,etc	Real
If INDTRP ≠0 and IAXCOD ≠ 1 or 3.			
N	Number of probabilities = NGENTL	1-3	Integer
TRNSON(I)	Probability allele transmitted to son if parent genotype I, I = 1, NGENTL	4-18,19-33,etc	Real
Following line.			
NPARAM	Number of parameters to maximize, grid or compute standard errors	1-3	Integer
Repeated NPARAM times for each parameter to maximize, standard error or grid.			
X	Parameter value	1-15	Double
BOUND1	Lower bound for grid or maximization	16-30	Double
BOUND2	Upper bound for grid or maximization	31-45	Double
NPARSM	Number of parameters held to value	46-48	Integer
LPARAM	Array element of parameter	49-52	Integer
KPARAM	Code identifying parameter = 1, FREQ = 2, TRNDAU = 3, TRNSON = 4, RCMMTH = 5, RCMFTH = 6 to 5+NTRAIT, PARMD = 6+NTRAIT to 5+2 NTRAIT, PARMQ = 6+2 NTRAIT to 5+3 NTRAIT, PARWG	53-55	Integer
Next line.			
NCLIN	Number of linear constraints	1-3	Integer
Repeated NCLIN times.			
BOUND1	Lower bound for constraint	1-15	Double
BOUND2	Upper bound for constraint	16-30	Double
CONLIN(I)	Weight for parameter I, I = 1, NPARAM	31-45,46-60,etc	Double

NAME	DESCRIPTION	COLUMN	TYPE

B.8. Frequency Subroutine List: *freq.dat*

Initial line of the set repeated for each papfq

NUMSBR	Number stored as JFRSBR in *papfq*	1-3	Integer
NLDESC	Number of lines of description	4-6	Integer
NLTOTL	Number of additional lines in set	7-9	Integer
JXACOD	Genotype-assignment code indicator > 0, equals the code allowed = 0, unrestricted < 0, equals the negative of the code not allowed	10-12	Integer
NUMLOC	Number of loci indicator > 0, equals the number allowed = 0, unrestricted	13-15	Integer
NUMALL	Number of alleles indicator > 0, equals the number allowed = 0, unrestricted	16-18	Integer
IFRTYP	Frequency parameter code = 1, alleles = 2, haplotypes = 3, genotypes = 4, conditional alleles	19-21	Integer

Second line.

FRNAME	Name of frequency subroutine	2-11	Char

Line repeated NLDESC lines.

DESCR	Description of subroutine	1-50	Char

Next line.

IFRTYP	Frequency parameter code = 1, alleles = 2, conditional alleles = 3, haplotypes = 4, genotypes	1-3	Integer
NFRPR	Number of frequency parameters	4-6	Integer

Repeated NPARFR times.

IFRNG	Range code for frequency parameter = 4, 0 to 1 = 7, 0 to 1, sum restricted	1-3	Integer
DSFR	Name of frequency parameter	5-24	Char

NAME	DESCRIPTION	COLUMN	TYPE

B.9. Transmission Subroutine List: *tran.dat*

Initial line of set for each version of paptc.

NUMSBR	Number stored as JTRSBR in *paptc*	1-3	Integer
NLDESC	Number of lines of description	4-6	Integer
NLTOTL	Number of additional lines in the set	7-9	Integer
JXACOD	Genotype-assignment code indicator > 0, equals the code allowed = 0, unrestricted < 0, equals the negative of the code not allowed		
NUMLOC	Number of loci indicator > 0, equals the number allowed = 0, unrestricted	10-12	Integer
NUMALL	Number of alleles indicator > 0, equals the number allowed = 0, unrestricted	13-15	Integer
IRQHW	Hardy-Weinberg requirement indicator = 0, doesn't require Hardy-Weinberg for the frequencies = 1, requires Hardy-Weinberg for the frequencies	16-18	Integer

Second line.

TRNAME	Name of subroutine	1-10	Char

Repeated NLDESC times.

DESCR	Description of subroutine	1-50	Char

Following line.

INDTRP	Transmission probability indicator = 0, TRNDAU, TRNSON not used = 1, TRNDAU, TRNSON used = 2, TRNDAU, TRNSON used except for heterozygotes	1-3	Integer
INDRCM	Recombination probability indicator = 0, RCMMTH, RCMFTH not used = 1, RCMMTH, RCMFTH used	4-6	Integer

NAME	DESCRIPTION	COLUMN	TYPE

B.10. Major Locus Discrete Subroutine List: *dmlp.dat*

Initial line of set for each dmlpr.

NUMSBR	Number stored as JMDSBR in *dmlpr*	1-3	Integer
NLDESC	Number of lines of description	4-6	Integer
NLTOTL	Number of additional lines in the set	7-9	Integer
NUMLOC	Number of loci indicator > 0, equals the number allowed = 0, unrestricted	10-12	Integer
NUMALL	Number of alleles indicator > 0, equals the number allowed = 0, unrestricted	13-15	Integer
IXA	Mode of inheritance indicator = 0, unspecified = 1, X-linked inheritance = 2, autosomal inheritance	16-18	Integer
IPOPN	Population data indicator = 0, population data not necessary = 1, population data required	19-21	Integer

Second line.

MDNAME	Name of subroutine	1-10	Char

Repeated NLDESC times.

DESCR	Description of subroutine	1-50	Char

Following line.

NCNTP	Number of concomitant types	1-3	Integer
NPRTP	Number of parameter types	4-6	Integer

Repeated NCNTP times.

NC	Number of concomitants > 0, equals the number required < 0, equals the negative of the corresponding parameter	1-3	Integer
DSCN	Name of concomitant	5-24	Char

Repeated NPARPN times.

IPRCD	Number of parameter values = -99, equal to the number of trait genotypes = -96, # of trait loci if not sex-specific, twice otherwise = -95, # of trait loci if not parent-specific, twice otherwise	1-3	Integer
IPRNG	Range code for parameter = 1, -∞ to ∞ = 4, 0 to 1	4-6	Integer
DSPR	Name of parameter	8-27	Char

NAME	DESCRIPTION	COLUMN	TYPE

B.11. Major Locus Quantitative Subroutine List: *qmlp.dat*

Initial line of set for each qmlpr.

NUMSBR	Number stored as JMQSBR in *qmlpr*	1-3	Integer
NLDESC	Number of lines of description	4-6	Integer
NLTOTL	Number of additional lines in the set	7-9	Integer
NUMLOC	Number of loci indicator > 0, equals the number allowed = 0, unrestricted	10-12	Integer
NUMALL	Number of alleles indicator > 0, equals the number allowed = 0, unrestricted	13-15	Integer
IXA	Mode of inheritance indicator = 0, unspecified = 1, X-linked inheritance = 2, autosomal inheritance	16-18	Integer
IPOPN	Population data indicator = 0, population data not necessary = 1, population data required	19-21	Integer

Second line.

MQNAME	Name of subroutine	1-10	Char

Repeated NLDESC times.

DESCR	Description of subroutine	1-50	Char

Following line.

IVRTP	Discrete data indicator = 0, no discrete data = 1, discrete data for some individuals	1-3	Integer

Following line.

NCNTP	Number of concomitant types	1-3	Integer
NPRTP	Number of parameter types	4-6	Integer

Repeated NCNTP times.

NCN	Number of concomitants > 0, equals number < 0, equals negative of corresponding parameter	1-3	Integer
DSCN	Name of concomitant	5-24	Char

NAME	DESCRIPTION	COLUMN	TYPE
Repeated NPARPN times.			
IPRCD	Number of parameter values = -99, equals number of trait genotypes = -97, 1 if not sex-specific, 2 if sex-specific = -96, # of trait loci if not sex-specific, twice otherwise = -95, # of trait loci if not parent-specific, twice otherwise	1-3	Integer
IPRNG	Range code for parameter = 1, -∞ to ∞ = 2, 0 to ∞ = 3, 0.01 to ∞ = 4, 0 to 1	4-6	Integer
DSPR	Name of parameter	8-27	Char

B.12. Within Genotype Subroutine List: *wgen.dat*

Initial line of set for each papwg.

NUMSBR	Number stored as JWGSBR in *papwg*	1-3	Integer
NLDESC	Number of lines of description	4-6	Integer
NLTOTL	Number of additional lines in the set	7-9	Integer

Second line.

WGNAME	Name of subroutine	1-10	Char

Repeated NLDESC times.

DESCR	Description of subroutine	1-50	Char

Following line.

MLMX	Major locus/mixed model indicator 0 = major locus 1 = mixed	1-3	Integer

Following line.

NCNTP	Number of concomitant types	1-3	Integer
NPRTP	Number of parameter types	4-6	Integer

Repeated NCNTP times.

NCN	Number of concomitants > 0, equals number < 0, equals negative of corresponding parameter	1-3	Integer
DSCN	Name of concomitant	5-24	Char

NAME	DESCRIPTION	COLUMN	TYPE
Repeated NPARPN times.			
IPRCD	Number of parameter values > 0, equals number of values = -98, equals trait number - 1 < 0, user-specified number	1-3	Integer
IPRNG	Range code for parameter = 4, 0 to 1 = 5, -1 to 1 = 6, -1 to 1, not equal = 7, 0 to 1, sum restricted	4-6	Integer
DSPR	Name of parameter	8-27	Char

C. SUBROUTINE LIBRARIES

C.1. Frequency: *papfq*

C.1.1. *papfqc*

Genotype Assignments:	Not autosomal/X-linked admixture
Number of Loci:	At least 2
Number of Alleles:	2 at loci with conditional frequencies, unrestricted at other loci
Assumptions:	Hardy-Weinberg equilibrium
Parameters:	Conditional allele frequencies at some loci
	Unconditional allele frequencies at other loci

C.1.2. *papfqd*

Genotype Assignments:	Not autosomal/X-linked admixture
Number of Loci:	2
Number of Alleles:	2 at each locus
Assumptions:	Hardy-Weinberg equilibrium
Parameters:	Allele frequencies
	D (disequilibrium)

C.1.3. *papfqg*

Genotype Assignments:	Not autosomal/X-linked mixed
Number of Loci:	No restrictions
Number of Alleles:	No restrictions
Assumptions:	None
Parameters:	Genotype frequencies

C.1.4. *papfqh*

Genotype Assignments:	Autosomal
Number of Loci:	2+
Number of Alleles:	No restrictions
Assumptions:	Hardy-Weinberg equilibrium
Parameters:	Haplotype frequencies

C.1.5. *papfqhw*

Genotype Assignments:	Not Autosomal/X-linked mixed
Number of Loci:	No restrictions
Number of Alleles:	No restrictions
Assumptions:	Hardy-Weinberg equilibrium
	Linkage equilibrium
Parameters:	Allele Frequencies

C.1.6. *papfqxa* [Hasstedt & Skolnick 1984]

Genotype Assignments:	Autosomal/X-linked mixed
Number of Loci:	1
Number of Alleles:	2
Assumptions:	Hardy-Weinberg equilibrium in female Generational equilibrium in males
Parameters:	Allele frequencies

C.2. Transmission: *paptc*

C.2.1. *paptcal*

Genotype Assignments:	Autosomal
Number of Loci:	2+
Number of Alleles:	No restrictions
Assumptions:	Mendelian segregation
Parameters:	Sex-specific recombination probabilities

C.2.2. *paptcapl*

Genotype Assignments:	Parent-specific autosomal
Number of Loci:	2+
Number of Alleles:	No restrictions
Assumptions:	Mendelian segregation
Parameters:	Sex-specific recombination probabilities

C.2.3. *paptcasl*

Genotype Assignments:	Category-specific autosomal
Number of Loci:	2+
Number of Alleles:	No restrictions
Assumptions:	Mendelian segregation
Parameters:	Sex-specific recombination probabilities

C.2.4. *paptcaxl*

Genotype Assignments:	Autosomal/X-linked admixture
Number of Loci:	2+
Number of Alleles:	No restrictions
Assumptions:	Mendelian segregation
Parameters:	Sex-specific recombination probabilities

C.2.5. *paptce*

Genotype Assignments:	No restrictions
Number of Loci:	No restrictions
Number of Alleles:	No restrictions
Assumptions:	Transmission probabilities equal genotype frequencies
Parameters:	None

C.2.6. *paptcet*

Genotype Assignments:	Not autosomal/X-linked mixed model
Number of Loci:	1
Number of Alleles:	2
Assumptions:	Generational equilibrium
Parameters:	Allele transmission probabilities

C.2.7. *paptcms*

Genotype Assignments:	No restrictions
Number of Loci:	No restrictions
Number of Alleles:	No restrictions
Assumptions:	Mendelian segregation Unlinked loci
Parameters:	None

C.2.8. *paptctdt*

Genotype Assignments:	Category-specificAutosomal
Number of Loci:	1
Number of Alleles:	No restrictions
Assumptions:	None
Parameters:	Allele transmission probabilities

C.2.9. *paptctp*

Genotype Assignments:	Not autosomal/X-linked mixed
Number of Loci:	No restrictions
Number of Alleles:	2
Assumptions:	None
Parameters:	Allele transmission probabilities

C.2.10. *paptcxl*

Genotype Assignments:	X-linked
Number of Loci:	2+
Number of Alleles:	No restrictions
Assumptions:	Mendelian segregation
Parameters:	Female recombination probabilities

C.2.11 *paptcxpl*

Genotype Assignments:	Parent-specific X-linked
Number of Loci:	2+
Number of Alleles:	No restrictions
Assumptions:	Mendelian segregation
Parameters:	Sex-specific recombination probabilities

C.3. Major Locus Discrete: *dmlpr*

C.3.1. *dmlprin*

Number of Loci:	Unrestricted
Number of Alleles:	2
Mode of Inheritance:	Autosomal
Population Data:	Incidence rates
Phenotype:	Disease dichotomy
Assumptions:	Earlier onset corresponds to higher liability
	Additivity across loci
Parameters:	Dominance
	Displacement in within-genotype standard deviations

C.3.2. *dmlprinc*

Number of Loci:	Unrestricted
Number of Alleles:	2
Mode of Inheritance:	Autosomal
Population Data:	Incidence rates by category
Phenotype:	Disease dichotomy
Assumptions:	Earlier onset corresponds to higher liability within categories
	Additivity across loci
Parameters:	Dominance
	Displacement in within-genotype standard deviations

C.3.3. *dmlprpn*

Number of Loci:	Unrestricted
Number of Alleles:	Unrestricted
Mode of Inheritance:	Unrestricted
Population Data:	None
Phenotype:	Disease dichotomy
Assumptions:	None
Parameters:	Penetrance for each genotype

C.3.4. *dmlprpr*

Number of Loci:	Unrestricted
Number of Alleles:	2
Mode of Inheritance:	Autosomal
Population Data:	Prevalence rates
Phenotype:	Disease dichotomy
Assumptions:	Earlier onset corresponds to higher liability
	Additivity across loci
Parameters:	Dominance
	Displacement in within-genotype standard deviations

C.3.5. *dmlprsv*

Number of Loci:	Unrestricted
Number of Alleles:	2
Mode of Inheritance:	Autosomal
Population Data:	Prevalence by disease severity
Phenotype:	Severity code
Assumptions:	Penetrance increases across the genotypes
	Additivity across loci
Parameters:	Dominance
	Displacement in within-genotype standard deviations

C.4. Major Locus Quantitative: *qmlpr*

C.4.1. *qmlprdd*

Number of Loci:	Unrestricted
Number of Alleles:	2
Mode of Inheritance:	Autosomal
Population Data:	None
Assumptions:	Equal variance within genotypes
	Additivity across loci
Parameters:	Total mean
	Total standard deviation
	Dominance
	Displacement in within-genotype standard deviations
	Power for transformation
	Coefficients of linear effects (0-5)

C.4.2. *qmlprddp*

Number of Loci:	Unrestricted
Number of Alleles:	2
Mode of Inheritance:	Autosomal
Population Data:	None
Assumptions:	Equal variance within genotypes
	Additivity across loci
Parameters:	Total mean
	Total standard deviation
	Dominance
	Displacement in within-genotype standard deviations
	Proportion of population affected
	Power for transformation

C.4.3. *qmlprmv*

Number of Loci:	Unrestricted
Number of Alleles:	Unrestricted
Mode of Inheritance:	Unrestricted
Population Data:	None
Assumptions:	None
Parameters:	Mean for each genotype
	Standard deviation for each genotype
	Power for transformation
	Coefficients of linear effects (0-5)

C.4.4. *qmlprmvt*

Number of Loci:	Unrestricted
Number of Alleles:	Unrestricted
Mode of Inheritance:	Unrestricted
Population Data:	None
Assumptions:	Affection corresponds to a value above a specified threshold
Parameters:	Mean for each genotype
	Standard deviation for each genotype
	Threshold
	Power for transformation

C.5. Within Genotype: *papwg*

C.5.1. *papwgam* [Hasstedt 1995]

Assumptions:	Within-genotype correlations modeled as variance components
	Assortative mating
Parameters:	Assortative mating correlation
	Polygenic heritability
	Variance components for shared environments
	Genetic correlation between traits
	Environmental correlation between traits

C.5.2. *papwgfc*

Assumptions:	Within-genotype correlation modeled as familial correlations
	Familial correlations zero outside nuclear family
Parameters:	Spouse correlation
	Parent-offspring correlation
	Sibling correlation
	Correlation between traits

C.5.3. *papwgml*

Assumptions:	No within-genotype correlations
Parameters:	Correlation between traits

C.5.4. *papwgvc* [Hasstedt 1993]

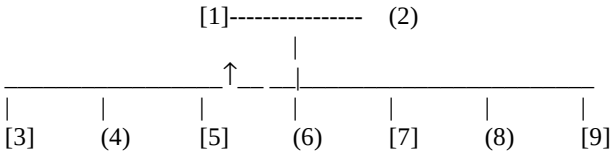
Assumptions:	Within-genotype correlations modeled as variance components
Parameters:	Polygenic heritability
	Variance components for shared environments
	Genetic correlation between traits
	Environmental correlation between traits

D. TEST EXAMPLES

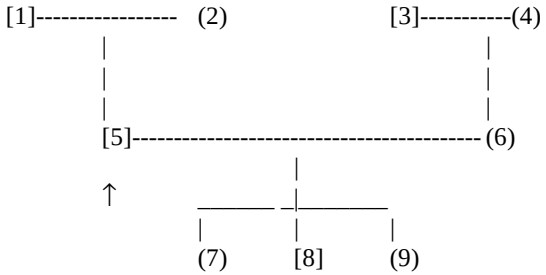
The PAP test data comprise seven data files (*trip.tst, header.tst1, header.tst2, header.tst3, header.tst4, phen.tst, ascer.tst*) and 59 parameter files (listed in section A.5). File *test.sh* contains the commands and *papout.tst* contains the output produced by successively executing *papdr* using each of the 59 parameter files. In this section I hand compute the likelihood for each example; Section IV describes each example.

The test data apply to the four 9-member pedigrees shown below. Pedigree 1 contains a large sibship; Pedigree 2 contains two ancestral couples; Pedigree 3 contains an inbreeding loop; and Pedigree 4 contains a multiple marriage. File *trip.tst* specifies these pedigree structures for input to PAP; each ID number in *trip.tst* equals the pedigree number followed by the individual number, so that each individual has a unique number. For descriptions of the contents and format of file *trip.dat* see section II.1 and section B.1, respectively.

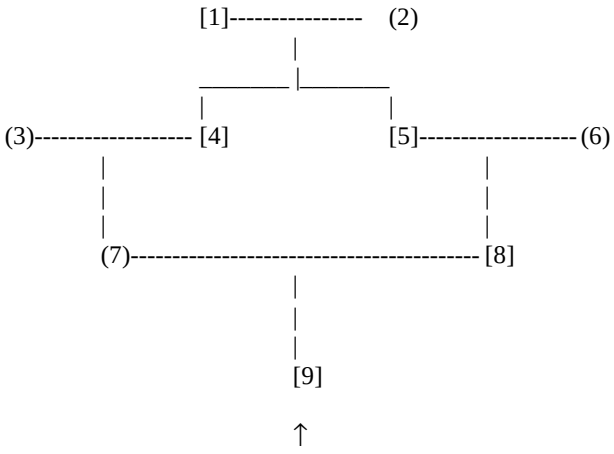
Pedigree 1



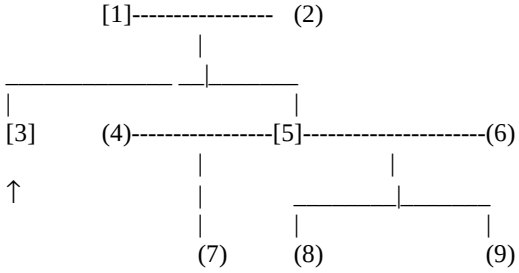
Pedigree 2



Pedigree 3



Pedigree 4



The test data comprise phenotypes for markers and quantitative and discrete trait traits as shown below. Each ID number equals the pedigree number followed by the individual number corresponding to the ID numbers in *trip.tst*. File *phen.tst* contains the listed phenotypes. For descriptions of the contents and format of *phen.dat* see section II.4 and section B.4, respectively.

ID	SEX	AGE	GENOTYPE	DISEASE	QUANTIT	MARITAL		MARKER
11	1	20	3	2	5.5	1	2	2
12	2	40	2	1	-0.5	1	1	2
13	1	40	2	1	0.0	-9999	1	2
14	2	40	2	1	0.5	-9999	1	2
15	1	20	3	2	5.0	-9999	2	2
16	2	20	3	2	4.5	-9999	2	2
17	1	40	2	1	-0.5	-9999	1	2
18	2	20	3	2	4.5	-9999	2	2
19	1	40	2	1	0.0	-9999	1	2
21	1	40	2	1	0.5	1	1	2
22	2	40	2	1	-0.5	1	1	2
23	1	40	1	1	-5.5	2	1	1
24	2	40	1	1	-5.0	2	1	1
25	1	20	3	2	5.0	3	2	2
26	2	40	1	1	-4.5	3	1	1
27	2	40	2	1	0.0	-9999	1	2
28	1	40	2	1	0.5	-9999	1	2
29	2	40	2	1	-0.5	-9999	1	2
31	1	20	3	2	5.5	1	2	2
32	2	40	1	1	-5.5	1	1	1
33	2	40	1	1	-5.0	2	1	1
34	1	40	2	1	0.0	2	1	2
35	1	40	2	1	0.5	3	1	2
36	2	40	1	1	-4.5	3	1	1
37	2	40	2	1	-0.5	4	1	2
38	1	40	2	1	0.0	4	1	2
39	1	20	3	2	4.5	-9999	2	2
41	1	40	2	1	0.5	1	1	2
42	2	40	2	1	-0.5	1	1	2
43	1	20	3	2	5.0	-9999	2	2
44	2	40	1	1	-5.5	-9999	1	1
45	1	20	3	2	5.5	2	2	2
46	2	40	1	1	-5.0	2	1	1
47	2	40	2	1	0.0	-9999	1	2
48	2	40	2	1	0.5	-9999	1	2
49	2	40	2	1	-0.5	-9999	1	2

We assume each pedigree was ascertained through a single affected proband. File *ascer.tst* contains the four entries listed below; missing values were assigned for variables other than the discrete trait designated DISEASE. For descriptions of the contents and format of *ascer.dat* see section II.4 and section B.4, respectively.

ID	SEX	AGE	GENOTYPE	DISEASE	QUANTIT	MARITAL		MARKER
11	1	20	-9999	2	-9999	-9999	-9999	-9999
25	1	20	-9999	2	-9999	-9999	-9999	-9999
31	1	20	-9999	2	-9999	-9999	-9999	-9999
43	1	20	-9999	2	-9999	-9999	-9999	-9999

Each individual's phenotypes were assigned to reflect the same autosomal genotype in order to simplify the following hand computation of the likelihoods. Each is computed as the $\log_{10}$ likelihood; you can obtain -2 ln likelihoods by multiplying by -2 $\log_e$ (10) or approximately -4.6.

D.1. Execution Options

For the execution option examples, the genetic model comprises 1 locus with 2 alleles; the variable comprises the marker GENOTYPE. Since GENOTYPE is a codominant marker, each individual's phenotype corresponds to his/her genotype at the locus. Since there are no missing values, the genotypes of all pedigree members are known. The likelihood equals the product of the genotype frequency for each founder and the genotype transmission probability for each nonfounder. Since all probands have missing values for GENOTYPE, the ascertainment correction equals 1.

Letting p = frequency of allele 1, q = 1 - p, and assuming Hardy-Weinberg equilibrium, the probability for each individual and pedigree equals:

	Pedigree			
Individual	1	2	3	4
1	q^2	2pq	q^2	2pq
2	2pq	2pq	p^2	2pq
3	½	p^2	p^2	¼
4	½	p^2	1	p^2
5	½	1/4	1	¼
6	½	¼	p^2	p^2
7	½	1	½	1
8	½	1	½	1
9	½	1	¼	1
Total	$pq^3/2^6$	p^6q^2	$p^6q^2/2^4$	$p^6q^2/4$

Total likelihood = $p^{19}q^9/2^{12}$

D.1.1. Single likelihood (*papfqhw, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_1_1*, option 1)

For p = 0.7, q = 0.3, the $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-3.5297
2	-1.9752
3	-3.1793
4	-2.5772
Total	-11.2614

D.1.2. Genotype probabilities (*papfqhw, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_1_2*, option 2, response y and y)

For each individual, the probability equals 1 for the genotype specified in *phen.dat* and 0 for the other 2 genotypes.

D.1.3. Simulation and output to a file (*papfqhw, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_1_3_1*, option 3, response 2 and 1)

There is no single correct result for simulated phenotypes. That parents and offspring have compatible genotypes is confirmed through the computation of a nonzero likelihood (using file *m_1_3_2*).

D.1.4. Search for maximum likelihood (*papfqhw, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_1_4*, option 4, response n)

Setting the derivative of the ln likelihood to 0 and solving the equation produces the maximum likelihood estimator.
ln likelihood = 19 ln (p) + 9 ln (1 - p)
d ln likelihood/dp = 19/p - 9/(1 - p) = 0
$\hat{p}$ = 19/28 = 0.6786

Computing the total $\log_{10}$ likelihood using $\hat{p}$ produces the maximum likelihood.

$\log_{10}$ likelihood = -11.2483

D.1.5. Maximize likelihood (*papfqhw, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_1_5*, option 5, response n, n and n)

See Section D.1.4.

D.1.6. Standard error of a parameter (*papfqhw, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_1_6*, option 6, response n)

Since the allele frequency is a proportion, the variance of the estimate equals $\hat{p}\hat{q}/n$. Therefore the standard error of $\hat{p}$ equals:
Sqrt($\hat{p}\hat{q}/n$) = 0.08826

Since *papdr* approximates the standard error, it may differ more from the theoretical value than occurs in this case.

D.1.7. Simulation and estimation (*papfqhw, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_1_7*, option 7, responses 2, 10, and n)

Again, there is no single correct answer. The mean estimate across the replicates should approximately equal the simulated value of 0.7, and the standard deviation of the estimate should be approximately 0.145. However, expect a lot of variation for only 10 replicates.

D.1.8. Grid on a parameter (*papfqhw, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_1_8*, option 8)

Dividing the range from 0.625 to 0.75 into five equal intervals produces six values of p with corresponding $\log_{10}$ likelihoods equal to:

p	Log_{10} Likelihood
0.625	-11.3244
0.65	-11.2704
0.675	-11.2486
0.7	-11.2614
0.725	-11.3119
0.75	-11.4047

D.2. Transmission Parameters

D.2.1. Segregation Analysis

For the segregation analysis examples, the genetic model comprises 1 autosomal locus with 2 alleles; the variable comprises the marker GENOTYPE. Designating the two alleles as A and a, we represent the probability that a parent transmits allele A to an offspring as $\tau_1 = \Pr\{AA \rightarrow A\}$, $\tau_2 = \Pr\{Aa \rightarrow A\}$, $\tau_3 = \Pr\{aa \rightarrow A\}$ for the three genotypes. As in D.1, the likelihood equals the product of the genotype frequency for each founder and the genotype transmission probability for

each nonfounder, where the transmission probabilities now depend on τ_1, τ_2, and τ_3. Since all the probands have missing values for GENOTYPE, the ascertainment correction equals 1. The probability for each individual equals:

	Pedigree			
ID	1	2	3	4
1	q^2	$2pq$	q^2	$2pq$
2	$2pq$	$2pq$	p^2	$2pq$
3	$\tau_3(1-\tau_2)+\tau_2(1-\tau_3)$	p^2	p^2	$(1-\tau_2)^2$
4	$\tau_3(1-\tau_2)+\tau_2(1-\tau_3)$	p^2	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$	p^2
5	$(1-\tau_2)(1-\tau_3)$	$(1-\tau_2)^2$	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$	$(1-\tau_2)^2$
6	$(1-\tau_2)(1-\tau_3)$	τ^2_1	p^2	p^2
7	$\tau_3(1-\tau_2)+\tau_2(1-\tau_3)$	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$
8	$(1-\tau_2)(1-\tau_3)$	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$
9	$\tau_3(1-\tau_2)+\tau_2(1-\tau_3)$	$\tau_1(1-\tau_3)+\tau_3(1-\tau_1)$	$(1-\tau_2)^2+\tau_3(1-\tau_1)$	$\tau_1(1-\tau_3)$

And the likelihood of each pedigree equals:

Pedigree	Likelihood
1	$2pq^3\,(1-\tau_2)^3\,(1-\tau_3)^3\,[\tau_2(1-\tau_3)+\tau_3(1-\tau_2)]^4$
2	$4p^6q^2\,\tau^2_1\,(1-\tau_2)^2\,[\tau_1(1-\tau_3)+\tau_3(1-\tau_1)]^3$
3	$p^6q^2\,(1-\tau_2)^2\,[\tau_1(1-\tau_3)+\tau_3(1-\tau_1)]^2\,[\tau_1(1-\tau_2)+\tau_2(1-\tau_1)]^2$
4	$4p^6q^2\,(1-\tau_2)^4\,[\tau_1(1-\tau_3)+\tau_3(1-\tau_1)]^3$

D.2.1.1. Mendelian segregation (*papfqhw, paptctp, dmlpr0, qmlpr0, papwg, papen, papcr, m_2_1_1*, option 1)

For $\tau_1 = 1$, $\tau_2 = ½$, $\tau_3 = 0$, p = 0.7, q = 0.3, the $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-3.5297
2	-1.9752
3	-3.1793
4	-2.5772
Total	-11.2614

D.2.1.2. Environmental nontransmission (*papfqhw, paptctp, dmlpr0, qmlpr0, papwg, papen, papcr, m_2_1_2*, option 1)

For $\tau_1 = 0.7$, $\tau_2 = 0.7$, $\tau_3 = 0.7$, p = 0.7, q = 0.3, the $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-6.0668
2	-3.8589
3	-4.5279
4	-4.5949
Total	-19.0485

D.2.1.3. General transmission model (*papfqhw, paptctp, dmlpr0, qmlpr0, papwg, papen, papcr, m_2_1_3*, option 1)

For $\tau_1 = 0.9$, $\tau_2 = 0.4$, $\tau_3 = 0.1$, p = 0.7, q = 0.3, the $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-3.7323
2	-2.1669
3	-3.0644
4	-2.5191
Total	-11.4827

D.2.2. Transmission/Disequilibrium Test

For the transmission/disequilibrium test examples, the genetic model comprises 1 autosomal locus with 2 alleles; the variable comprises the marker GENOTYPE. Designating the two alleles as A and a, the probability that a parent transmits allele A to an offspring, $\Pr\{AA \rightarrow A\} = 1$, $\tau = \Pr\{Aa \rightarrow A\}$ if the offspring is affected, $\Pr\{Aa \rightarrow A\} = ½$ if the offspring is un affected, $\Pr\{aa \rightarrow A\} = 0$. As in D.1, the likelihood equals the product of the genotype frequency for each founder and the genotype transmission probability for each nonfounder, where the transmission probabilities now depend on τ_1, τ_2, and τ_3. No ascertainment correction is made. The probability for each individual equals:

	Pedigree			
Individual	1	2	3	4
1	q^2	2pq	q^2	2pq
2	2pq	2pq	p^2	2pq
3	½	p^2	p^2	$(1-\tau)^2$
4	1/2	p^2	1	P^2
5	1-τ	$(1-\tau)^2$	1	$(1-\tau)^2$
6	1-τ	1	p^2	p^2
7	½	1	½	1
8	1-τ	1	½	1
9	1/2	1	$(1-\tau)^2$	1

And the likelihood of each pedigree equals:

Pedigree	Log_{10} Likelihood
1	$pq^3 (1-\tau)^3/2^3$
2	$4p^6q^2 (1-\tau)^2$
3	$p^6q^2 (1-\tau)^2/2^2$
4	$4p^6q^2 (1-\tau)^4$

D.2.2.1. No Linkage (*papfqhw, paptctdt, dmlpr0, qmlpr0, papwg, papen, papcr, m_2_2_1*, option 1)

For $\tau = 0.5$, p = 0.7, q = 0.3, the $\log_{10}$ likelihood of each pedigree and the total equals (as in D.2.1.1):

Pedigree	Log_{10} Likelihood
1	-3.5297
2	-1.9752
3	-3.1793
4	-2.5772
Total	-11.2614

D.2.2.2. Linkage (*papfqhw*, *paptctdt*, *dmlpr0*, *qmlpr0*, *papwg*, *papen*, *papcr*, *m_2_2_2*, option 1)

For $\tau = 0.3$, $p = 0.7$, $q = 0.3$, the $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-3.0913
2	-1.6829
3	-2.8870
4	-1.9927
Total	-9.6539

D.3. Recombination Parameters

For the recombination parameter examples, the genetic model comprises 2 loci with 2 alleles each; the variables comprise the markers GENOTYPE and MARKER. Designating the alleles for GENOTYPE as A and a, and the alleles for MARKER as B and b, there are four haplotypes and 10 genotypes, numbered as follows:

	AB	Ab	aB	Ab
AB	1			
Ab	2	3		
AB	4	5	6	
Ab	7	8	9	10

For each individual, phenotypes for GENOTYPE and MARKER fix the identical 1-locus genotype and all 2-locus genotypes except 5 and 7, the double heterozygotes. The possible genotypes include 1 set for Pedigree 3, 2 sets for Pedigree 1, and four sets for Pedigrees 2 and 4. Since all the probands have missing values for both GENOTYPE and MARKER, the ascertainment correction equals 1. Subscripting f to represent the corresponding haplotype frequency and letting θ = recombination probability, the probability by pedigree and individual equals:

Pedigree 1: Two possible sets of genotypes

ID	Genotype Set 1		Genotype Set 2	
	Genotype	Probability	Genotype	Probability
1	ab/ab	f_{10}	ab/ab	f_{10}
2	Ab/aB	f_5	AB/ab	f_7
3	AB/ab	$\theta/2$	AB/ab	$(1-\theta)/2$
4	AB/ab	$\theta/2$	AB/ab	$(1-\theta)/2$
5	ab/ab	$\theta/2$	ab/ab	$(1-\theta)/2$
6	ab/ab	$\theta/2$	ab/ab	$(1-\theta)/2$
7	AB/ab	$\theta/2$	AB/ab	$(1-\theta)/2$
8	ab/ab	$\theta/2$	ab/ab	$(1-\theta)/2$
9	AB/ab	$\theta/2$	AB/ab	$(1-\theta)/2$

Likelihood = $f_{10} [f_5 \theta^7 + f_7 (1-\theta)^7]/2^7$

Pedigree 2: Four possible sets of genotypes

ID	Genotype Set 1		Genotype Set 2		Genotype Set 3		Genotype Set 4	
	Genotype	Probability	Genotype	Probability	Genotype	Probability	Genotype	Probability
1	Ab/aB	f_5	Ab/aB	f_5	AB/ab	f_7	AB/ab	f_7
2	Ab/aB	f_5	AB/ab	f_7	Ab/aB	f_5	AB/ab	f_7
3	AB/AB	f_1	AB/AB	f_1	AB/AB	f_1	AB/AB	f_1
4	AB/AB	f_1	AB/AB	f_1	AB/AB	f_1	AB/AB	f_1
5	ab/ab	$\theta^2/4$	ab/ab	$\theta(1-\theta)/4$	ab/ab	$\theta(1-\theta)/4$	ab/ab	$(1-\theta)^2/4$
6	AB/AB	1	AB/AB	1	AB/AB	1	AB/AB	1
7	AB/ab	1	AB/ab	1	AB/ab	1	AB/ab	1
8	AB/ab	1	AB/ab	1	AB/ab	1	AB/ab	1
9	AB/ab	1	AB/ab	1	AB/ab	1	AB/ab	1

Likelihood = $f^2_1 [f_5 \theta + f_7 (1-\theta)]^2/4$

Pedigree 3: One possible set of genotype

ID	Genotype	Probability
1	AB/AB	f_{10}
2	ab/ab	f_1
3	ab/ab	f_1
4	ab/AB	1
5	ab/AB	1
6	ab/ab	f_1
7	ab/AB	$(1-\theta)/2$
8	ab/AB	$(1-\theta)/2$
9	AB/AB	$(1-\theta)^2/4$

Likelihood = $f^3_1 f_{10} (1-\theta)^4/16$

Pedigree 4: Four possible sets of genotypes

ID	Genotype Set 1		Genotype Set 2		Genotype Set 3		Genotype Set 4	
	Genotype	Probability	Genotype	Probability	Genotype	Probability	Genotype	Probability
1	Ab/aB	f_5	Ab/aB	f_5	AB/ab	f_7	AB/ab	f_7
2	Ab/aB	f_5	AB/ab	f_7	Ab/aB	f_5	AB/ab	f_7
3	ab/ab	$\theta^2/4$	ab/ab	$\theta(1-\theta)/4$	ab/ab	$\theta(1-\theta)/4$	ab/ab	$(1-\theta)^2/4$
4	AB/AB	f_1	AB/AB	f_1	AB/AB	f_1	AB/AB	f_1
5	ab/ab	$\theta^2/4$	ab/ab	$\theta(1-\theta)/4$	ab/ab	$\theta(1-\theta)/4$	ab/ab	$(1-\theta)^2/4$
6	AB/AB	f_1	AB/AB	f_1	AB/AB	f_1	AB/AB	f_1
7	AB/ab	1	AB/ab	1	AB/ab	1	AB/ab	1
8	AB/ab	1	AB/ab	1	AB/ab	1	AB/ab	1
9	AB/ab	1	AB/ab	1	AB/ab	1	AB/ab	1

Likelihood = $f^2_1 [f_5 \theta^2 + f_7 (1-\theta)^2]^2/4^2$

D.3.1. Unlinked (*papfqhw, paptcal, dmlpr0, qmlpr0, papen, papcr, m_3_1*, option 1)

For $\theta = ½$, $f_1 = 0.3136$, $f_5 = 0.0672$, $f_7 = 0.0672$, $f_{10} = 0.0036$, the $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-7.5297
2	-3.9546
3	-6.3628
4	-5.1587
Total	-23.0058

D.3.2. Linked (*papfqhw, paptcal, dmlpr0, qmlpr0, papen, papcr, m_3_2*, option 1)

For $\theta = 0.0$, $f_1 = 0.3136$, $f_5 = 0.0672$, $f_7 = 0.0672$, $f_{10} = 0.0036$, the $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-5.7235
2	-3.9546
3	-5.1587
4	-4.5566
Total	-19.3934

D.4. Frequency Parameters

For the frequency parameter examples, the genetic model comprises 2 unlinked loci with 2 alleles each; the variables comprise the markers GENOTYPE and MARKER. Setting $\theta = ½$, $f_h = f_5 + f_7$ in the expression in section D.3, the likelihood of each pedigree equals:

Pedigree	Log_{10} Likelihood
1	$f_h f_{10}/4^7$
2	$f^2{}_1 f_h{}^2/16$
3	$f^3{}_1 f_{10}/4^4$
4	$f^2{}_1 f_h{}^2/4^4$

D.4.1. Allele frequencies

Let p_A represent the frequency of allele A, p_B represent the frequency of allele B, D represent disequilibrium, $q_a = 1 - p_A$, and $q_b = 1 - p_B$. Then the genotype frequencies equal:

$f_1 = (p_A p_B + D)^2$
$f_5 = 2(p_A q_b - D)(p_B q_a - D)$
$f_7 = 2(p_A p_B + D)(q_a q_b + D)$
$f_{10} = (q_a q_b + D)^2$

D.4.1.1. Equilibrium (*papfqd, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_4_1_1*, option 1)

Linkage equilibrium is defined as D = 0. Setting $p_A = 0.7$, $q_a = 0.3$, $p_B = 0.8$, $q_b = 0.2$, then $f_1 = 0.3136$, $f_5 = 0.0672$, $f_7 = 0.0672$, $f_{10} = 0.0036$, and the likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-7.5297
2	-3.9546
3	-6.3628
4	-5.1587
Total	-23.0058

D.4.1.2. Disequilibrium (*papfqd, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_4_1_2*, option 1)

Letting D = - 0.02, and $p_A = 0.7$, $q_a = 0.3$, $p_B = 0.8$, $q_b = 0.2$, then $f_1 = 0.2916$, $f_5 = 0.0832$, $f_7 = 0.0432$, $f_{10} = 0.0016$, and the likelihood of each pedigree and the total equals:

Pedigree	Log_{10} Likelihood
1	-7.9086
2	-4.0711
3	-6.8098
4	-5.2752
Total	-24.0645

D.4.2. Conditional allele frequencies

Letting $p_{A|B}$ represent the frequency of allele A conditional on allele B, $p_{A|b}$ represent the frequency of allele A conditional on allele b, $q_{A|B} = 1 - p_{A|B}$, $q_{A|b} = 1 - p_{A|b}$, p_B = frequency of allele B, $q_b = 1 - p_B$ the genotype frequencies equal:

$f_1 = (p_{A|B}p_B)^2$
$f_h = p_{A|B}p_B\, q_{A|b}q_b + p_{A|b}q_b\, q_{A|B}p_B$
$f_{10} = (q_{A|b}q_b)^2$

D.4.2.1. Equilibrium (*papfqc, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_4_2_1*, option 1)

Letting $p_{A|B} = 0.7$, $q_{A|B} = 0.3$, $p_{A|b} = 0.7$, $q_{A|b} = 0.3$, $p_B = 0.8$, $q_b = 0.2$, then $f_1 = 0.3136$, $f_5 = 0.0672$, $f_7 = 0.0672$, $f^1_0 = 0.0036$. See section D.4.1.1 for the likelihood computation.

D.4.2.2. Disequilibrium (*papfqc, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_4_2_2*, option 1)

Letting $p_{A|B} = 0.675$, $q_{A|B} = 0.325$, $p_{A|b} = 0.8$, $q_{A|b} = 0.2$, $p_B = 0.8$, $q_b = 0.2$, then $f_1 = 0.2916$, $f_5 = 0.0832$, $f_7 = 0.0432$, $f_{10} = 0.0016$. See section D.4.1.2 for the likelihood computation.

D.4.3. Haplotype frequencies

Letting h_{AB} represent the frequency of haplotype AB, h_{Ab} represent the frequency of haplotype Ab, h_{aB} represent the frequency of haplotype aB, and h_{ab} represent the frequency of haplotype ab, the genotype frequencies equal:

$f_1 = h^2_{AB}$
$f_h = h_{AB}h_{ab} + h_{Ab}h_{aB}$
$f_{10} = h^2_{ab}$

D.4.3.1. Equilibrium (*papfqh, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_4_3_1*, option 1)

Letting $h_{AB} = 0.56$, $h_{Ab} = 0.14$, $h_{aB} = 0.24$, $h_{ab} = 0.06$, then $f_1 = 0.3136$, $f_5 = 0.0672$, $f_7 = 0.0672$, $f^1_0 = 0.0036$. See section D.4.1.1 for the likelihood computation.

D.4.3.2. Disequilibrium (*papfqh, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_4_3_2*, option 1)

Letting $h_{AB} = 0.54$, $h_{Ab} = 0.16$, $h_{aB} = 0.26$, $h_{ab} = 0.04$, then $f_1 = 0.2916$, $f_5 = 0.0832$, $f_7 = 0.0432$, $f_{10} = 0.0016$. See section D.4.1.2 for the likelihood computation.

D.4.4. Genotype frequencies

D.4.4.1. Equilibrium (*papfqg, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_4_4_1*, option 1)

Let $f_1 = 0.3136$, $f_5 = 0.0672$, $f_7 = 0.0672$, $f_{10} = 0.0036$. See section D.4.1.1 for the likelihood computation.

D.4.4.2. Disequilibrium (*papfqg, paptcms, dmlpr0, qmlpr0, papwg, papen, papcr, m_4_4_2*, option 1)

Let $f_1 = 0.2916$, $f_5 = 0.0832$, $f_7 = 0.0432$, $f_{10} = 0.0016$. See section D.4.1.2 for the likelihood computation.

D.5. Discrete Trait Parameters

For the discrete trait examples, the genetic model comprises 1 autosomal locus with 2 alleles; the variable comprises the discrete trait DISEASE. In this case, the genotype cannot be unequivocally inferred from the phenotype. However, a rare allele frequency and recessive inheritance with affection probabilities of 0 and 1 combine to produce a high probability for the assigned genotype for each individual. The small probability of other genotypes is ignored in the hand likelihood computation, making it differ slightly from the likelihood computed by *papdr*, but identical to the computation in section D.1. Therefore, the likelihood of each pedigree and the total equals:

Pedigree	Likelihood
1	$pq^3/2^6$
2	p^6q^2
3	$p^6q^2/2^4$
4	$p^6q^2/4$
Total	$p^{19}q^9/2^{12}$

The proband in each pedigree is affected. Therefore, the ascertainment correction for each pedigree equals the probability that a random individual is affected or q^2.

D.5.1. Affection probability (*papfqhw, paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_5_1*, option 1)

Letting $p = 0.999$, $q = 0.001$, the $\log_{10}$ likelihood of each pedigree and the total equals:

	Log_{10} Likelihood	
Pedigree	Uncorrected	Corrected
1	-10.8066	-4.8066
2	-6.0026	-0.0026
3	-7.2067	-1.2067
4	-6.6047	-0.6047
Total	-30.6206	-6.6206

D.5.2. Prevalence (*papfqhw, paptcms, dmlprpr, qmlpr0, papwgml, papend, papcr, m_5_2*, option 1)

Let the prevalence equal 1 X 10^{-6}, dominance equal 0, and displacement equal 10. See section D.5.1 for the likelihood calculation.

D.5.3. Incidence (*papfqhw, paptcms, dmlprin, qmlpr0, papwgml, papend, papcr, m_5_3*, option 1)

Let the incidence equal 1 X 10^{-6}, dominance equal 0, and displacement equal 10. See section D.5.1 for the likelihood calculation.

D.5.4. Severity classes (*papfqhw, paptcms, dmlprsv, qmlpr0, papwgml, papend, papcr, m_5_4*, option 1)

Let the incidence equal 1 X 10^{-6}, dominance equal 0, and displacement equal 10. See section D.5.1 for the likelihood calculation.

D.6. Quantitative Trait Parameters

For the quantitative trait examples, the genetic model comprises 1 locus with 2 alleles; the variable comprises the quantitative trait QUANTIT. Again the genotype cannot be unequivocally inferred from the phenotype. However, a rare allele frequency and recessive inheritance with a small standard deviation combine to produce a high probability for the assigned genotype for each individual. The likelihood equals the product of the genotype frequency for each founder, the transmission probability for each nonfounder, and the penetrance for each measured individual. Assuming the assigned genotype, the penetrance for each individual equals:

	Pedigree			
	1	2	3	4
1	$N((5.5-\mu_{aa})/\sigma_{aa},\sigma_{aa})$	$N((.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((5.5-\mu_{aa})/\sigma_{aa},\sigma_{aa})$	$N((.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$
2	$N((-.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((-.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((-5.5-\mu_{AA})/\sigma_{AA},\sigma_{AA})$	$N((-.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$
3	$N((0-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((-5.5-\mu_{AA})/\sigma_{AA},\sigma_{AA})$	$N((-5.-\mu_{AA})/\sigma_{AA},\sigma_{AA})$	$N((5.-\mu_{aa})/\sigma_{aa},\sigma_{aa})$
4	$N((.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((-5.-\mu_{AA})/\sigma_{AA},\sigma_{AA})$	$N((0-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((-5.5-\mu_{AA})/\sigma_{AA},\sigma_{AA})$
5	$N((5.-\mu_{aa})/\sigma_{aa},\sigma_{aa})$	$N((5.-\mu_{aa})/\sigma_{aa},\sigma_{aa})$	$N((.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((5.5-\mu_{aa})/\sigma_{aa},\sigma_{aa})$
6	$N((4.5-\mu_{aa})/\sigma_{aa},\sigma_{aa})$	$N((-4.5-\mu_{AA})/\sigma_{AA},\sigma_{AA})$	$N((-4.5-\mu_{AA})/\sigma_{AA},\sigma_{AA})$	$N((-5.-\mu_{AA})/\sigma_{AA},\sigma_{AA})$
7	$N((-.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((0-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((-.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((0-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$
8	$N((4.5-\mu_{aa})/\sigma_{aa},\sigma_{aa})$	$N((.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((0-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$
9	$N((0-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((-.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$	$N((4.5-\mu_{aa})/\sigma_{aa},\sigma_{aa})$	$N((-.5-\mu_{Aa})/\sigma_{Aa},\sigma_{Aa})$

where $N(\delta,\sigma)$ represents the height of a normal density for deviation δ and standard deviation σ, the number is the individual's quantitative phenotype, μ_i and σ_i represents the mean and standard deviation, respectively, of the assigned genotype, i = AA, Aa, or aa. The $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	$\log_{10}$ Likelihood		
1	3 n(0) + 6 n(.5) -	10.80	-14.7243
2	3 n(0) + 6 n(.5) -	6.0026	-9.9203
3	3 n(0) + 6 n(.5) -	7.2067	-11.1244
4	3 n(0) + 6 n(.5) -	6.6047	-10.5224
Total			-46.2914

where $n(0) = \log_{10} N(0) = -.3991$ and $n(.5) = \log_{10} N(.5) = \log_{10} N(-.5) = -.4534$, and the $\log_{10}$ likelihood of the product of the frequencies and transmission probabilities comes from section D.5.1. The small probability of other genotypes than the assigned genotype is ignored in the hand likelihood computation, making it differ slightly from the likelihood computed by *papdr*.

D.6.1. Means/Standard Deviations (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgml, papenq, papcr, m_6_1*, option 1)

Setting p = 0.999, q = 0.001, μ_{AA} = -5., μ_{Aa} = 0., μ_{aa} = 5., and $\sigma_{AA} = \sigma_{Aa} = \sigma_{aa} = 1$ produces the above $\log_{10}$ likelihood. Since all the probands have missing values for QUANTIT, the ascertainment correction equals 1.

D.6.2. Dominance/Displacement (*papfqhw, paptcms, dmlpr0, qmlprdd, papwgml, papenq, papcr, m_6_2*, option 1)

Setting the total mean to -4.99, the total standard deviation to 1.0246710, dominance to 0.5 and displacement to 10 produces the same parameter values and consequently the same $\log_{10}$ likelihood as section D.6.1. Since all the probands have missing values for QUANTIT, the ascertainment correction equals 1.

D.6.3. Means/Standard Deviations/Threshold (*papfqhw, paptcms, dmlpr0, qmlprmvt, papwgml, papendq, papcr, m_6_3*, option 1)

Setting μ_1 = -5., μ_2 = 0., μ_3 = 5., $\sigma_1 = \sigma_2 = \sigma_3 = 1$, and threshold = 5 produces the same $\log_{10}$ likelihood for the pedigrees as in section 6.1. With the threshold equal to μ_3, the affection probability equal 0.5 which when multiplied by the genotype frequency of 0.000001 equals a $\log_{10}$ likelihood for the ascertainment correction of -6.3.

D.6.4. Dominance/Displacement/Proportion (*papfqhw, paptcms, dmlpr0, qmlprddp, papwgml, papenq, papcr, m_6_4*, option 1)

Setting the total mean to -4.99, the total standard deviation to 1.0246710, dominance to 0.5 and displacement to 10 as for D.6.2 produces the same parameter values and consequently the same $\log_{10}$ likelihood as section D.6.1. Setting the proportion to 0.0000005 produces the same ascertainment correction as D.6.3.

D.7. Within-Genotype Parameters

For the within-genotype examples, the genetic model comprises 1 locus with 2 alleles; the variable comprises the quantitative trait QUANTIT. Again the genotype cannot be unequivocally inferred from the phenotype. However, a rare allele frequency and recessive inheritance with a small standard deviation combine to produce a high probability for the assigned genotype for each individual. As in section D.6, the likelihood equals the product of the genotype frequency for each founder, the transmission probability for each nonfounder, and the penetrance for each measured individual. However, the penetrances differ from those in section D.6 because of the correlation between individuals. Let d_1 and d_2 represent the deviations from the genotype mean and s_1 and s_2 represent the within-genotype standard deviations for individuals 1 and 2, and r represent the correlation between them. As before, the penetrance for individual 1 equals $N(d_1, s_1)$ where $N(\delta, \sigma)$ represents the normal density with deviation δ and standard deviation σ. The penetrance for individual 2 equals $N(d_2', s_2')$ where $d_2' = (d_2 - \rho\, d_1)/(1 - r^2)$ and $s_2' = s_2\,(1 - r^2)$. See section VI.12 for extension beyond two individuals.

D.7.1. Exact Variance Components

Heritability h^2 represents the proportion of the within-genotype variance attributed to polygenes and parameter m^2 represents the proportion of the within-genotype variance attributed to a spouse effect. Variable MARITAL identifies spouses through assignment of the same number to both except that in pedigree 4, individual 5 shares a marital effect only with individual 6 and not individual 4.

D.7.1.1. Marital Effect (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgvc, papenqe, papcrqe, m_7_1_1*, option 1)

Setting $h^2 = 0$ and $m^2 = 0.1$, produces a correlation r = 0.1 between spouse pairs (except individuals 4 and 5 in pedigree 4) and r = 0 between all other pairs. The penetrance for each individual equals:

	1	2	3	4
1	N(.5, 1|-.5)	N(.5. 1|-.5)	N(.5, 1|-.5)	N(.5, 1|-.5)
2	N(-.5, 1)	N(-.5, 1)	N(-.5, 1)	N(-.5, 1)
3	N(0, 1)	N(-.5, 1|0)	N(0, 1|0)	N(0, 1)
4	N(.5, 1)	N(0, 1)	N(0, 1)	N(-.5, 1)
5	N(0, 1)	N(0, 1|.5)	N(.5, 1|.5)	N(.5, 1|0)
6	N(-.5, 1)	N(.5, 1)	N(.5, 1)	N(0, 1)
7	N(-.5, 1)	N(0, 1)	N(-.5, 1|0)	N(0, 1)
8	N(-.5, 1)	N(.5, 1)	N(0, 1)	N(.5)
9	N(0, 1)	N(-.5, 1)	N(-.5, 1)	N(-.5, 1)

where $N(\delta,\sigma)$ represents the height of a normal density for deviation δ and standard deviation σ, and $N(\delta,\sigma|\gamma)$ represents the height of a normal density for deviation δ and standard deviation σ after conditioning on deviation γ. Note that the husband in each spouse pair is conditioned on his wife's deviation. Consequently, the $\log_{10}$ likelihood of each pedigree and the total equals:

Pedigree	Penetrance	Frequency/ Transmission	$\log_{10}$ likelihood
1	3 n(0) + 5 n(.5) + n(.5|-.5)	-10.8066	-14.7348
2	2 n(0) + 4 n(.5) + n(.5|-.5) + n(-.5|0) + n(0|.5)	-6.0026	-9.9233
3	2 n(0) + 3 n(.5) + n(-.5|0) + n(.5|.5) + n(0|0) + n(.5|-.5)	-7.2067	-11.1200
4	3 n(0) + 4 n(.5) + n(.5|0) + n(.5|-.5)	-6.6047	-10.5318
Total			-46.3099

where $n(d) = \log_{10} N(d,1)$, $n(d_1|d_2) = \log_{10} N(d_1,1|d_2)$, n(0) = -.3991, n(.5) = -.4534, n(.5|-.5) = -.4639, n(-.5|0) = -.4523, n(0|.5) = -.3975, n(.5|.5) = -.4418, n(0|0) = -.3969, n(.5|0) = -.4523, and the $\log_{10}$ likelihood of the product of the frequencies and transmission probabilities comes from section D.5.1. The small probability of other genotypes than the assigned genotype is ignored in the hand likelihood computation, making it differ slightly from the likelihood computed by *papdr*.

D.7.1.2. Marital Effect and Heritability (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgvc, papenqe, papcrqe, m_7_1_2*, option 1)

Setting $h^2 = 0.2$ and $m^2 = 0.1$ produces correlations between all pairs in each pedigree except individuals 4 and 5 in pedigree 4. To compute the penetrance requires successively conditioning each individual's deviation on all previous individuals' deviations, a tedious calculation by hand. Consequently, a hand-calculated likelihood is not given.

D.7.2. Approximate Variance Components

As in section D.7.1, heritability h^2 represents the proportion of the within-genotype variance attributed to polygenes and parameter m^2 represents the proportion of the within-genotype variance attributed to a spouse effect. Variable MARITAL identifies spouses through assignment of the same number to both. Within-genotype correlations between individuals preclude recursive computation of the likelihood. The exact likelihood equals the sum, over all sets of genotypes, of the probability of each set of genotypes for all pedigree members. Approximation [Hasstedt 1993] allows recursive computation and much greater speed.

D.7.2.1. Marital Effect (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgvc, papenqa, papcrqa, m_7_2_1*, option 1)

Setting $h^2 = 0$ and $m^2 = 0.1$ produces exactly the same $\log_{10}$ likelihood as in section D.7.1.1, despite approximation in the calculation. Only spouses share the within-genotype effect, and recursive calculation considers the genotypes of spouse pairs simultaneously.

D.7.2.2. Marital Effect and Heritability (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgvc, papenqa, papcrqa, m_7_2_2*, option 1)

Setting $h^2 = 0.2$ and $m^2 = 0.1$ produces an approximate likelihood. Compare to the likelihood obtained in D.7.1.2.

D.7.3. Approximate Familial Correlations

The gender-specific familial correlations comprise husband-wife ρ_{hw}, mother-daughter ρ_{md}, mother-son ρ_{ms}, father-daughter ρ_{fd}, father-son ρ_{fs}, sister-sister ρ_{ss}, sister-brother ρ_{sb}, and brother-brother ρ_{bb}. Outside the nuclear family, zero correlation is assumed for all relative pairs.

D.7.3.1. Husband-Wife Correlation (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgfc, papenqa, papcrqa, m_7_3_1*, option 1)

Setting $\rho_{hw} = 0.1$, $\rho_{md} = \rho_{ms} = \rho_{fd} = \rho_{fs} = \rho_{ss} = \rho_{sb} = \rho_{bb} = 0$ produces a likelihood identical to the likelihood from section D.7.1.1 for pedigrees 1 through 3. The likelihood for pedigree 4 differs because this parameterization includes a correlation between individual 5 and both spouses; section D.7.1.2 excluded the correlation between individuals 5 and 4.

D.7.3.2. Spouse, Parent-offspring and Sibling Correlations (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgfc, papenqa, papcrqa, m_7_3_2*, option 1)

Setting $\rho_{hw} = \rho_{md} = \rho_{ms} = \rho_{fd} = \rho_{fs} = \rho_{ss} = \rho_{sb} = \rho_{bb} = 0.1$ produces a likelihood identical to the likelihood from section D.7.2.2 for pedigree 1 since parent-offspring and sibling correlations of 0.1 corresponds to a heritability of 0.2. Pedigrees 2-4 contain more distant relatives for whom this parameterization assumes zero correlation but the variance components parameterization does not.

D.8. Assortative Mating

For the assortative mating examples, the genetic model comprises 1 locus with 2 alleles; the variable comprises the quantitative trait QUANTIT. The model differs from D.7 in that assortative mating alters the genotype frequencies away from Hardy-Weinberg equilibrium and independence between spouse pairs and that that model ignored the marital correlation between spouses 4 and 5 in pedigree 4, but the assortative mating model does not.

When there is assortative mating, the genotype frequencies for a spouse pair are not independent. Rewriting the frequencies and transmissions from D.1 in terms of joint genotype frequencies we obtain:

Pedigree	Frequency/Transmission
1	$f_{23}/2^7$
2	$f_{11}f_{22}/4$
3	$f_{12}^2f_{13}/f_2^2/2^4$
4	$f_{13}^2f_{22}/f_3^2/2^4$

where $f_{ij} = f_{ji}$ represents the joint genotype frequency for a spouse pair with genotypes i and j and f_i represents the marginal frequency of genotype i.

D.8.1. Assortative Mating (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgam, papenqa, papcrqa, m_9_1*, option 1)

Hasstedt [1995] derives a solution for the joint genotype frequencies for spouses assuming assortative mating. For p = 0.999 and $a^2 = 0.1$, the genotype frequencies equal $f_1 = 0.9980536$, $f_2 = 0.0018928$, and $f_3 = 0.0000536$ and the joint genotype frequencies equal $f_{11} = 0.9963004$, $f_{12} = f_{21} = 0.0017099$, $f_{22} = 0.0001733$, $f_{13} = f_{31} = 0.0000433$, $f_{23} = f_{32} = 0.0000097$, $f_{33} = 0.0000005$. Setting $h^2 = 0$, the penetrance terms as in section D.7.1.1. Therefore, for $h^2 = 0$ and $a^2 = 0.1$ and computing the frequencies and transmissions from the expressions above and taking the penetrances from D.7.1.1 produces the following:

Pedigree	Frequency/Transmission	Penetrance	Log_{10} Likelihood
1	-7.1204	-3.9282	-11.0486
2	-4.3649	-3.9207	-8.2856
3	-5.6559	-3.9133	-9.5692
4	-5.0580	-3.9271	-8.9851
Total			-37.8885

The difference for pedigree 4 results because the computation in D.7.1.1 assumed zero correlation between individuals 4 and 5 of pedigree 4.

D.8.2. Assortative Mating and Heritability (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgam, papenqa, papcrqa, m_9_2*, option 1)

Setting $h^2 = 0.2$ and $a^2 = 0.1$, using the frequencies and transmissions from D.8.1 and computing the penetrances by subtracting the frequencies and penetrances from D.7.1.1 from the total log likelihoods from D.7.1.2 produces the following:

Pedigree	Frequency/Transmission	Penetrance	Log_{10} Likelihood
1	-7.1204	-3.8849	-11.0046
2	-4.3649	-3.9203	-8.2852
3	-5.6559	-3.8909	-9.5469
4	-5.0580	-3.9131	-8.9712
Total			-37.8079

Again, the difference for pedigree 4 results because the computation in D.7.1.1 assumed zero correlation between individuals 4 and 5 of pedigree 4.

D.9. Multivariate Parameters

D.9.1. Measured genotype

The variables comprise the quantitative trait QUANTIT and the marker GENOTYPE. Since all the probands have missing values for QUANTIT and GENOTYPE, the ascertainment correction equals 1.

D.9.1.1. 1-locus (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgml, papenq, papcr, m_9_1_1*, option 1)

The genetic model comprises 1 locus with 2 alleles. The genotypes assigned by GENOTYPE are equivalent to the genotypes used in previous quantitative trait likelihood computations. Therefore, see section D.6 for the likelihood computation.

D.9.1.2. 2-locus (*papfqhw, paptcms, dmlpr0, qmlprmv, papwgml, papenq, papcr, m_9_1_2*, option 1)

The genetic model comprises 2 loci each with 2 alleles; QUANTIT is assigned to both loci and GENOTYPE is assigned to locus 1. Means for QUANTIT are assigned to reflect the genotype at locus 1. Therefore, see section D.6 for the likelihood computation.

D.9.1.3. 2-locus (*papfqhw, paptcms, dmlpr0, qmlprdd, papwgml, papenq, papcr, m_9_1_3*, option 1)

The genetic model comprises 2 loci each with 2 alleles; QUANTIT is assigned to both loci and GENOTYPE is assigned to locus 1. For locus 1, the dominance and displacement are assigned as in D.6.2; for locus 2 the dominance and displacement are set equal to zero. Therefore, see section D.6 for the likelihood computation.

D.9.2. Bivariate

The variables comprise the quantitative trait QUANTIT and the discrete trait DISEASE. Parameter ρ represents the correlation between QUANTIT and DISEASE. Since each proband is affected, the ascertainment correction equals the disease frequency or 10^{-6}.

D.9.2.1. 1-locus (*papfqhw, paptcms, dmlprpn, qmlprmv, papwgml, papendq, papcr, m_9_2_1*, option 1)

The genetic model comprises 1 autosomal locus with 2 alleles. The genotypes assigned by DISEASE are equivalent to the genotypes used in previous quantitative trait likelihood computations and $\rho = 0$. Therefore, see section D.6 for the likelihood computation.

D.9.2.2. 2-locus (*papfqhw, paptcms, dmlprpn, qmlprmv, papwgml, papendq, papcr, m_9_2_2*, option 1)

The genetic model comprises 2 autosomal loci with 2 alleles each; QUANTIT is assigned to both loci and DISEASE is assigned to locus 1. The genotypes assigned by DISEASE are equivalent to the genotypes used in previous quantitative trait likelihood computations, the means for QUANTIT correspond to locus 1 genotypes, and $\rho = 0$. Therefore, see section D.6 for the likelihood computation.

D.10. Genotype-Assignment Codes

For the genotype-assignment code examples, the genetic model comprises 1 or 2 loci with 2 alleles; the variable comprises the discrete trait DISEASE.

D.10.1. Autosomal (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_1*, option 1)

See section D.5.1 for the likelihood computation.

D.10.2. X-linked (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_2*, option 1)

For p = allele frequency, q = 1 - p, the probability by individual and pedigree equals:

Individual	Pedigree			
	1	2	3	4
1	q	p	q	p
2	2pq	2pq	p^2	2pq
3	½	p	2pq	½
4	½	p^2	1	p^2
5	½	½	1	½
6	½	1	p^2	p^2
7	½	1	½	1
8	½	1	1	1
9	½	1	½	1
Total	$pq^2/2^6$	p^4q	$p^5q^2/2$	$p^6q/2$

Since p = 0.999, q = 0.001, the ascertainment correction = -3 for each pedigree, and the $\log_{10}$ likelihood by pedigree equals:

Pedigree	Log_{10} Likelihood	
	Uncorrected	Corrected
1	-7.8066	-4.8066
2	-3.0017	-0.0017
3	-6.3032	-3.3032
4	-3.3036	-0.3036
Total	-20.4152	-8.4152

D.10.3. Parent-specific autosomal

D.10.3.1. Equal penetrances (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_3_1*, option 1)

For equal penetrance in heterozygotes regardless of the parental origin of each allele, the model is equivalent to the standard autosomal model. See section D.10.1 for the likelihood computation.

D.10.3.2. Unequal penetrances (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_3_2*, option 1)

For pen(Aa) = 1, pen(aA) = 0, where the alleles are ordered as maternal/paternal origin, the likelihood by individual and pedigree equals:

Individual	Pedigree			
	1	2	3	4
1	pq	p^2	pq	p^2
2	pq	pq	p^2	pq
3	½	p^2	p^2	½
4	½	p^2	½	p^2
5	½	½	1	1/2
6	½	1	p^2	p^2
7	½	1	½	1
8	½	1	1	1
9	½	1	½	1
Total	$p^2q^2/2^7$	$p^7q/2$	$p^7q/2^3$	$p^7q/2^2$

For p = 0.999, q = 0.001, the ascertainment correction = -3, and the $\log_{10}$ likelihood by pedigree equals:

Pedigree	Log_{10} Likelihood	
	Uncorrected	Corrected
1	-8.1081	-5.1081
2	-3.3041	-0.3041
3	-3.9061	-0.9061
4	-3.6051	-0.6051
Total	-18.9234	-6.9234

D.10.4. X-linked parental origin

D.10.4.1. Equal penetrances (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_4_1*, option 1)

For equal penetrance in heterozygotes regardless of the parental origin of each allele, the model is equivalent to the standard X-linked model. See section D.10.2 for the likelihood computation.

D.10.4.2. Unequal penetrances (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_4_2*, option 1)

For pen(Aa) = 1, pen(aA) = 0, where the alleles are ordered as maternal/paternal origin, the likelihood by individual and pedigree equals:

Individual	Pedigree			
	1	2	3	4
1	q	p	q	p
2	pq	pq	p^2	pq
3	½	p	p^2	½
4	½	p^2	1	p^2
5	½	½	1	½
6	½	1	p^2	p^2
7	½	1	1	1
8	½	1	1	1
9	½	1	0	1
Total	$pq^2/2^7$	$p^5q/2$	0	$p^6q/2^2$

For p = 0.999, q = 1 - p, the ascertainment correction = -3 for each pedigree, and the $\log_{10}$ likelihood by pedigree equals:

Pedigree	Log_{10} Likelihood	
	Uncorrected	Corrected
1	-8.1076	-5.1076
2	-3.3032	-0.3032
3	-∞	-∞
4	-3.6047	-0.6047
Total	-∞	-∞

The model is impossible in pedigree 3 because: **(1)** for 9 to be affected he must receive the allele from his mother 7; **(2)** for 7 to be unaffected she must receive the allele from her father 4; **(3)** for 4 to be unaffected, he must not have the allele.

D.10.5. Category-specific autosomal inheritance

D.10.5.1. Equal penetrances (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_5_1*, option 1)

For equal affection probabilities regardless of sex, the model is equivalent to the standard autosomal model. See section D.10.1 for the likelihood computation.

D.10.5.2. Unequal penetrances (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_5_2*, option 1)

For pen(Aa) = 1 in males, pen(Aa) = 0 in females, the likelihood by individual and pedigree equals:

Individual	Pedigree			
	1	2	3	4
1	1	p^2	2pq	p^2
2	2pq	2pq		2pq
3	2pq	p^2	2pq	½
4	¼	p^2	½	p^2
5	¾	½	½	½
6	¼	1	p^2	p^2
7	¼	1	½	1
8	¼	½	1	1
9	¼	1	1/2	1
Total	$p^2q^2 3^2/4^6$	$p^7q/2$	$p^6q^2/2^2$	$p^7q/2$

For p = 0.999, q = 1 - p, the ascertainment correction equals $1 - p^2 = -2.6992$, and the $\log_{10}$ likelihood by pedigree equals:

Pedigree	Log_{10} Likelihood	
	Uncorrected	Corrected
1	-8.6590	-5.9598
2	-3.3041	-0.6049
3	-6.6047	-3.9055
4	-3.3041	-0.6049
Total	-21.8718	-11.0751

D.10.6. Autosomal/X-linked mixed model

In the autosomal/X-linked mixed model, τ_f represents father-daughter transmission and τ_m represents father-son transmission.

D.10.6.1. Autosomal (*papfqxa paptctp, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_6_1*, option 1)

For $\tau_m = \tau_f = 0.5$, the model is equivalent to the standard autosomal model. See section D.10.1 for the likelihood computation.

D.10.6.2. X-linked (*papfqxa paptctp, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_6_2*, option 1)

For $\tau_f = 0$, $\tau_m = 1$, the model is equivalent to the standard X-linked model. See section D.10.2 for the likelihood computation.

D.10.6.3. Mixed (*papfqxa paptctp, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_6_3*, option 1)

For $\tau_f = 0.5$, $\tau_m = 1$, the likelihood by individual and pedigree equals:

Individual	Pedigree			
	1	2	3	4
1	q	p	q	p
2	2pq	2pq	p^2	2pq
3	½	p	p^2	½
4	½	p^2	1	p^2
5	½	½	1	½
6	½	½ ½	p^2	p^2
7	½	½ 1	½	1
8	½	½ 1	1	1
9	½	½ 1	½	1
Total	$pq^2/2^6$	$p^5q/2(1+1/2^3)$	$p^6q/4$	$p^6q/2$

For p = 0.999, q = 1 - p, the ascertainment correction = -3, and the $\log_{10}$ likelihood by pedigree and total equals:

Pedigree	Log_{10} Likelihood	
	Uncorrected	Corrected
1	-7.8066	-4.8066
2	-3.2521	-0.2521
3	-3.6047	-0.6047
4	-3.3036	-0.3036
Total	-17.9670	-5.9670

D.10.7. Autosomal/X-linked admixture

In the autosomal/X-linked admixture model, q_x represents the frequency of the X chromosome allele and q_a represents the frequency of the autosomal allele.

D.10.7.1. Autosomal allele only (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_7_1*, option 1)

For $q_x = 0$, the model is equivalent to the standard autosomal model. See section D.10.1 for the likelihood computation.

D.10.7.2. X-linked allele only (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_7_2*, option 1)

For $q_a = 0$, the model is equivalent to the standard X-linked model. See section D.10.2 for the likelihood computation.

D.10.7.3. Both alleles (*papfqhw paptcms, dmlprpn, qmlpr0, papwgml, papend, papcr, m_10_7_3*, option 1)

The likelihood of the admixture model can be computed by summing the likelihood of each set of genotypes for affected individuals. In pedigrees 1, 2, and 4, only exclusively autosomal or exclusively X-linked genotypes require fewer than three alleles. Pedigree 3 also requires only two disease alleles if individual 1 has the X-linked form and individual 9 has the autosomal form. The approximate likelihood by pedigree equals:

Pedigree	Autosomal	X-Linked	X/Autosomal	Sum by Pedigree
1	$10^{-10.8066}$	$10^{-7.8066}$		$10^{-7.8062}$
2	$10^{-6.0026}$	$10^{-3.0017}$		$10^{-3.0013}$
3	$10^{-7.2048}$	$10^{-6.3032}$	$10^{-7.2098}$	$10^{-6.2519}$
4	$10^{-6.6047}$	$10^{-3.3036}$		$10^{-3.3034}$
Total	$_{10}$-30.6170	$_{10}$-20.4152		$_{10}$-20.4152

E. ERROR MESSAGES

Four subroutines, *filerr*, *daterr*, *moderr*, and *paperr*, output error messages to the console, then terminate execution. Subroutine *paperr* also outputs its error messages to *pap.out*.

E.1. Errors in the Files: *filerr*

Subroutine *filerr*, called by programs *preped* (section III.1), *descstat*, (section III.2), *prepap* (section III.3), *simrk* (section III.4), *simul* (section III.5) *papdr* (section III.6), and *gpe* (section III.7) reports errors in the files in opening, reading, or premature end of file. The error messages follow:

1. Error opening *trip.dat*.
2. Error reading *trip.dat*.
3. Error reading *order.dat*.
4. Error opening *header.dat*.
5. Error reading *header.dat*.
6. End of file on *header.dat*.
7. Error opening *phen.dat*.
8. Error reading *phen.dat*.
9. Error opening *ascer.dat*.
10. Error reading *ascer.dat*.
11. Error opening *papin.dat*.
12. Error reading *papin.dat*.
13. End of file on *papin.dat*.
14. Error opening *freq.dat*.
15. Error reading *freq.dat*.
16. End of file on *freq.dat*.
17. Error opening *tran.dat*.
18. Error reading *tran.dat*.
19. End of file on *tran.dat*.
20. Error opening *dmlp.dat*.
21. Error reading *dmlp.dat*.
22. End of file on *dmlp.dat*.

23. Error opening *qmlp.dat*.

24. Error reading *qmlp.dat*.

25. End of file on *qmlp.dat*.

26. Error opening *wgen.dat*.

27. Error reading *wgen.dat*.

28. End of file on *wgen.dat*.

29. Error opening *popln.dat*.

30. Error reading *popln.dat*.

31. End of file on *popln.dat*.

32. Error opening *model.dat*.

33. Error reading *model.dat*.

34. End of file on *model.dat*.

35. Error opening *prob.dat*.

36. Error reading *prob.dat*.

E.2. Errors in the Data Files: *daterr*

Subroutine *daterr*, called by programs *preped* (section III.1), *descstat* (section III.2), *prepap* (section III.3), *simrk* (section III.4), *simul* (section III.5), *papdr* (section III.6), and *gpe* (section III.7) reports errors in either the pedigree structure or the amount of storage allocated for the data. The pedigree structure errors require correction of *trip.dat* (section II.1 and B.1). The storage errors require increasing the indicated parameter value in the include files and recompiling. The error messages follow:

37. The number of individuals with phenotype data exceeds MINDPH. Increase MINDPH and recompile.

38. Invalid ID number in *trip.dat*.

39. Individual own parent. Correct *trip.dat*.

40. The number of pedigrees exceeds MPEDGR. Increase MPEDGR and recompile.

41. The number of triplets exceeds MTRIPS. Increase MTRIPS and recompile.

42. File *trip.dat* has not been sorted.

43. Individual duplicated in *trip.dat*. Correct *trip.dat*.

44. Individual occurs as an offspring more than once. Correct *trip.dat*.

45. Individual missing from *trip.dat*.

46. The number of variables exceeds MDATA. Increase MDATA and recompile.

47. Invalid type code for variable. Correct *header.dat*.

48. The number of columns exceeds MCOL. Increase MCOL and compile.

49. Marker missing allele count. Enter number of alleles in *header.dat*.

50. NCOL and NDATA inconsistent. Correct *header.dat*.

51. Marker coded as autosomal in *header.dat* and X-linked in *popln.dat*.

52. Marker coded as X-linked in *header.dat* and autosomal in *popln.dat*.

53. Invalid code. Correct *header.dat*.

54. The standard deviation for variable equals zero. Correct *header.dat*.

55. Individual has an invalid allele code. Correct *phen.dat*.

56. The number of categories exceeds MCATEG. Increase MCATEG and recompile.

57. Marker phenotype allows all alleles. Genotype collapsing will not work.

58. Marker genotype corresponds to no phenotype. Correct *popln.dat*.

59. Marker genotype corresponds to more than one phenotype. Correct *popln.dat*.

60. The number of ancestors exceeds MFMANC. Increase MFMANC and recompile.

61. The number in the nuclear family exceeds MFAMEM. Increase MFAMEM and recompile.

62. The number of nuclear families in the sample exceeds MFAMTL. Increase MFAMTL and recompile.

63. The number of nuclear families in a pedigree exceeds MFAMPD. Increase MFAMPD and recompile.

64. Pedigree missing from *order.dat*.

65. The number of offspring requiring joint likelihoods exceeds MCHJT. Increase MCHJT and recompile.

66. The number in one pedigree exceeds MINDPD. Increase MINDPD and recompile.

67. The family is never used. Correct *trip.dat*.

68. The number of individuals exceeds MINDT. Increase MINDT and recompile.

69. The gender designation is incorrect for individual. Correct *phen.dat*.

70. The number of spouses exceeds MSPOUS. Increase MSPOUS and recompile.

71. Number of marker genotypes too low for marker. Correct *popln.dat*.

72. Number of marker phenotypes too low for marker. Correct *popln.dat*.

73. Variable not standardized as necessary for power transformation. Standardize in *header.dat*.

74. Genotype probabilities do not add to one for individual. Correct number of genotypes or *prob.dat*.

E.3. Errors in the Genetic Model: *moderr*

Subroutine *moderr*, called by *descstat* (section III.2), *prepap* (section III.3), *simrk* (section III.4), *simul* (section III.5), and *papdr* (section III.6) reports errors in either the genetic model and parameter values or in the amount of storage allocated. Correct errors in the genetic model or parameter values by executing *prepap*. Correct errors in storage by increasing the parameter values in included file. The error messages follow:

75. The number of variables included in the model exceeds MVAR. Increase MVAR and recompile.

76. Invalid genotype assignment code. Execute *prepap* it write *model.dat* correctly.

77. The number of loci exceeds MLOCI. Increase MLOCI and recompile.

78. An X-linked locus is not allowed for the genotype assignment code selected. Execute *prepap* to write *model.dat* correctly.

79. The number of markers included in the model exceed MMARKR. Increase MMARKR and recompile.

80. The number of variables included in the model exceeds MTRAIT. Increase MTRAIT and recompile.

81. Location of the marker in the variable list is less than 1 or greater than NVAR. Execute *prepap* to write *model.dat* correctly.

82. Locus assigned for the marker is less than 1 or greater than NLOCI. Execute *prepap* to write *model.dat* correctly.

83. Location of the marker in phen.dat is less than 1 or greater than NDATA. Execute *prepap* to write *model.dat* correctly.

84. Location of the trait in the variable list is less than 1 or greater than NVAR. Execute *prepap* to write *model.dat* correctly.

85. Values input for the penetrance paramter are too few or too many. Execute *prepap* to write *model.dat* correctly.

86. The number of frequency sets exceeds MFREQS. Increase MFREQS and recompile.

87. The number of parameters required for the frequency exceeds MRFPR. Increase MFRPR and recompile.

88. The number of values input for the frequencies exceeds MFREQT. Increase MFREQT and recompile.

89. The mode of inheritance of the marker does not match the mode of inheritance of the locus. Execute *prepap* to write *model.dat* correctly.

90. The number of marker alleles in *popln.dat* does not equal the number in *model.dat*. Execute *prepap* to write *model.dat* correctly.

91. Frequency count number is less than zero. Execute *prepap* to write *model.da*t correctly.

92. Frequency is less than zero or greater than one.

93. The number of frequencies does not equal the number required by this model and subroutine. Execute *prepap* to write *model.dat* correctly.

94. Female recombination probability is incorrect. Execute *prepap* to write *model.dat* correctly.

95. Male recombination probability is incorrect. Execute *prepap* to write *model.dat* correctly.

96. The number of transmission probabilities exceeds MXTRAN. Increase MXTRAN and recompile.

97. Daughters transmission probability is incorrect. Execute *prepap* to write *model.dat* correctly.

98. Daughters transmission probability is less than zero or greater than one. Execute *prepap* to write *model.dat* correctly.

99. Sons transmission probability is incorrect. Execute *prepap* to write *model.dat* correctly.

100. Sons transmission probability is less than zero or greater than one. Execute *prepap* to write *model.dat* correctly.

101. The number of parameters for maximization, standard errors, or gridding exceeds MPARAM. Increase MPARAM and recompile.

102. The number of parameters held to the same value exceeds MPARSM. Increase MPARSM and recompile.

103. Parameter is already be maximized.

104. The number of linear constraints exceeds MCLIN. Increase MCLIN and recompile.

105. The number of haplotype exceeds MHAPTL. Increase MHAPTL and recompile.

106. The number of genotypes exceeds MGENTL. Increase MGENTL and recompile.

107. The number of trait genotypes exceeds MGTRAT. Increase MGTRAT and recompile.

108. The number of available subroutines exceeds MSUBRT. Increase MSUBRT and recompile.

109. The frequency subroutine is inappropriate for the model as specified. Execute *prepap* to write *model.dat* correctly.

110. The transmission subroutine is inappropriate for the model as specified. Execute *prepap* to write *model.dat* correctly.

111. The major locus discrete subroutine is inappropriate for the model as specified. Execute *prepap* to write *model.dat* correctly.

112. The major locus quantitative subroutine is inappropriate for the model as specified. Execute *prepap* to write *model.dat* correctly.

113. The within genotype subroutine is inappropriate for the model as specified. Execute *prepap* to write *model.dat* correctly.

114. No frequency subroutines are available for the model as specified. Execute *prepap* to write *model.dat* correctly.

115. No transmission subroutines are available for the model as specified. Execute *prepap* to write *model.dat* correctly.

116. No major locus discrete subroutines are available for the model as specified. Execute *prepap* to write *model.dat* correctly.

117. No major locus quantitative subroutines are available for the model as specified. Execute *prepap* to write *model.dat* correctly.

118. No within genotype subroutines are available for the model as specified. Execute *prepap* to write *model.dat* correctly.

119. The number of frequency types exceeds MFRPR. Increase MFRPR and recompile.

120. The number of concomitant types required for the penetrance exceeds MCNTP. Increase MCNTP and recompile.

121. The number of parameters required for the penetrance exceeds MPRTPD. Increase MPRTPD and recompile.

122. The number of parameters required for the penetrance exceeds MPRTPQ. Increase MPRTPQ and recompile.

123. The number of parameters required for the penetrance exceeds MPRTPW. Increase MPRTPW and recompile.

124. The number of parameters required for the penetrance exceeds MPRD. Increase MPRD and recompile.

125. The number of parameters required for the penetrance exceeds MPRQ. Increase MPRQ and recompile.

126. The number of parameters required for the penetrance exceeds MPRW. Increase MPRW and recompile.

127. Program simrk does not simulate traits. Run *prepap* to rewrite *model.dat*.

128. The number of marker alleles in *model.dat* is incorrect. Execute *prepap* to write *model.dat* correctly.

129. Lower bound larger than upper bound for parameter.

130. The number of parameters to maximize, compute standard error or grid exceed MPARAM. Increase MPARAM and recompile. Program *prepap* will allow continued parameter input.

131. The parameter values to not conform to linear constraints. Correct *model.dat*.

132. The number of parameters required for the penetrance exceeds MPRTP. Increase MPRTP and recompile.

133. The number of parameters required for the penetrance exceeds MPR. Increase MPR and recompile.

E.4. Execution Errors: *simerr*

Subroutine *simerr*, called by *simrk* (section III.4), and *simul* (section III.5) reports simulation execution errors. The errors include the linkage of subroutines inappropriate for *model.dat*, data and parameter values wrong for the subroutine, arrays too small for the data/genetic model combination, and execution errors. The error messages follow:

134. Program *simrk* only simulates markers. Run *prepap* to rewrite *model.dat*.

135. Program *simrk* does not simulate non-codominant markers. Run *prepap* to rewrite *model.dat*.

136. Program *simul* does not allow parameter estimation. Run *prepap* to rewrite *model.dat* without parameter estimation.

137. Assortative mating models cannot be simulated.

138. This analysis must be performed using *papdr*.

139. The frequency subroutine linked is not the one for which *model.dat* was written.

140. The transmission subroutine linked is not the one for which *model.dat* was written.

141. The major locus discrete subroutine linked is not the one for which *model.dat* was written.

142. The major locus quantitative subroutine linked is not the one for which *model.dat* was written.

143. The within genotype subroutine linked is not the one for which *model.dat* was written.

144. The penetrance subroutine linked is not the one for which *model.dat* was written.

145. The correlated penetrance subroutine linked is not the one for which *model.dat* was written.

146. Correlation impossible for individual.

147. Frequencies sum to more than one.

148. The number of genotypes possible for offspring exceeds MGENTP. Increase MGENTP and recompile.

149. At least one transmission probability must not be estimated. Run prepap to rewrite *model.dat*, leaving one or more fixed.

150. Individual has an age out of range.

151. Individual is missing age.

152. Individual has an invalid disease code.

153. The prevalence exceeds 1 for trait. Correct *popln.dat*.

154. Classification missing for pedigree member.

155. Individual has classification out of range.

156. The prevalence is not increasing.

157. Individual is missing a covariate.

158. Assortative mating is restricted to autosomal inheritance.

159. Assortative mating is restricted to 1- or 2-allele models.

160. Assortative mating is restricted to one trait.

161. Assortative mating requires that the means increase across genotypes.

162. Assortative mating requires equal variances for all genotypes.

163. The number of measurement exceeds MMEAS. Increase MMEAS and recompile.

164. Age and onset age must be in separate columns when simulating phenotypes.

165. Chromosome length of zero.

166. Recombination probability greater than 0.5 between loci.

167. Location past end of chromosome in males.

168. Location past end of chromosome in females.

169. Category ranges are not increasing. Correct *popln.dat*.

E.5. Execution Errors: *paperr*

Subroutine *paperr*, called by *simrk* (section III.4), *simul* (section III.5), and *papdr* (section III.6), reports execution errors. The errors include the linkage of subroutines inappropriate for *model.dat*, data and parameter values wrong for the subroutine, arrays too small for the data/genetic model combination, and execution errors. The error messages follow:

170. The number of alleles for the marker exceeds MMKALL. Increase MMKALL and recompile.

171. Marker phenotype exceeds the number of phenotypes.

172. No marker genotypes possible for individual.

173. The number of possible genotype combinations exceeds MADPH. Increase MADPH and recompile.

174. The number of genotypes to compute probabilities for exceeds MGPRB. Increase MGPRB and recompile.

175. The input value of the parameter is not the maximum likelihood estimate.

176. No genotypes possible for individual.

177. Gender or other discrete categorization is missing for pedigree member.

178. The number of stored likelihoods exceeds MPRLST. Increase MPRLST and recompile.

179. The cutset size is larger than MCUTST. Increase MCUTST and recompile.

180. File papin.dat is not appropriate for assortative mating. Rerun preped.

181. The initial values produce a zero likelihood.

F. CHANGES FROM VERSION 4

F.1. Input Files

The user-supplied input files have not changed from version 4. Files *papin.dat* and *model.dat* have changed; before using version 5 of PAP on existing data files, you will need to run program *preped* to create new versions of *papin.dat* and run program *prepap* to create new versions of *model.dat*.

F.2. Programs

Program *simrk* is the only new program. It rapidly simulates marker genotypes by chromosome using the method of Terwilliger et al [1993]. See section III.4 for more information on *simrk* and Elbein et al [1999] for an example of using *simrk* to simulate genome scan data to obtain the lod score distribution under the assumption of no linkage to obtain a significance level for a lod score.

F.3. Genetic Models

The mixed model has been extended to include assortative mating. The assortative mating model is limited to univariate single-locus analysis. Program *preped* writes a different version of *papin.dat* when assortative mating is used. See section D.8, Hasstedt [1995], and Hasstedt et al [1995] for more information and examples of the use of the assortative mating model.

The transmission disequilibrium test is available for application to pedigree data. Any number of marker allles may be included. See section VI.9 and Bovill et al [2000] for more information and an example of the use of the transmission disequilibrium test.

Multi-locus models may assume additivity across loci when using any penetrance subroutine parameterized with dominance and displacement. Displacement for a multi-locus genotype is assumed to equal the sum of the displacements of the contributing loci. This reduces the number of parameters in the model. See Hasstedt et al [1997] and Hasstedt et al [1998] for locus heterogeneity and epistasis, respectively, examples of the use of the model.

Subroutine *dmlprinc* allows pedigree members to be categorized according to a trait that affects the disease incidence. File *popln.dat* repeats the age ranges across the categories. For example, diabetes incidence may be entered separately for non-obese and obese individuals. The appropriate incidence is used for each individual according to obesity status. Dominance and displacement are assumed equal for all categories. Alternatively, the use of genotype coding option 5 allows for two categories with different dominance and displacement. Another alternative is to code as different traits by category and perform a bivariate analysis, fixing the correlation at 1.

F.4. Options

Program *papdr* Version 5 still has eight execution options, but they differ from Version 4. Option 7, expected lod scores, was eliminated since option 6, simulation and estimation, performs the same function. Option 4, search over parameter values, was added, and option 4-6 shifted to 5-7. See section III.6 for more information on the execution option in *papdr*.

Lod scores may be computed for any multi-locus model for which the recombination fraction is a parameter. There cannot be an ascertainment correction for the lod score to be computed.

Any time you choose to maximize when running program *papdr*, you are given the option of having starting values selected at random. Random starting values are selected from a uniform distribution across the range between the lower

and upper bounds given in *model.dat*. Therefore, before using this option, it is necessary to specify boundary values that are reasonable for the parameters.

F.5. Modifications

GEMINI has been modified to perform better when the maximum is in the vicinity of a boundary.