SIB−PAIR manual

SIB−PAIR 1.00a17 (21 Jun 2006)

by

David L. Duffy

(1995−2006)

A program for elementary genetical analyses

David L. Duffy, MBBS PhD.
Queensland Institute of Medical Research,
300 Herston Road,
Herston, Queensland 4029, Australia.
Email: davidD@qimr.edu.au

CONTENTS

INTRODUCTION

Program Sib−pair performs a number of simple analyses of family data that tend to be "nonparametric" or "robust" in nature. It is modelled to some extent on the Genetic Analysis System [Young, 1995] in terms of the command language and types of analysis. Included are routines for:

- Imputation of genotypes where unequivocal.
- Estimation of allele frequencies in codominant genetic systems.
- Simple and complex segregation analysis of a binary trait.
- Estimation of familial correlations and sibship variances for a quantitative trait. Variance components analysis of quantitative and binary traits using a variety of likelihoods.
- Haseman−Elston sib−pair regression of a quantitative (or binary) trait using full and half−sib data, and variance components linkage analysis for normally distributed quantitative traits.
- Carrying out multiple versions of the transmission−disequilibrium test.
- Testing allelic association with a binary or quantitative trait −− Monte Carlo simulation of null distribution of simple tests, or now "measured genotype" familial analysis including combined association and linkage analysis.
- Single locus Affected Pedigree Method identity−by−state and identity−by−descent linkage analysis. This includes Wards [1993] extensions to include unaffected pedigree members.
- Writing out of locus and pedigree files in the formats used by the programs APM, Arlequin, ASPEX, CRIMAP, FISHER, GAS, GDA, Genehunter, LINKAGE, LOKI, MENDEL, MERLIN, PAP and

SAGE.

An example of use

Sib−pair is command line oriented, and writes output only to the standard output (the screen if you are using it interactively). Therefore output can be saved to a file only in batch mode, or via another program that can copy from the standard output, such as *tee*, or the Tcl/Tk based GUI program *isp* which can found on the same site as Sib−pair itself.

Bearing this in mind, here is a sample interactive session. We start at the command line of your operating system shell:

```
> sib−pair
```

|||| SIB−PAIR: A program for simple genetic analysis
|\/| Version : Version 1.00a16 (01−Jul−2006)
|/\| Author : David L Duffy (c) 1995−2006
|||| Job run : Thu Jun 1 15:28:03 2006

>>

A double arrow command prompt appears. We read in a previously prepared script:

>> include ex.in

The contents of ex.in are a series of Sib−pair commands:

```
# declare four loci
set loc a affection
set loc b quantitative
set loc m1 marker 0.0 cM
set loc m2 marker 5.1 cM
# read the pedigree data
read ped inline
ex1 1a x x m n x 1 3 1 2
ex1 1b x x f n x 1 2 3 4
ex1 2a 1a 1b m n 3.5 1 2 1 3
ex1 2b x x f n 1.1 2 2 2 3
ex1 3a 2a 2b m y 4.3 1 2 1 2
ex1 3b 2a 2b m n 2.0 2 2 2 3
ex1 3c 2a 2b f n 0.8 2 2 3 3
ex1 4a 3c 3d f y x 1 2 2 3
ex1 3d x x m n x x x x
ex1 4b 3b 3e m y 4.7 1 2 3 4
ex1 4c 3b 3f m n 1.6 2 2 1 3
;;;;
# The four semicolons ends the in−line data
run
```

The "run" command actually starts the initial processing of the pedigree.

```
NOTE:  Imputation level 1.  Imputing untyped parental
        genotypes where unequivocal.

Pedigree file        = inline.ped
Number of loci       =     4

Locus      Type Position
---------- ---- --------
a           a     6
b           q     7
m1          m     8--  9
m2          m    10-- 11

Number of marker loci=    2
Bonferroni corr. 5%  =    0.025321
Bonferroni corr. 1%  =    0.005013
Bonferroni corr. 0.1%=    0.000500



NOTE:  Creating dummy record for ex1−3e.
NOTE:  Creating dummy record for ex1−3f.

NOTE: Father and mother of person ex1−4a appear to be swapped around.  Reordering.
NOTE:  Person ex1−3e appears as a mother and sex was unspecified.
       Setting sex to female.
NOTE:  Person ex1−3f appears as a mother and sex was unspecified.
       Setting sex to female.
Pedigree: ex1        No. members:   13 No. founders:    6 No. sibships:    5



Total number of pedigrees  =     1
Number with only 1 member  =     0
Total number of sibships   =     5
Total number of subjects   =    13
Total subjects genotyped   =    10 ( 76.9%)
Total number of genotypes  =     20
Largest pedigree (members) =    13 (Pedigree ex1)
Deepest pedigree (genrtns) =     4 (Pedigree ex1)

Mean size of pedigrees     =    13.0
Mean pedigree depth        =     4.0
Mean size where >1 members =    13.0
Mean depth where >1 geners =     4.0
Number of imputed genotypes=     0

Number of pedigree errors  =     0
Number of deleted records  =     0
```

We obtain a few routine messages and summary statistics. The small table of Bonferroni corrections is a reminder that Sib−pair does not usually correct P−values for multiple tests; it is up to the user to decide what the appropriate significance thresholds are.

Blank records are created for named but missing parents. This is necessary, as a formal pedigree should have neither or both parents present for each person.

>> ls

a* b* m1 m2

The "ls" commands shows trait loci (with an asterisk appended), and marker loci.

>> drop $m
>> ls
>> undelete

a* b* (m1) (m2)

The "drop" commands drops loci from the scope of subsequent commands, while the "undelete" command returns access to all loci.

>> describe m1

```
------------------------------------------------
Allele frequencies for locus "m1        "
------------------------------------------------
   Allele  Frequency   Count  Histogram
       1     0.3000        6  ******
       2     0.6500       13  *************
       3     0.0500        1  *

Number of alleles    =    3
Heterozygosity (Hu)  =    0.5105
Poly. Inf. Content   =    0.4064
Number persons typed =   10 ( 76.9%)
```

>> describe snp

```
Marker          NAll  Allele(s)    Freq   Het    Ntyped
--------------  ----  ----------   ------ ------ ------
m1                 3    1 ..   3    −     0.5105    10
m2                 4    1 ..   4    −     0.7211    10
```

The "describe" command gives summary information about loci. For marker loci, it tabulates simple allele counts and proportions in the dataset. Given the keyword "snp", it gives a summary for all markers, one line per locus. For traits, "describe" gives familial correlations or recurrence risks. The "tabulate" gives simpler tables:

>> tab a

```
a          x:   2      y:   3      n:   8
```

>> tab a m1

```
------------------------------------------------
Cross-tabulation of "a         " ... "m1        "
------------------------------------------------

                     m1
a                 1/2          1/3          2/2
          ----------------------------------------
     n         2 (.286)     1 (.143)     4 (.571)
     y         3 (1.00)     0 (.000)     0 (.000)

    No. complete observations =    10
    LR contingency chi-square =   5.5
           Degrees of freedom =   2
                      P-value =0.0643
```

There are a few other simple descriptive commands. The "count" command gives information about families and individuals:

>> count b>1

```
Count where "b > 1":
<Pedigree   Con=T   Num  ASPs Trios    4+
---------- ----- ----- ----- ----- -----
ex1            6    13     1     0     0
Total          6    13     1     0     0
```

The "select" command selects pedigrees containing a specified number (or one or more) individuals meeting the criterion. The "print" command is individual oriented:

>> print b>1

```
Print where "b > 1":

ped=ex1 id=2b fa=x mo=x  sex=f b=1.1000 m1=2/2 m2=2/3
ped=ex1 id=4c fa=3b mo=3f sex=m b=1.6000 m1=2/2 m2=1/3
ped=ex1 id=4b fa=3b mo=3e sex=m b=4.7000 m1=1/2 m2=3/4
ped=ex1 id=3a fa=2a mo=2b sex=m b=4.3000 m1=1/2 m2=1/2
ped=ex1 id=3b fa=2a mo=2b sex=m b=2.0000 m1=2/2 m2=2/3
ped=ex1 id=2a fa=1a mo=1b sex=m b=3.5000 m1=1/2 m2=1/3

Number of matched persons   =    6 out of    13 ( 46.2%)
Number of matched pedigrees =    1 out of     1 (100.0%)
```

If we had instead issued:

>> keep $m
>> print b>1

we would obtain:

```
Print where "b > 1":

ped=ex1 id=2b fa=x mo=x  sex=f m1=2/2 m2=2/3
ped=ex1 id=4c fa=3b mo=3f sex=m m1=2/2 m2=1/3
...
```

As of 2006−Mar−01, we can write expressions containing genotypes:

>> undelete
>> print m2=="3/4"

```
Print where "m2 = = 3/4":

ped=ex1 id=1b fa=x mo=x  sex=f a=n b=x m1=1/2 m2=3/4
ped=ex1 id=4b fa=3b mo=3e sex=m a=y b=4.7000 m1=1/2 m2=3/4
```

The "associate" command gives results from family based tests of allelic association for a trait versus all active marker loci:

>> ass a

```
------------------------------------------------
Allelic association testing for trait "a         "
------------------------------------------------

Marker     Typed  Allels Chi−square Asy P  Emp P  Iters
---------- ------ ------ ---------- ------ ------ ------
m1             10      3        1.9 0.3930 0.1575    127 AssX2−HWE .
m1              1      3        1.1 0.5978 0.5978     23 RC−TDT    .
m2             10      4        0.9 0.8192 0.7692     26 AssX2−HWE .
m2              1      4        2.1 0.4471 0.4471    160 RC−TDT    .
```

The "sibpair" command gives results from regression based tests of linkage for a trait versus all active marker loci. The P−values for these tests can be "empirically" estimated by gene−dropping marker alleles under the null hypothesis of no linkage:

>> sib b simulate

```
-----------------------------------------------------------
Sham S+D H−E for trait "b         " v. all markers
-----------------------------------------------------------

Marker     FSibs  HSibs  t−value    Asy P  Emp P  Iters
---------- ------ ------ ---------- ------ ------ ------
m1              3      1        2.6 0.1180 0.0288    695 H−E +
m2              3      1        1.5 0.1908 0.2128     94 H−E .
```

Finally, we log transform the quantitative trait values and write out a Genehunter type pedigree file, so we can

further examine our data using a multipoint program:

>> b=log(b+1)
>> keep b −− m2
>> write locus gh b.loc dummy
>> write gh b.pre dummy

The "dummy" keyword adds a dummy binary trait variable to the Genehunter pedigree file, so that program will calculate *ibd*s for us. If we like, Genehunter can be run from within Sib−pair:

>> $ gh

The "$" command shells out to run another program. When we exit from Genehunter, we will be returned to Sib−pair:

>> quit

```
This job took  1.8 minutes
```

There are a number of operations useful for manipulating pedigree data prior to analysis. These allow you to prune ("prune") pedigrees down to selected individuals and only those relatives needed to connect the index people, to reduce families to unrelated cases and controls ("case"), to break up large pedigrees into component nuclear families ("nuclear") or into unrelated cliques ("subped"), if the pedigree file does not specify all the connecting relatives (between different branches say). One can also select ("select" and "unselect") particular groups of pedigrees, and specify different values of a variable in each group:

Select out ASP nuclear families
>> nuclear
>> select containing exactly 2 where dementia and isnon and anytyp
or EDAC nuclear families
>> unselect
>> select containing 1 where IgE>1000 and isnon and anytyp
>> select containing 1 where IgE<50 and isnon and anytyp

Sib−pair also can be used as a calculator, and has a few genetics utilities, notably the "sml" and "grr" commands which gives expected recurrence risks, *ibd*'s and genotype frequencies for a specified diallelic model:

>> sml 0.01 0.5 0.2 0.1

```
-----------------------------------------------
Single Major Locus Recurrence Risk Calculation
-----------------------------------------------

 Frequency(A): 0.010000; Pen(AA): 0.500; Pen(AB): 0.200; Pen(BB): 0.100
 Trait Prev  : 0.102020; Pop AR:   2.0%; Var(Add): 0.000206; Var(Dom): 0.000004

 Measure      MZ Twin       Sib−Sib        Par−Off      Second
----------   ----------    ----------     ---------    ----------
 Rec risk         0.104         0.103         0.103         0.103
 Rel risk         1.023         1.011         1.011         1.006
 Odds rat         1.025         1.012         1.012         1.006
 PRR              1.020         1.010         1.010         1.005
```

```
 ibd|A−A          1.000         0.502         0.500         0.251
 ibd|A−U          1.000         0.500         0.500         0.250

 Freq of A if Affected: 0.019898 (0.000,0.039,0.961)
 Freq of A if Unaffctd: 0.008875 (0.000,0.018,0.982)

 Mating       Proportion    Risk to offspring
----------   -----------   ------------------
UnA x UnA         0.806         0.102
Aff x UnA         0.183         0.103
Aff x Aff         0.010         0.104
```

METHODS

Analytic Methods

Imputation of unobserved genotypes. This is performed using the algorithm described by Lange & Goradia [1987]. Firstly, (*0*) A phenoset (all possible genotypes) for one locus is generated for each individual in the pedigree. Then, iterate by nuclear families, repeating the next two steps until no further updates: (*1*) Parental genotypes inconsistent with their offspring are removed; (*2*) child genotypes inconsistent with their parents are removed. Finally, (*3*) If zero genotypes remain, report an inconsistency; if one genotype remains, this becomes the imputed genotype; if the joint spouse genotype is unambiguous, but the specific genotype each spouse carries is ambiguous, if requested, randomly assign a genotype to each parent. These latter genotypes might be used only for calculation of statistics for offspring of the pair, but not for the parents themselves. A further extension is to sequentially (founders then nonfounders) impute the remaining missing genotypes as the most likely member of the phenoset. This or a faster randomised algorithm is always run (unless the imputation flag is set to "−1", see below) to give starting values for the MCMC methods, but these simulated genotypes will not be saved unless the imputation level is set to "3" or "full".

Sex imputation. Likelihood of observed homozygosity at multiple sex−linked loci is calculated under hypothesis of male and female sex, assuming 0.1% of male and female genotypes are miscalled as heterozygotes.

Allele frequencies. These are for codominant systems only. Either a straight allele count is used, or the contribution of each pedigree is weighted by the number of founders it contains. Alternatively, the imputed and observed genotypes in the founders can be counted.

Hardy−Weinberg proportions. These are tested by a Pearson chi−square, with the P−value estimated via "gene−dropping" (Monte−Carlo,MC) simulation. These are based on the genotypes in the founders of that pedigree, where typed, or on the gene frequencies in the total sample where the founder is untyped, and the structure of the pedigree. For unrelated individuals, the exact test of Hardy−Weinberg equilibrium is calculated for diallelic markers.

Segregation ratios. The default analysis assumes the pedigrees are unascertained, and gives the naive estimators. The ascertainment corrected analysis is for nuclear families and follows Davie [1979]: $p=(r-j)/(t-j)$, where p is the risk, r is the total number of affected children, j, the number of sibships containing exactly one proband, t, the number of children. A fairly efficient approximate sampling variance is also given.

Haplotype reconstruction. This is performed on a nuclear family by family basis, though incorporating grandparental information where available. Initial reconstruction is performed using a simulated annealing algorithm that maximizes a sharing based criterion based on length of runs of the same alleles on a putative

chromosome among sib pairs, parent−offspring pairs, and grandparent−child pairs. The order of loci in the pedigree file is treated as the linkage order, and map distance information is *not* used. Missing parental haplotypes are filled in using the childrens' haplotypes in a simple fashion. Mendelian inconsistencies are flagged in the printout.

The second algorithm is more ambitious and attempts to construct recombination minimized haplotypes, again on a nuclear family by family basis, and dealing more intelligently with missing data. A simulated annealing algorithm using multiple restarts is used. Recombination events and Mendelian errors are flagged in the output.

Admixture analysis. This refers to testing for a mixture of specified distributions in the empirical distribution of a quantitative trait, usually a mixture of normals, though Sib−pair also offers mixtures of exponential and Poisson distributions. Information from the relationship between family members is not utilised. The usual EM approach is used.

Test for normality. The Filliben correlation [Filliben 1975] is calculated as a test for normality. This is the correlation between the observed data and the rankits expected under the assumption of normality. The P−value for this statistic is approximate, and is produced using an approach modelled on that used by Royston [1993] for the related Shapiro−Francia W':

$log(1-r_f) \sim N(m,s)$
$m:=1.0402\ (log(log(n))-log(n)) - 1.99196$
$s:=0.788392/log(n) + 0.31293$

where *n* is the sample size. This performs reasonably well versus the empirical percentiles:

	Coverage							
n	5	10	20	50	100	500	1000	5000
Nominal P=0.05	0.021	0.044	0.055	0.057	0.059	0.052	0.045	0.033
Nominal P=0.01	0.00	0.006	0.009	0.014	0.019	0.007	0.007	0.007

A common use of the Filliben correlation is as the criterion for selecting an optimal transformation of the data.

The $J_{0.02}$ statistic is a variance−corrected order−statistic based skewness measure:

$[(P_{0.02}+P_{0.98})/2-P_{0.50}]/[P_{0.75}-P_{0.25}]$

[David & Johnson 1956]. A test of normality can be constructed, using the standard error of $J_{0.02}$. This is based on interpolation of results from Monte−Carlo simulations ($SE(J_{0.02}):=1.36/sqrt(N)$ cf Resek [1974]).

Sibship variance test. This is the linear model suggested by Fain [1977] for the detection of the phenotypic effects of quantitative trait loci. Briefly, if parental trait values are at the extreme of the population distribution, then they will be carrying multiple increasing or decreasing alleles at the QTLs. As a result, the trait variance among their offspring is decreased compared to sibships whose parents have trait values close to the population mean. This U−shaped relationship between midparent value and sibship variance can be detected by fitting linear or quadratic curves.

Variance components analysis. This is the usual mixed effects analysis of quantitative traits assuming multivariate normality. The log−likelihood:

$LL = -0.5\ [log(det(S)) + (y-u)'\ inv(S)\ (y-u)]$

is maximized, where S is the variance−covariance matrix for the trait values for each phenotyped pedigree member, and y and u are the trait values and their expected values:

$u = B'X$,

reducing to the grand mean in the absence of covariates. The main diagonal elements of S takes the value:

$V_A+V_Q+V_D+V_E$,

and the off−diagonal elements:

$R_{i,j}V_A+ibd_{i,j}V_Q+K_{i,j}V_D$,

where,

$R_{i,j}$ is the coefficient of relationship for the $i-j$th pair of relatives
K is the coefficient of fraternity
ibd is the average *ibd* sharing at the marker location being tested for linkage to a QTL.

The maximization is performed using AS319 (variable metric minimizer with numerically estimated gradients). The phenometric models (ADE, AE, E) are fitted to the intact entire pedigree, but the models including V_Q can be fitted to the entire pedigree, or to sibships only.

Binary trait association analysis. This is the Pearson goodness−of−fit based test for equality of allelic gene frequencies at a marker locus in individuals expressing or not expressing a binary trait (2xN table). Both the nominal (ignoring relatedness of the sample) and empirical P−values for the test are output. The empiric P−value is estimated via gene−dropping simulation. These are based on the gene frequencies in the total sample, and the structure of the pedigree, and are conditional on the observed occurrence of the binary trait in the families.

Formerly, a sibship permutation based test was provided, calculating the same chi−square statistic for members of sibships that contain at least one affected and one unaffected typed individual, and generating a (within−sibship) permutation P−value. This is now replaced by a more powerful score test combining the within−sibship association and transmission−disequilibrium tests after Knapp [1999] and Laird et al [2000]. In this approach, the complete or partial genotype of untyped parents is reconstructed from the genotypes of the affected and unaffected children. The transmission of alleles from these reconstructed parents is conditioned on the genotypes of the children used in the reconstruction. The appropriate conditional distribution under the null hypothesis is approximated via Monte Carlo simulation and rejection sampling.

Population genetic F statistics (F_{IS}, F_{IT} and F_{ST}) are also calculated for each marker assuming affected and unaffected individuals come from separate related demes. These are estimated following the approach of Pons and Chaouche [1995], as described by Excoffier [2001].

Quantitative trait association analysis. This fits an additive (allelic means) model predicting an individual's trait value from his/her genotype at a marker locus. The residual sum−of−squares is compared to those obtained via a gene−dropping simulation of the pedigrees, giving an empiric P−value.

Two−locus linkage disequilibrium estimation. This algorithm finds informative founder matings, or informative matings where all the grandparents are untyped, and imputes the two−locus haplotypes

transmitted to the offspring. The loci are assumed to be tightly linked, so that four parental haplotypes are counted, as opposed to the more usual two haplotypes from the child. Both *D* and *D'* measures are calculated.

Homozygosity analysis. This tests for an increase in observed homozygosity at a marker locus in individuals expressing a binary trait, comparing this to the predicted homozygosity based on the allele frequencies in the total sample. This may occur in the presence of allelic association with a recessive trait locus, and/or deletional loss of heterozygosity (the parents would be untyped for such an individual not to have been flagged as a Mendelian inconsistency of course). A one−tailed empiric P−value is estimated via "gene−dropping" simulation, based on the gene frequencies in the total sample, and the structure of the pedigree. The multipoint homozygosity analysis uses the mean maximum marker homozygosity run length in the set of cases. Again, gene dropping is used to produce the distribution of this statistic under the null given the marker map, allele frequencies and observed pedigrees.

Transmission−disequilibrium test. The original formulation of the TDT is for a diallelic marker [Spielman et al 1993]. The TDT statistic calculated by the program is the Pearson goodness−of−fit based test of symmetry in the square table of transmitted versus nontransmitted alleles to each affected child [Haberman, 1979]. Empiric P−values are produced by randomization of the table. This global allelic form does not correct for the (usually small) correlation in parental genotypes induced by linkage disequilibrium (absent of course under the null hypothesis). Another allelic test provided is the marginal allelic test suggested by Spielman and Ewens [1996], which is probably slightly more powerful than the global symmetry test [Kaplan et al 1997].

The genotypic TDT P−value is estimated via gene−dropping based on the genotypes of typed ancestors of probands (where both parents of the proband and all antecedents must be typed), and the structure of the pedigree. The test statistic compares the observed number of each genotype transmitted with the number expected based on the parental genotypes. Pairs of cells whose total count is less than a given cutoff may be excluded from the analysis to increase power. The P−value for the TDT testing each allele versus all others in turn ("allele−by−allele") is the exact two−sided binomial probability (via the beta distribution). When the *plevel* is zero, only the best of the allele−by−allele test results are printed, and the P−value is Bonferroni corrected for the number of alleles at that marker.

Note that the default option is to use probands for the TDT only one parent is typed. For the diallelic marker case with unequal allele frequencies, using one parent families does lead to biased results. The unified transmission test available via the "assoc" command does not suffer from this problem (see above).

The Schaid and Sommer [1993] genotypic risk ratio tests for familial association under the assumption of Hardy−Weinberg equilibrium, or conditional on parental genotypes, is also offered. This uses log−linear modelling (implemented as iteratively reweighted least squares) of a biallelic locus with both parents genotyped. The attributable risk is also produced.

Haplotype Relative Risk analysis. This is the original familial association statistic comparing transmitted and nontransmitted allele frequencies unmatched on family (Falk and Rubinstein 1987; Knapp et al 1993). As usual, an empiric P−value is estimated via "gene−dropping" simulation, based on the gene frequencies in the total sample, and the structure of the pedigree.

Sib−pair analysis. Identity by descent estimation is based on the sib pair and parental genotypes when available. In the case of untyped parents, the full−sib sharing is the sum of sharing for each possible set of parental genotypes weighted by their likelihood based on all children in the sibship. Half−sib sharing is estimated based only on known genotypes, whether observed or unequivocally imputed. The effective degrees of freedom for the t−test of the slope of the regression line is given as the number of individuals in the sample (counted once only) who are in a sib pair where both members are typed at trait and marker, minus two. For a sample made up of nuclear families (no halfsibs), this will be equivalent to the £(sibship size−1)−2 value used

by SIBPAL 2.6, and originally suggested by Hodge [1984]. For binary traits, the same ordinary−least−squares analysis is performed −− the t−statistics from these results are only really applicable to large samples, and tend to be too liberal. The quantity regressed is not the usual Haseman−Elston squared trait difference, but a function of the squared trait sums and differences following Sham and Purcell [2001]. This approach is supposed to approach the power of the variance components approach according to those authors, and gives appropriate Type 1 error rates.

For the Fulker & Cardon methods, the expected *ibd* through the interval between the two markers is estimated using the equation given in Olson [1995]. Haseman−Elston regressions are performed at a series of points across the interval using the *ibd* sharing of the two flanking markers, and the given size of the interval. The Haldane mapping function is used.

Affected sib pair analysis. This is the original IBS based approach described by Lange [1986a], extended to half−sibs as per Bishop and Williamson [1990]. No correction for sibship size is made −− that is all possible pairs are treated as independent. The usual two d.f. chi−square is calculated, with expected counts being calculated based on the observed gene frequencies in the total sample. The IBD based mean test is also calculated.

Affected Pedigree Method. This uses the measures of genetic similarity described by Weeks and Lange [1988; see also, Lange, 1986a, 1986b], Whittemore and Halpern [1994], and Ward [1993, 1995]. The expected mean and variance for each pedigree is estimated via gene−dropping simulation. These are based on the observed gene frequencies in the total sample, and the structure of the pedigree. Both ibs and ibd based family scores can be estimated. The original APM statistics, the APM statistic of Whittemore and Halpern and the T(AB) and GPM statistics of Ward are calculated.

Multilocus IBS sharing statistics. These are used to confirm the pedigree structure using marker data. One approach calculates the overall probability of sharing two alleles IBS for full and half sibs, summing over all loci, and ignoring any linkage between markers. The second approach uses gene dropping based on the given marker map. Two lists are generated, one by individual, the other by relative pair.

Martingale residuals. The elegant approach of Commenges [1994] to genetic analysis of age−at−onset is to analyse the residuals obtained from a nonparametric or semiparametric survival analysis. Sib−pair implements the former, calculating the martingale residuals using the Nelson−Aalen estimator for the integrated hazard [eg Andersen et al 1993], which are then transformed following Therneau et al [1990] to give a more symmetrical distribution.

Generalized linear models and survival regression. These are the usual IRLS algorithms (using AS 164, [Stirling 1981]). The exponential and Weibull regressions are implemented as Poisson regressions (with log time as offset) as per Aitken and Clayton [1980].

Other standard statistical tests. A number of classical tests for independent data (eg unrelated cases) are implemented, such as contingency chi−square test (with Monte Carlo "exact" P−values, see below), ordinal by ordinal trend test for contingency tables [Yates 1948], and nonparametric one way analysis of variance via the Kruskal−Wallis test. Sib−pair can calculate the Pearson correlation coefficient for between−trait association, the Kaplan−Meier estimator for the survivor function, and Nelson−Aalen estimator for the hazard function, and measures of agreement for contingency tables.

Monte Carlo Algorithms

Gene−dropping. The "unconditional" algorithm producing null distributions is as follows. Repeat the following *1−3* steps a large number of times. (*1*) Founder genotypes are assigned using the allele frequencies

in the observed sample, assuming panmixia and Hardy− Weinberg equilibrium (HWE). Iterate, until all genotypes are assigned: (*2*) If both parental genotypes are nonmissing, randomly assign the index a genotype based on Mendelian autosomal inheritance (ie if parental genotypes are {1/2} and {3/4}, a child's genotype is randomly selected from {{1/3}, {1/4}, {2/3}, {2/4}}, with each genotype having a probability of selection of 0.25). Once complete, (3) calculate the test statistic based on the family's simulated genotype. Following completion of the outer loop, (4) summarize the distribution of the resulting test statistic. This procedure is used to generate null distributions for the association Pearson chi−square.

For the *share* command, this also allows for recombination between markers based on the given linkage map

The null distributions for the genotypic marginal TDT is generated using a "conditional on founders" algorithm, that takes observed founder/ancestor genotypes as given. Only typed nonfounders genotypes where both parents were typed are simulated.

ibd estimation. The modification to calculate *ibd* distributions gives each founder (two) unique alleles in his/her "typing genotype". A simple gene−drop gives the null distributions for the *ibs* and *ibd* statistics of the APM method.

I based the "conditional on observed genotypes" (gene−drop with rejection sampling) algorithm for calculating *ibd* on that described by Blangero et al [1995]. As before, each founder is assigned a typing genotype made up of two unique alleles. Offspring are only assigned *ibd*−typing genotypes that are consistent with the observed genotype at the observed locus of interest. For example, say the observed genotypes in the parents are 100/102 and 100/102, and the typing genotypes associated with these are set to {1/2} and {3/4} respectively. If the child is 100/102, assignment of typing genotypes {1/3} and {2/4} will be rejected. This is equivalent to a child's genotype being randomly selected from {{1/4}, {2/3}}, with each genotype having a probability of selection of 0.5. The resulting *ibd* statistics based on the typing genotype will approximate *ibd* for the marker locus of interest.

Missing genotype simulation by Monte−Carlo Markov Chain (MCMC). The calculation of *ibd* in the presence of missing genotypes is performed via a Metropolis algorithm. This algorithm is a multiallelic extension of that described by Lange and Matthysse [1989]. One iteration of the generation of a legal constellation of imputed and observed genotypes is produced by:

(1) Perform (a) (b) (c) or (d):

(a) Simulate *ibd*, then "mutating" up to four imputed founder alleles. These propagate through the pedigree using the current pattern of *ibd* transmission as indicated by the *ibd*−typing alleles, and are rejected and resimulated if an (unordered) inconsistency with an observed genotype occurs.

(b) Simulate *ibd*, then switch the parent of origin for an individual heterozygote. Propagate this change up through the pedigree to the originating founder(s), but not below the chosen "pivot" individual.

(c) A "conditional on observed genotypes" dropping of *ibd*−typing alleles, with the refinement that this ignores imputed genotypes. Inconsistencies thus generated for imputed genotypes are resolved by changing the imputed genotypes. This procedure will be slow in the presence of a large number of untyped nonfounders.

For (a)−(c), additional local proposals (as below) are compounded to these, and the resulting proposed constellation is accepted or rejected via the Metropolis criterion.

(d) Resimulate all untyped x untyped founder mating joint genotypes conditional on their offspring and other spouses, then other pedigree members singly, again conditional on surrounding genotypes. This is a simple Gibbs sampler, and is more efficient than the above when there are many missing genotypes in larger pedigrees.

MCMC burn−in. In releases prior to version 0.96.0, there was no burn−in for this Metropolis algorithm, as preliminary empirical tests had found the results from this program agreed well with "exact" results from programs such as GENEHUNTER. Subsequently, I have found some pedigrees where using the starting genotypes from the Lange−Goradia approach does lead to biased *ibd* estimates for certain pairs of relatives. Therefore, the program now performs a number of burn−in iterations (default 100) prior to those used to estimate *ibd*. The required number of such iterations depends on the number of missing genotypes in the pedigree.

Metropolis generalized linear mixed model and finite polygenic model sampler. This is either a "standard" or "slice" Metropolis sampler, where the simulated variables include diallelic QTL genotypes, Gaussian breeding values, a single QTL allele frequency (shared by all QTLs in the FPM), up to three genotypic means (shared by all QTLs in the FPM), polygenic and environmental variances (including pedigree ("VC") and maternally−derived sibship ("VS") variances).

The trait model can be gaussian, binomial (with identity, probit or logit link), poisson (including log link), weibull or MFT.

Proposals for diallelic QTLs genotypes are straightforward to generate, and do not usually give rise to noncommunication between sets of legal genotype proposals. Proposals for continuous variables are generated from random normal deviates, and a tuning parameter can be set that alters the variance of these proposal distributions.

The likelihood contribution from the *ith* individual to the Metropolis criterion for these models is (see for example, Guo and Thompson [1993]):

$$LL = F*£(\log(P(G_j))) + F*\log(f(a|V_A)) + (1-F)*\log(f(a|a_{FA}, a_{MO})) + \log(c|V_A) + I*\log(f(y|G_1,...G_j, a, c, V_E))$$

where,
P(x) denotes the probability of *x*,
f(x) denotes the density of *x*,
y is the trait value,
a is the breeding value,
c is the pedigree−specific intercept,
G_j is the genotype at the *jth* QTL,
V_A is the additive polygenic variance,
V_C is the familial environmental variance,
V_E is the error variance,
F=1 when a founder, 0 when a nonfounder
I=1 when phenotype observed, 0 when unobserved.

The conditional density for the breeding values of offspring includes the correction for inbreeding (the segregation variance being $1-0.5*(F_{FA}+F_{MO})$. The random effects are modelled as zero−mean gaussian.

The realizations of the parameters are summarized as means, and approximate standard errors produced by batching (default $B=\text{iter}^{1/2}$ [Jones et al 2005]). The interbatch lag−1 serial correlation is calculated as a diagnostic for the appropriate number of values to simulate [Ripley 1987].

The implementation of the generalized linear mixed models is quite straightforward in the chosen Metropolis paradigm (it would be more work to produce a Gibbs sampler, I believe), but for the "standard" sampler, it is important to check that the proposal acceptance rates are in the optimum range (usually stated as 0.2−0.6, Ripley 1987). This is less critical for the slice sampler, where the tabulated acceptance rates are actually the ratio of accepted proposals to the number of function evaluations (and so are just a measure of algorithm efficiency). Models fitting V_C and V_S or V_A are two−level GLMMs and so I have fitted a number of test datasets from the literature. There are surprising differences between results from standard software for some of these datasets, so although Sib−pair sometimes does not give identical results to that from non−simulation−based maximum likelihood methods, this may reflect approximations used by other programs.

Increasing the number of random effects chains is realized by duplicating the families the appropriate number of times and correcting the likelihood and standard errors. One is essentially averaging over multiple estimates of the random effects for each individual, as global parameters such the fixed effects regression coefficients and overall variances are the same over the replicate chains at that iteration. This seems to reduce bias in the estimation of the random effects, but with the side effect of increasing the between−batch correlation, and so slowing estimation. The tabulated results below generally used 4 chains run for 10000 iterations after a 1000 iteration burnin.

Binomial GLMM analysis of seed germination dataset of Crowder et al (1978) using different approaches. PQL1 is the penalised quasilikelihood approach implemented as glmmPQL() in the MASS package [Venables and Ripley 2002], while PQL2, AGQ are results from lmer() in the lme4 package of Bates and Sarkar [2005] using penalized quasilikelihood, adaptive Gaussian Quadrature respectively. The BUGS results are from the examples manual.

	Parameter Estimate (SE)				
Method	Sib−pair	AGQ	PQL1	PQL2	BUGS
SD of Plate Effect	0.30 (0.07)	0.24 (0.09)	0.23	0.24	0.29 (0.15)
Intercept	−0.51 (0.12)	−0.54 (0.17)	−0.54 (0.17)	−0.54 (0.16)	−0.51
Seed	0.06 (0.17)	0.10 (0.28)	0.09 (0.27)	−0.09 (0.28)	
Root	1.31 (0.18)	1.33 (0.24)	1.32 (0.23)	1.32 (0.24)	
Seed x Root	−0.79 (0.27)	−0.81 (0.38)	−0.81 (0.38)	−0.81 (0.38)	

Binomial GLMM analysis of "bacteria" dataset from the R MASS package [Venables and Ripley 2002] using 5 different approaches. PQL1 is the penalised quasilikelihood approach implemented as glmmPQL() by Ripley in the MASS package, while PQL2, AGQ and Laplace are results from lmer() in the lme4 package of Bates and Sarkar [2005] using penalized quasilikelihood, adaptive Gaussian Quadrature and the Laplace approximation respectively.

	Parameter Estimate (SE)				
Method	Sib−pair	AGQ	PQL1	PQL2	Laplace
RE Variance	2.05 (0.65)	1.70 (1.05)	1.98	3.27	1.66
Intercept	3.70 (0.42)	2.86 (0.48)	2.74 (0.38)	2.75 (0.48)	2.81 (0.48)
Low dose	−1.46 (0.45)	−1.36 (0.82)	−1.25 (0.64)	−1.25 (0.82)	−1.35 (0.82)
High dose	0.60 (0.42)	0.58 (0.85)	0.49 (0.67)	0.49 (0.85)	0.58 (0.85)
Week>2	−1.66 (0.28)	−1.63 (0.46)	−1.61 (0.36)	−1.61 (0.46)	−1.57 (0.46)

Binomial GLMM analysis of contraception usage data from the 1988 Bangladesh Fertility Survey [Steele et al 1996].

	Parameter Estimate (SE)	
Method	Sib−pair	PQL
RE Variance	0.25 (0.06)	0.22
Intercept	−1.67 (0.16)	−1.66 (0.15)
Age	−0.03 (0.01)	−0.03 (0.01)
Urban	0.72 (0.07)	0.72 (0.12)
1 child	1.10 (0.12)	1.09 (0.16)
2 children	1.36 (0.15)	1.35 (0.17)
3+ children	1.32 (0.17)	1.32 (0.18)

Poisson GLMM analysis of epileptic seizure count data of Thall and Vail [1990] using different approaches. PQL1 is the penalised quasilikelihood approach implemented as glmmPQL() in the MASS package [Ripley and Venables 2002], while PQL2, AGQ are results from lmer() in the lme4 package of Bates and Sarkar (2005).

	Parameter Estimate (SE)			
Method	Sib−pair	AGQ	PQL1	PQL2
RE variance	0.268 (0.029)	0.252	0.197	0.101
Intercept	1.834 (0.094)	1.833 (0.074)	1.870 (0.106)	1.870 (0.074)
Progabide	−0.346 (0.129)	−0.334 (0.105)	−0.310 (0.149)	−0.309 (0.105)
log basal rate	0.861 (0.119)	0.883 (0.091)	0.882 (0.129)	0.882 (0.091)
Base:therapy	0.394 (0.157)	0.339 (0.143)	0.342 (0.203)	0.342 (0.143)
log age	0.513 (0.219)	0.481 (0.244)	0.534 (0.346)	0.533 (0.244)
Period 4	−0.159 (0.019)	−0.160 (0.055)	−0.160 (0.077)	−0.160 (0.143)

Poisson GLMM analysis of European male melanoma death rate dataset of Langford et al (1998) using different approaches. PQL1 is the penalised quasilikelihood approach implemented as glmmPQL() by Ripley and Venables [2002] in the MASS package, while PQL2, AGQ are results from lmer() in the lme4 package of Bates and Sarkar (2005). The STATA result used the xtpois command, and comes from the review article at *http://www.mlwin.com/softrev/revstata.html*.

	Parameter Estimate (SE)				
Method	Sib−pair	AGQ	PQL1	PQL2	STATA
Region variance	0.188 (0.012)	0.170 (−)	0.161	0.125	0.102 (−)
Intercept	−0.151 (0.038)	−0.139 (0.043)	−0.129 (0.049)	−0.129 (0.043)	−0.138 (0.017)
UVB insolation	−0.035 (0.005)	−0.034 (0.009)	−0.038 (0.010)	−0.038 (0.009)	

					−0.056 (0.004)

The final results are sometimes sensitive to the choice of starting values for the random effects (the fixed effects are started automatically from the marginal model parameter estimates using "reg"), and to the proposal step size. Because of the correlation between random and fixed effects in GLMM's other than Gaussian (since the intercept affects variance), differences in the estimated random effect size do alter the fixed effects regression coefficients.

The output (with *plevel* set to 1) allows plotting of parameter estimates to assess convergence of the chain.

Randomized TDT. The randomization test for the global allelic TDT permutes the transmission table by randomly selecting a single proband−parent pair and reversing the transmitted and nontransitted alleles. One "shuffle" of the table involves N such permutations, where N is the number of such informative parent−proband pairs in the observed pedigrees (this reduces the correlation between successive tables in the random walk to close to zero).

Sequential empiric P−values. The Monte−Carlo P−values provided for the various MC−based tests are produced using the sequential approach described by Besag and Clifford [1991]. In this refinement, we only generate as many pseudosamples as is necessary to give a P−value numerator of size *mincount*; the denominator is the number of pseudosamples. The practical effect of this procedure is that if the true P−value is large, then relatively few pseudosamples are generated to give a less precise estimate of this uninteresting value. Besag and Clifford suggest a value for *mincount* of 10−20. It is necessary to set a maximum denominator to avoid excessive computation for "highly significant" results.

Other empiric P−values. An exception to this is the algorithm used for empirical P−values for the APM. Here, a P−value for each family is simulated at the same time as the mean and variance. These P−values were previously combined using the procedure due to Fisher [Hedges and Olkin 1985], that is, twice the sum of the natural logarithms of the P−values was treated as a chi−square variate with 2*N degrees of freedom, where N is the number of contributing families. This does not seem to be particularly powerful, so each P−value is now inverse−normal transformed to a Z−score, and these combined in an unweighted fashion [Hedges and Olkin 1985].

USAGE

The program reads commands from standard input, and writes results to standard output. Therefore, the program can be run interactively, or if a series of commands is to be found in a file, in batch mode. If the input file was *test.in*, entering "**`sib−pair <test.in >test.out`**" would perform the commands in test.in, and write results to test.out.

A command is a single line of keywords, locus names and/or variable values. If a "\" character is the last *word* of a line, the next line is interpreted as a continuation of the previous command. Sib−pair is case−sensitive, so that the keyword "READ" is not equivalent to "read". Commands are either global, which can be entered at any time; descriptive (*set impute*, *set locus*, *read pedigree*), which must precede the *run* statement; the *run* statement, that causes the dataset to be read and processed; or analytic, which act only after the *run* statement.

One command, *set plevel*, controls verbosity of output. Some useful descriptive tables are only printed if *plevel* is at least 1.

Sib−pair's parser can evaluate simple algebraic and logical expressions for each record in a datafile, but does

not allow complex programming. For example, multiple actions contingent on a single condition usually have to each have their own *if* statement.

Global commands

1. **!**|**#**. The rest of the line is a comment, and is echoed to standard output.
2. **%**|**$**. The rest of the line (up to position 80) is a command, and is passed to the shell for execution.
3. **clear**. Restarts the program, closing all workfiles and zeroing all arrays.
4. **help** [**All**|**Globals**|**Operators**|**Data**| **Analysis**|<*search_string*>]. Prints a brief description of the commands −− either all, a subset, or all matching the search string.
5. **info**. Information about program settings and the current dataset. For the latter, gives counts of active and inactive pedigrees, individuals, and loci and a table of numbers observed for every trait and marker.
6. **list**|**ls** [**markers**|**$m**|**$x**|**traits**|**$a**|**$q**]. List of loci in current analysis.
7. **show loci**. List of active loci along with table of numbers available (as per **info**).
8. **show pedigrees**. Same as **gener**, with print level 0.
9. **show map**. Shows the current marker map.
10. **time**. Print time elapsed since start of the program.
11. **set timer** [**on**|**off**]. Show time taken by each command.
12. **include** <*command_file*>. Read in Sib−pair commands from a file.
13. **last** [<*line_number*>]. With no argument, displays the command history, otherwise submits that line of the history for reevaluation. A negative line number counts backwards from the current line. The command history is saved to a file "sib−pair.log".
14. **quit**|**bye**. Halts the program.
15. **set prompt** [**on**|**off**]. Displays a prompt, and activates/resets the command line history.
16. **set ndecimal_points** [<*nwid*>] <*ndec*>. The total width (number of characters) of a quantitative variable written to a new pedigree file defaults to 9 (and is fixed to 8 for some files, notably MENDEL and FISHER) but can be set as high as 20 for GAS and LINKAGE format files. The number of decimal places can be set to *ndec*.
17. **set epoch** [**iso**|**jul**|<*epoch*>]. Set or show the epoch used for julian dates. Defaults to "iso" epoch of 1970−01−01.
18. **set out**|**plevel** <*level*>|**verbose**|**on**|**off**. Increasing the print level causes more information to be printed by almost all procedures. Print level *1* prints out the identities and genotypes of parents imputed where the genotype was missing, raw counts of genotypes for the *hwe* procedure, expected *ibs* probabilities for the *asp* procedure etc. Print level *2* (or *verbose*) writes out the statistics for each simulated dataset for the MC based procedures, the intrapair variance and ibd sharing for each pair in the sib pair analysis, etc. Print level *−1* omits outputting a list of pedigrees.
19. **set weight founders**|**imputed**. Weights contribution of each pedigree to the allele frequencies by the number of typed founders, or alternatively gives the count of the founder alleles, observed and imputed.
20. **set burn−in** <*number of iterations*>. Controls the number of Markov Chain Monte−Carlo iterations used by the *apm* algorithm discarded before estimation commences. Default is 100 iterations. Setting *bur* to zero means no burn−in is performed (the old default).
21. **set iteration**<*number of iterations*>. Controls the number of iterations used by the various Monte Carlo algorithms. Default is 200 iterations. Setting *ite* to zero means the Monte Carlo procedures are not performed.
22. **set emit** <*number of iterations*>. Controls the number of (Monte−Carlo) Expectation Maximization iterations used by the *mcf* algorithm.
23. **set batch** <*number of batches*>. Controls the number of batches used by the *fpm* algorithm for the estimation of parameter standard errors.
24. **set chain** <*number of MCMC chains*>. Controls the number of chains used by the *fpm* algorithm.

25. **set tune** <*MCMC tuning parameter*>. Controls the multiplier for the MCMC proposal step size used by the *fpm* algorithm. The base step size is usually the fixed effects model standard error for that parameter, and **tune** defaults to 0.3.
26. **set mincount** <*minimum numerator of P−value*>. Controls the number used for Monte Carlo simulation of a P−value. Default is 20 pseudosamples with a test statistic more extreme than that for the observed statistic. Set *mincount* equal to *iter* if this is not desired.
27. **set seeds** <*seed1*> <*seed2*>< *seed3*>. Initializes random number generator seeds to given values, rather than via system time.
28. **set tdt bot[h parents]**|**one [parent]**|**first**. Limits TDT statistic to cases where either both parents or at least one parent is typed, or one proband per family where both parents typed.
29. **set hre zero**|**children**. Assume zero recombinants between markers for **dis** command where parents genotyped, thus counting four imputed parental haplotypes. Alternatively, only utilize two haplotypes from children.
30. **set map function kosambi**|**haldane**. Set the mapping function used by multipoint analytic and locus file output routines.
31. **pchisq** <*chi−square*> <*degrees of freedom*>. Calculate P−value for central chi−square distribution.
32. **chisq** <*nrows*> <*ncols*>. Calculate contingency chi−square and permutation P for flat table entered via keyboard.
33. **proportion** <*numerator*> <*denominator*> <*confidence interval width*>. Calculate accurate confidence interval following Wilson (as described by Agresti and Coull) for a proportion.
34. **sml** <*Frequency of **A** allele*> <*Penetrance of **AA** genotype*> <*Penetrance of **AB** genotype*> <*Penetrance of **BB** genotype*>. Calculates recurrence risks and segregation ratios under a specified diallelic generalized single major locus model.
35. **sml** <*Frequency of **A** allele*> <*Mean for **AA** genotype*> <*Mean for **AB** genotype*> <*Mean for **BB** genotype*> <*standard deviation for **AA** genotype*>. [<***AB** SD*> [<***BB** SD*>]]. Calculates mean, variance components and parent−offspring regression results under a specified diallelic generalized single major locus model.
36. **grr** <*trait prevalence*> <*Frequency of **A** allele*> <*genotypic risk ratio*> [**add**|**dom**|**rec**]. Calculates recurrence risks and segregation ratios under a diallelic generalized single major locus model specified via trait prevalence, ratio of penetrances and pattern of inheritance (codominant multiplicative, dominant or recessive).

Algebraic operators and functions

37. "<*allele1*>/<*allele2*>". Double quotes mark the contained text for special evaluation by the parser. A constant genotype is written as two numbers (1−999) or letters (a−zA−Z) separated by a slash and surrounded by quotes. Other quoted items are passed intact to be read, either as a reserved command or as a single Fortran real, so "1+3" is evaluated as 1000, and "1 1" as 11.
38. (<*value*>|<*locus*>) *****|/|+|−|^ (<*value*>|<*locus*>). Arithmetic operations combining numerical constants and/or trait values. The result of an operation involving constants is a single constant, but an operation involving a trait value results in *nobs* results (where *nobs* is the number of individuals in the pedigree file.
39. (<*value*>|<*locus*>) **<**|**>**|**=<**|**=>**|**ge**|**le**| **eq**|**==**|**ne**|**^=**|**and**|**or** (<*variable*>|<*locus*>). Logical operations comparing numerical constants and/or trait values. when operating on genotypes, the equality and inequality operators require both pairs of alleles to meet the criterion, but the comparison operators test true if *either* pair of alleles meets the criterion. That is "2/2">"1/3" evaluates to True, but "1/2"=="2/2" evaluates to False.
40. **if** <*logical expression*> **then** <*action*> [**else if** <*logical expression*> **then** <*action*>]... [**else** <*action*>]. Conditional evaluation of expressions. Note that **if** statements cannot be nested.
41. **log**|**sqrt**|**exp**|**sin**|**cos**|**tan**|**asin**|**acos**|**atan**|**abs**|**int**|**round** (<*variable*>|<*locus*>). Functions acting on numerical constants and/or trait values.
42. **rand**|**rnorm**. Produce a random value from U(0..1) or N(0,1).

43. **istyp**|**untyp** <*marker*>. Test if genotyped at given marker. Necessary since if imputation is higher than −1, all untyped individuals have a genotype containing negative allele numbers (used to start MCMC algorithm).
44. **ishom**|**ishet** <*marker*>. Test if homozygous (or heterozygous) at given marker.
45. **alla**|**allb** <*marker*>. Return the first or second allele for each individual at the given marker.
46. **marcom**. Show the maximum of the number of markers an individual and any of his relatives are both genotyped at.
47. **numtyp**|**anytyp**|**alltyp**. Show number of markers an individual is genotyped at, or indicate whether genotyped at any one or all marker loci.
48. **male**|**female**|**isfou**|**isnon**. Test sex and founder status of individual.
49. **num**|**nfound**. Number of members and number of founders of the pedigree containing an individual.
50. **famnum**|**index**. Position of pedigree and of individual in the *active* dataset.

Data Declaration commands

51. **set datadirectory** <*pathname*>. Sets directory to be searched for pedigree files.
52. **set workdirectory** <*pathname*>. Sets directory to which temporary files are written.
53. **set impute off**|**on**|<*level*>. Toggles imputation routine.
    - Imputation level *0* (**off**) does not impute genotypes. It does generate legal genotypes for all untyped individuals, that are used for the MCMC IBD estimation. This approach is slowed by, and can fail (stochastically) in large pedigrees. In this case, the imputation level can be set to *−1* (*completely off*!).
    - Imputation level *1* (**on**) imputes single individual's genotypes if unambiguous. All missing genotypes are silently imputed for the use of the MCMC IBD routine via sequential imputation and application of the Lange−Goradia algorithm (see level *3*).
    - Imputation level *2* imputes the genotypes of a spouse pair where the exact owner of each genotype is ambiguous eg both parents unknown, children 1/2, 3/4 so parental phenosets are each {{1/3},{2/4}}. Of limited use.
    - Imputation level *3* makes available all the imputed values for the missing genotypes (in pedigrees where at least one member is typed at the locus). The missing genotype is replaced by the most likely genotype conditional on the typed members of the pedigree, and those genotypes imputed at that time −− single locus sequential imputation. Imputation proceeds through the untyped founders, according to the collation order of the ID, then nonfounders, according to the collation order of parental IDs.
54. **set errordrop off**|**on**|<*level*>. Toggles automatic deletion of genotypes that give rise to a Mendelian inconsistency, either an entire nuclear family (level 1), or an entire pedigree (level 2, the default).
55. **set checking off**|**on**. Toggles the first level testing for Mendelian inconsistencies within nuclear families.
56. **set locus** <*locus name*> <*locus type*> [<map position> [<description...>]]. Declares position (by order within list), name and type of locus within pedigree file. Locus type may be either:

marker	an autosomal (fully) codominant marker
xmarker	X−linked codominant marker
quantitative	quantitative (or interval or ordinal) trait
affection	binary trait

It is best to avoid a locus name containing reserved characters (eg "+−*/()^\"), if algebraic manipulation of that variable will be required (otherwise quotation of the name is required). Names identical to commands also cause trouble unless protected by brackets.

The fourth column (optionally) contains the genetic map position. All subsequent words (up to a total of 40 characters) are stored as an annotation. The annotation is appended to the long form of output of some commands (eg **show loci** or **list**), and is searchable by some commands (currently **keep|drop where**).

57. **rename** <*locus name*> [**to**] <*new name*>. Change name of previously declared locus.
58. **loci** <command file>. Read in Sib−pair locus and pedigree file declarations from a file.
59. **read locus linkage** <*locus file name*>. Read locus names, types and map positions from a Linkage−format locus (*.dat*) file. Does not recognise factor coding of genotypes, but does create a new quantitative trait for liability class
60. **read locus merlin** <*locus file name*>. Read locus names, types from a Merlin−format locus (*.dat*) file.
61. **read pedigree** <*pedigree file name*>|**inline**. Reads a GAS type pedigree file either from an external file, or inline following the command. The inline data is terminated by a line containing ";;;;".
62. **read linkage** <*pedigree file name*>|**inline**. Reads a LINKAGE type pedigree file.
63. **set sex on**. Creates a quantitative dummy variable for sex (field 6 in pedigree file).
64. **set skiplines <slines>**. Skip first *slines* lines in pedigree file) when reading in.
65. **order** <*loc1*>...[<*locB*> **to** <*locC*>]... [$(**m**|**x**|**q**|**a**)[**r**|**m**]]...<*locN*>. Set order of loci. Addition of *r* to a class eg *$mr*, reverses the order of all members of that class, while the *m* modifier causes the order to be the genetic map order. You may have to revise the genetic map order (by *set map* or *set dist* to get sensible export files for some programs such as Linkage (Sib−pair assumes a map position lower than the preceding position implies the markers are unlinked).
66. **set map** <pos1>...<posN>. Set map positions for the marker loci. This will overwrite any original map positions.
67. **set distances** <dis1_2> <dis2_3>...<posN−1_N>. Set interlocus map distances map positions for the marker loci. Distances are in centiMorgans. This will overwrite any original map positions.
68. **read map** <map file name>. Read in map positions for loci from a file, matching via names of previously declared markers. Should recognize most formats of map file automatically eg Merlin, Mendel, Solar. Tests number of columns and whether column contents are numeric or alphabetical, skipping first row as possible header).
69. **run**. Reads in pedigree file and creates working pedigree file. Imputes genotypes if requested.

Analysis and data manipulation commands

70. **keep**|**drop** <*loc1*>...[<*locB*> **to** <*locC*>]... [**$(m|x|q|a)**]...<*locN*>. Retain or exclude loci for subsequent analysis. Consecutive loci can be summarized as a range, as can all members of a particular class of locus type (*marker*, *quantitative*, *affection*) via a class ($*type*) token. Note that dropped variables can still be used in algebraic and logical expressions.
71. **keep**|**drop where** (**monomorphic** | **max** <*frequency*> | **number_typed** <*ntyp*> | **distance** <smallest_gap> | **every** *number_skipped* | *search_string*>). Retain or exclude loci for analysis. Note that dropped variables can still be used in algebraic and logical expressions. The *where* condition can be used to match the set of loci meeting that condition. Available conditions are: test that a marker is monomorphic, that the commonest marker allele frequency exceeds a threshold, the number of individuals typed falls below a threshold, the marker is closer than a set amount to the last included marker, every Nth marker in list, or the marker annotation contains the search string.
72. **undelete** [<*loc1*>...[<*locB*> **to** <*locC*>]... [**$(m|x|q|a)**] ...<*locN*>]. Return previously dropped loci to analysis. Default is to undelete all dropped loci. This is not the reverse of the *delete* command.
73. **select** [**containing**|**exactly** <*nprobands*>] [**where**] <*a logical expression*>. Select pedigrees containing one or more individuals with a trait value meeting the criterion.
74. **select pedigree|id** [[**not**] **in**] <*ped1*>...<*pedN*>. Select pedigrees included or excluded from a list of pedigree or individual names. The names can contain wildcard characters: "." (match any character in the target at that position in the search string) and "*" (match any characters zero or more times in the target at that position in the search string).

75. **unselect**. Returns all pedigrees excluded by a select command back to the analysis.
76. **pack loci**|**pedigrees**. Permanently delete all loci currently excluded by a **drop** command, or all pedigrees currently excluded by a **select** command from the work file.
77. **edit** <*pedigree*> <*person*>|**all** <*trait*> **to** <*value*> [<*new value*>]. Allows editing of trait values or genotypes. The **all** keyword performs the action on all members of that pedigree: since wildcards can now be used, an equivalent is "edit <ped> *".
78. **delete** <*pedigree*> <*person*>|**all** Sets all data to missing for a specified individual. The **all** keyword performs the action on all members of that pedigree.
79. **delete** [<*locus1*>...<*locusN*>] **[whe|where]** <*a logical expression*>. Sets specified data to missing for all individuals meeting particular criteria.
80. **recode** (<*marker*>|$(m|x)) [**frequencies**]. Recodes alleles at that marker or set of markers to 1..N, where the ordering defaults to the allele size (or collation order for letter alleles). If the **freq** modifier is present, the numbering is by ascending allele frequency.
81. **recode** <*marker*> <*all1*|*value1*>...<*allN*|*valueN*> **to** <*new allele*|*new value*>. Allows pooling of marker alleles prior to subsequent analysis.
82. **combine** <*marker1*> [...<*markerN*>] [<threshold>]. Pool rare alleles for a marker into one new allele. "Rare" defaults to a frequency of 5%, but can be changed via the last parameter on the command line.
83. **date** <*quantitative_trait*> [**julian**|**gregorian**|**year**]. Convert a numeric variable from Julian to Gregorian, Gregorian to Julian, or Gregorian to "decimal" year. The "chronological" Julian date is the number of days since the epoch, usually 1970−01−01 or −4712−01−01. Gregorian dates are represented as 8 (or 9) digit integers of the form of (−)YYYYMMDD. The decimal years are YYYY.x, where the decimal part is the day of year number (from 1...366) divided by the length of that year (365 or 366).
84. **date** (<*yyyymmdd*> **julian**)| (<*juldate*> **gregorian**). Convert a single date from Julian to Gregorian or Gregorian to Julian.
85. **transform** <*xtrait*> <*divisor*> <*subtractand*> <*power*> <*lower threshold*> <*higher threshold*>. This transform the quantitative trait *xtrait* as:

    $$boxcox(\{xtrait - subtractand\}/divisor)$$

    where boxcox() is (a slightly altered) Box−Cox transformation, so that:

    - if *power*=0, the transformation is $log(\{x-s\}/d)$;
    - if *power*=1, it is $\{x-s\}/d$;
    - and otherwise $[(\{x-s\}/d)^p-1]/p$.

    The resulting transformed value can then be truncated above or below using a specified *low* or *high* threshold.
86. **standardize** <*trait*> [**familywise**]. Replace each trait value with its Z−score, ie (*x*−*xbar*)/*sd*, where *xbar* is the total sample mean, and *sd* the total sample standard deviation. This can also be performed using the individual's family mean and standard deviation, if the **fam** keyword is included.
87. **adjust** <*ytrait*> **on** <*xtrait*> [**to** <*adjustment value of xtrait*|*m*|*f*>]. Performs linear regression of quantitative trait ytrait on quantitative or binary trait xtrait (or sex, if **sex** is set to **on**), calculates residuals, and adds *adjustment value* or, if not specified, the mean value of *xtrait*. The residuals then replace the original values of *ytrait*. A multiple regressive adjustment of Y on X_1 and X_2 requires sequential adjustment of Y on X_2,X_1 on X_2, and then Y on the adjusted X_1.
88. **residuals** <*ytrait*> **on** <loc1>...[**to**]...<locN> [**complete_obs**]. Replace quantitative trait with the residuals from the multivariate regression on the list of predictors (which may include the average allele length of a marker locus). The **com** option means only individuals with no missing values for any of the listed traits will be updated. Otherwise, missing values are replaced with the sample mean

for that phenotype when calculating the predicted value.

89. **predict** <*ytrait*> **on** <loc1>...[**to**]...<locN> [**complete_obs**] Replace quantitative trait with the predicted value from the multivariate regression on the list of predictors (which may include the average allele length of a marker locus). The **com** option means only individuals with no missing values for any of the listed traits will be updated. Otherwise, missing values are replaced with the sample mean for that phenotype. when calculating the predicted value
90. **impute** <*ytrait*> **on** <loc1>...[**to**]...<locN> [**complete_obs**] Replace missing quantitative trait values with the predicted value from the multivariate regression on the list of predictors (which may include the average allele length of a marker locus). The **com** option means only individuals with no missing values for any of the listed predictor traits will be updated. Otherwise, missing values are replaced with the sample mean for that phenotype when calculating the predicted value.
91. **kaplan−meier** <*age−at−onset*> <*censor*> [**residuals**]. Prints the product−limit estimator for the survivor function for the quantitative trait *age−at−onset*, where *censor* is the binary outcome trait, which is *affected* when *age−at−onset* represents the age at which the individual first expressed the trait. The *age−at−onset* is replaced by a nonparametric residual when requested. If *affected*, this is:

$$sgn(1-H(t)).(-2(1-H(t)+log(H(t))))^{1/2}$$

If *unaffected*:

$$-(-2H(t))^{1/2}$$

where $H(t)$ is the Nelson−Aalen estimate of the integrated hazard function at that age t.

92. **rank** <*trait*> <*rank*>. Write the ranks of a quantitative trait to the quantitative variable *rank*.
93. **simulate** <*marker*> [<*linked extant marker*>] [<*number of equifrequent marker alleles*> | <*allele 1 frequency*>...<*allele N frequency*>]. The data for the named autosomal marker is replaced by simulated data. If a second marker name is given, the new marker is simulated as being completely linked to the second marker. Either a set of allele frequencies, or the number of (equifrequent) alleles, can be given for the simulation. If the sum of the given allele frequencies is less than 1, an extra allele will be added automatically.
94. **nuclear** [*maxsibs*] [**grandparents**]. Split pedigrees into component nuclear families, duplicating individuals as necessary. If *maxsibs* is set, then sibships containing more than *maxsibs* members are truncated. The *gra* option includes the grandparents as well.
95. **subpedigrees**. Split nominal pedigrees into component true pedigrees. Sib−pair normally can analyse a group of individuals with the same pedigree ID, even if they are not all related. This command splits such groups into uniquely named formal pedigrees.
96. **prune** [<*binary trait*> [|<*quantitative trait*> **over**|**under** <*threshold*>]]. Reduce pedigree to contain probands and minimum number of connecting relatives.
97. **cases** <*locus*>. Reduce pedigree to unrelated individuals with non−missing values at the trait i.e. the informative founders, and any informative nonfounders who are not directly related to any individuals already selected.
98. **unique_id** [**sequential**]. Generate unique consecutive (within family) numerical IDs for all individuals (as well as new numeric pedigree IDs). The **sequential** gives IDs from 1...total_records, instead of 10001, 10002...20001...
99. **print** [**where**] <*a logical expression*>. Print trait values for individuals, with a combination of trait values meeting the criterion.
100. **print** [**ped**] <*Ped1*>...<*PedN*> [**id**] <*Id1*>...<*IdN*> Print trait values for individuals, with specified combination of pedigree and individuals IDs. The pedigree and ID names can contain wildcard characters: "." (match any character at that position in the search string) and "*" (match zero or more

characters).

101. **write** [<*pedigree file name*>]. Writes a GAS type pedigree file from the current dataset. Default is to screen.
102. **head** [<*nrec*>]. Writes the first *nrec* records of a pedigree file to the screen.
103. **write pap**. Writes the required pedigree files *trip.dat* and *phen.dat* (note that you may have to sort *trip.dat*).
104. **write**

pedigree|**gas**	<*pedigree file name*>	
arl		[**par**|**all**]
asp|**tcl**		
crimap|**tcl**		
csv		[**nop**]
dot		
fisher		
gda		[**all**]
linkage|**ppd**|**gh**		[**dummy**] [**numbered_alleles**]
mendel		
phe		
sage		
solar		[**phe**] [**nop**]

Use of the keywords *pedigree* or *gas* writes a GAS type pedigree file from the current dataset. Quantitative values are written as F9.x or F8.4. The keyword *gda* writes a GDA Nexus datafile containing all current marker genotypes for founders. If the keyword *all* is added, nonfounders will be included as well, but the "gdatype" format will not differentiate between relatives. Similarly, *arlequin* writes a data file for the program Arlequin containing haplotypes from one informative child per family, or two parents of such a child if the *par* keyword is added. Only if the *all* keyword is added will all genotyped individuals be printed. The keywords *linkage* and *ppd* write a pedigree file from the current dataset suitable for use by the LINKAGE (and FASTLINK) programs, the former type requires preprocessing by the Makeped program (note that if a quantitative trait value is zero −− that is nonmissing −− it is recoded to 0.0001); *aspex* (or *tcl*) writes a linkage style pedigree file but with the marker locus names as the first line, as the ASPEX programs prefer; *gh* writes a linkage style pedigree file with a dummy affection trait as the first trait and all the quantitative traits last, with "−" for missing quantitative trait values. The *dummy* option added to *linkage* or *gh* writes a dummy affection locus as the the first trait (everybody affected). The *numbered_allele* option skips recoding alleles to numbered alleles which is slow as currently implemented. The *sage* keyword writes a pedigree file from the current dataset suitable for use by the program FSP included in the SAGE package; *mendel* writes a pedigree file from the current dataset suitable for use by the programs MENDEL or SIMWALK; *fisher* writes a pedigree file from the current dataset suitable for use by the program FISHER; *phe* writes the "pheno.dat" style file required by Mapmaker−Sibs; both *csv* and *solar* give a comma delimited file, with header naming columns, from which the parental ID columns can be dropped via the *nop* option, and the SOLAR phenotype file written by the *phe* option.

105. **write map mendel**|**merlin**|**loki** <*map file name*>. Writes out the map file required by MENDEL 4.0, MERLIN or LOKI.
106. **write locus pap**. Writes the required locus files *header.dat* and *popln.dat*.
107. **write locus**

aspex|**tcl**	<*locus file name*>	
fisher		
gas		

linkage|**gh** [**dummy**]
loki
mendel
merlin
sage
sib−pair

Use of the keyword *gas* writes a GAS type locus file from the current dataset; *linkage* writes a locus file from the current dataset suitable for use by the LINKAGE (and FASTLINK) programs; *gh* writes the same as *linkage* save that map distances are in cM. The *dummy* option is used when the first trait is a dummy trait generated by *write linkage <file> dummy*. The keyword *loki* writes a control file for LOKI's *prep* program; *sage* writes a locus file from the current dataset suitable for use by the program FSP included in the SAGE package; *mendel* writes a locus file from the current dataset suitable for use by the programs MENDEL or SIMWALK; *fisher* writes a locus file from the current dataset suitable for use by the program FISHER; *merlin* for MERLIN; *tcl* or *aspex* writes the tcl command file required by ASPEX programs such as SIB_PHASE; *sib−pair* writes a Sib−pair style script.

108. **write var** [**mendel**] <*var file name*>. Writes out the var file (list of quantitative traits) required by MENDEL.
109. **generations** [<*quantitative trait*>]. List founders/marry−ins and sibships by generation number for all pedigrees, (over)writing the generation number to a quantitative trait if requested.
110. **haplotypes** [<*binary trait*>]. This (still experimental) routine displays nuclear family (plus grandparental, where present) haplotypes as an ASCII pedigree drawing.
111. **triads** This routine lists haplotypes inferred from fully typed parent−offspring triads, along with counts of obligate recombinants.
112. **relatives** This routine lists relatives of an index individual: parents, sibs, spouses, offpsring and descendants.
113. **ancestors** <*binary trait*> |(<*quantitative trait*> >|>=|<|<=|==|^= < *threshold*>). This prints the IDs of the ancestor (and ancestral mating) shared by the greatest possible number of probands in a family. The mean intrafamilial inbreeding coefficient for the probands is also output.
114. **frequencies**|**describe** [[<*codominant marker*>| <*binary trait*>| <*quantitative trait*>]...[**to**]...<*trait*>] | **snp**. Print allele frequencies for marker loci, segregation ratios for binary trait, or means, variances, familial correlations and a sibship variance test for a quantitative trait. Default is to describe all loci. The *snp* option prints minor allele frequencies and number typed for all diallelic marker loci.
115. **count** [**where**] <*a logical expression*>. Count individuals, and sibships and pedigrees containing such individuals, with a combination of trait values meeting the criterion.
116. **print** [**where**] <*a logical expression*>. Print phenotype data for individuals with a combination of trait values meeting the criterion.
117. **tab** <*trait 1*>...[<*trait N*>]. Print contingency table for one, two or N traits, along with contingency chi−squares, Kruskal−Wallis test or odds ratio if appropriate. For RxC contingency tables where the second variable is a diallelic marker locus, allele frequencies and exact P−values testing Hardy−Weinberg Equilibrium are printed for each level of the first trait.
118. **kruskal−wallis** <*quantitative trait*> <*trait*>. Print table of means for the quantitative trait for each level of *factor*, along with the Kruskal−Wallis chi−square.
119. **regress** <*ytrait*> = <*x1*>...[**to**]... <*xN*> [**offset** <*offset*>] [**poisson**|(**exponential**|**weibull** [<*censoring_trait*>)]. Performs linear or logistic or poisson or weibull regression of trait *ytrait* on set of loci *x1...xN*. If an x variable is a marker genotype, that independent variable is the mean allele size in the genotype, with the exception of the first marker locus encountered in the list, which is fully allelic effect coded. The **offset** option reads an offset for the linear predictor from the specified trait. Addition of a binary trait name to the end of the keyword list when the regression is **weibull** or **exponential** declares this as the censoring indicator.

120. **mixture** <*quantitative trait*> [<Number of distributions> [normal|pooled_normal|exponential|poisson]]. Estimate mixing proportions, means and standard deviations for a 1..5 component mixture model describing the specified quantititative trait. The default is a mixture of Normal (Gaussian) distributions with different means and variances, but a common variance can alternatively be specified. Other distributions available are the exponential and Poisson. A line−printer type histogram is produced.
121. **kinship** [**inbreeding**|**pairwise**|<*binary trait*> [|<*quantitative trait*> [>|>=|<|<=|==|^= <*threshold*>]]. Write the numerator relationship matrix (matrix of coefficients of relationship) for each pedigree in a lower triangular form or as a list of pairs (in the latter case, the coefficient of fraternity is also printed). Alternatively, if requested, print a list of individuals with a non−zero inbreeding coefficient. If a binary trait is specified, the NRM is only for the affecteds if *plevel*=1; for *plevel*=0, only a summary for each pedigree is printed: number of affecteds, number of "sporadic" cases ie cases unrelated to any other affected family members (eg marry−ins with no affected descendants), mean coefficients of relationship for affected relative pairs and of inbreeding for cases.
122. **ibd** <*codominant marker*> [**pairwise**]. Write the estimated mean identity−by−descent sharing at a marker for all relative pairs in a pedigree as a lower triangular matrix or a list of pairs.
123. **ibs** <*codominant marker*> [**pairwise**]. Write the estimated mean identity−by−state sharing at a marker for all relative pairs in a pedigree as a lower triangular matrix or a list of pairs.
124. **hwe** [**founders**]. Prints chi−square statistic for Hardy−Weinberg equilibrium for all marker loci. Analysis may be restricted to founders, and if the marker is diallelic, an exact test is carried out. If nonfounders are included, then a gene−dropping simulated P−value is produced. The mean IBS sharing for all typed matings is also calculated, and compared to its expected value. This latter test may allow detection of homogamy or assortment.
125. **cksib**. Lists all sib pairs, and the mean of IBS at all *marker* loci where both members of the pair are typed at the marker, comparing this to that expected if related as specified by the pedigree structure. The output is to the standard output.
126. **share** [**pairs**]. Lists all relative pairs, and the mean of IBS at all *marker* loci where both members of the pair are typed at the marker, comparing this to that expected if related as specified by the pedigree structure, allele frequencies and linkage map. The output is to the standard output. The default lists individuals whose Z score measuring deviation from expected exceeds 1.65 with any other relative. The **pairs** option prints the statistic for each deviant pair, or all pairs if output is set to verbose.
127. **mztwin** <*monozygosity_indicator*> |(<*zygosity_score*> >|>=|<|<=|==|^= <*threshold*>) [**clean**|**delete**]. Using a binary or quantitative trait which indicates which sib pairs are monozygotic twin pairs, list markers at which the twins carry discordant genotypes. Gives proportion discordance for each marker. This is useful for estimating genotyping error rates. The **clean** option deletes genotypes for pairs where there is an inconsistency, and fills in missing genotypes where that for the cotwin is available. The **delete** option drops the member of the pair with the fewest nonmissing phenotypes, and averages (across the pair) quantitative phenotypes where both are observed.
128. **davie** <*binary trait*> [<*proband indicator*>]. Print segregation ratios and standard errors for a binary trait, adjusted for the ascertainment scheme. Probands are indicated by being *affected* at the proband indicator "locus". If no proband indicator variable is given, complete ascertainment is assumed (equivalent to "davie trait trait").
129. **varcomp** <*quantitative trait*> [**ae**] [**ce**] [**ace**] [**ade**] [**covariate** <*covariate trait 1*> ... <*covariate trait N*>]. Performs MVN variance components analysis for a quantitative trait using all phenotyped individuals. Currently fits ADE, ACE, AE and CE models. The **ae** option fits AE and E only, and so forth. Multiple covariates can be included in the fixed effects part of the model via adding the **cov** keyword and a list of covariates at the end of the command line. Only the first marker locus in the list of covariates is fully allelic effect coded, with subsequent markers included as the mean of their allele values (i.e. 1="1/1", 1.5="1/2", 2="2/2" for a diallelic marker).
130. **lrt**. Constructs likelihood ratio test comparing the last two variance components (**var**) or generalized linear (**reg**) or generalized linear mixed (**fpm**) models fitted. The compared likelihood statistics from

fpm are actually the mean model loglikelihoods. This allows one to carry out tests of linkage after conditioning out the effects of genotype at a candidate polymorphism, and traditional "measured genotype" association analysis in pedigrees, including non−normal data (currently binomial and poisson distibuted traits).

131. **blup** <*quantitative trait*> <*h2*>. Calculates BLUPs (best linear unbiased predictions of the breeding value) for a quantitative trait assuming the given heritability.
132. **fpm** <*quantitative trait*> [>|>=|<|<=|==|^= <*threshold*>]|<*binary trait*>
[**nqtl** <*number simulated QTLs*>]
[**link logit**|**probit**|**mft**|**ln**]
[**likelihood_family gaussian**|**binomial**|**poisson**]
[**p**|**fre**] [**a**|**va**] [**d**|**vd**] [**g**|**vg**|**h2**] [**c**|**vc**] [**s**|**vs**]
[**fix a**|**c**|**d**|**e**|**g**|**mu**|**s**|**var**]
[**aval**|**cval**|**dval**|**eval**|**gval**|**mu**|**pval**|**sval**| **var**|**AA**|**AB**|**BB**|**SD** <*value*>]
[**cov** <*x1*> [+ <*x2*>...]].
Uses a Monte Carlo Markov Chain (Metropolis) algorithm to perform Generalized Linear Mixed Model analysis (when **nqtl** set to zero), complex segregation analysis (**nqtl** set to one), "genetic" mixed model analysis (**nqtl** set to one, **g** estimated), or a finite polygenic model analysis for a quantitative or binary trait using all phenotyped individuals. The **a**, **d**, **g**, **c** and **s** options respectively include additive QTL effects, dominance QTL effects, additive Gaussian polygenic random effects, pedigree environmental effects and maternally derived sibship effects in the model. The **fix** option allows that variable to be held fixed. The **aval** (**dval** etc) keyword allows the starting value of that parameter to be set. Either a Gaussian, Binomial, Weibull or Poisson likelihood can be used. In addition to the canonical links for these three types, one can also fit the multifactorial threshold model to binary data. Because the identity link function for a binomial likelihood is of genetic interest, notably in the case of the single major locus model, special checks to reject models where predictions lie outside 0−1 are made, so that this model can be successfully fitted.
133. **dis** [[<*marker locus 1*>] <*marker locus 2*>]. Estimates frequencies of two locus haplotypes based on individuals with two informative parents. The loci are assumed to be tightly linked, so that four parental haplotypes may be inferred. If no markers are specified, the sequence of pairs of markers in the list of loci is produced (ie marker1 with marker2, marker2 with marker3...). If one marker is specified, then the pairing of all other markers with this index marker is analysed.
134. **assoc** <*binary trait*> [|<*quantitative trait*> [>|>=|<|<=|==|^= <*threshold*>]] [**founders**] [**covariate** <*covariate*>] [**genotypic**]. For a binary or dichotomised quantitative trait, prints chi−square statistics for association for all marker loci versus the trait, either *affected* versus *unaffected* if the trait is binary, or above or below the threshold if the trait is quantitative. A second table in the output shows the results from the reconstruction−combined TDT within informative sibships. Also prints F statistics assuming trait is marker for different subpopulations. For a quantitative trait, prints the model and residual sums−of−squares and allelic means with naive standard errors from an additive allelic ANOVA model. Monte−Carlo empiric P−values are produced for either analysis. Analysis may be restricted to founders. Genotypic rather than allelic analyses can also be specified, using the *genotype* flag. Covariates can be added to some analyses.
135. **homoz** [<*binary trait*> [|<*quantitative trait*> >|>=|<|<=|==|^= <*threshold*>]]. Prints the asymptotic Z statistic and a one−sided MC P−value for whether homozygosity at each marker locus is increased in probands, either affected if the trait is binary, or above or below the given threshold if the trait is quantitative.
136. **multihomoz** [<*binary trait*> [|<*quantitative trait*> >|>=|<|<=|==|^= <*threshold*>]]. Prints the asymptotic Z statistic and a one−sided MC P−value for whether the maximum length of runs of homozygosity at marker loci along the specified map is increased in probands, either affected if the trait is binary, or above or below the given threshold if the trait is quantitative.
137. **tdt** <*binary trait*>|(<*quantitative trait*> >|>=|<|<=|==|^= <*threshold*>) [**cutoff** <*cutoff*>] [*mat*|*pat*]. Prints transmission−disequilibrium statistics for all *marker* loci versus the *trait*, where an index

person is either *affected* with a binary trait, or whose value for a quantitative trait exceeds the given threshold. Since binary traits are coded internally as 2=*y* and 1=*n*, an analysis using unaffecteds as proband can be performed as *tdt <binary trait> under 2*. Similarly, in unascertained families, *tdt <binary trait> over 0* tests for segregation distortion. Calculation of the TDT statistic can be restricted to pairs of cells whose total is greater than *cutoff*, eg 5, and to the maternal or paternal contributions, if parent−of−origin effects are suspected.

138. **hrr** <*binary trait*>|(<*quantitative trait*> >|>=|<|<=|==|^= <*threshold*>). Performs the Haplotype Relative Risk test.
139. **schaid** <*binary trait*> <*marker*> [<*allele*>]. Performs the Schaid and Sommer [1993] genotypic risk ratio test for familial association under the assumption of Hardy−Weinberg equilibrium, as well as the "Conditional on Parental Genotypes" version that is equivalent to the genotypic TDT. Only one allele (defaulting to the commonest) is tested versus all others, and two penetrance ratios ($GRR_2=f_2/f_0$ and $GRR_1=f_1/f_0$) are estimated, along with the LR chi−square test that $GRR_2=GRR_1=1$.
140. **asp** <*binary trait*>|(<*quantitative trait*> >|>=|<|<=|==|^= <*threshold*>). Prints IBS−based affected (full and half) sib−pair statistics for all *marker* loci versus the *trait*. It also prints the mean IBD sharing for full sibs, along with the exact (binomial) two−tailed P−value for the "mean" test. All possible sib−pairs are used, and are treated as independent.
141. **apm** <*binary trait*>|(<*quantitative trait*> >|>=|<|<=|==|^= <*threshold*>) [**ibd**|**ibs**]. Prints APM statistics for all *marker* loci versus the *trait*.
142. **sibpair**|**he1**|**he2**|**vis** <*quantitative trait*>| <*binary trait*> [<*Weight variable*>] [**sim**] [**mean**<*trait mean*>] [**var** | **sd** <*trait variance or SD*>] [**cor**<*trait sibling correlation*>]. Performs Haseman−Elston regressions (Sham & Purcell [2000] as the default, but Visscher−Hopper [Visscher & Hopper 2001], traditional and "new" Haseman−Elston also available) for all marker loci versus the trait using full and half−sib relative pairs. The contribution of each pair can be weighted by the mean of their values at a quantitative trait. Empirical P−values can be simulated, if requested. For the S+P regression, the "true" population trait mean, variance and sibling correlation can be specified, to facilitate analysis of selected samples.
143. **twopair** <*quantitative trait*>| <*binary trait*> <*marker locus 1*> <*marker locus 2*> <*theta12*>. Performs Fulker & Cardon's Haseman−Elston interval regression for first and second marker loci versus the trait using full− and half−sib relative pairs. The recombination distance between the markers theta12 must be given. Haseman−Elston regression is performed using ibd estimated at ten points in the interval.
144. **qtlpair** <*quantitative trait*> [**full**] [**covariate** <*covariate trait 1*>]. Performs variance components linkage analysis for all marker loci versus the trait using full−sib relative pairs, or if the **full** option is active, all genotyped individuals. Both the polygenic background and the QTL are modelled as additive genetic. Covariates, which can include codominant marker loci, are added using as **cov** keyword−trait pairs. Only the first marker locus is fully allelic effect coded, with subsequent markers included as the mean of their allele values (i.e. 1="1/1", 1.5="1/2", 2="2/2" for a diallelic marker).
145. **linkage** [<*marker locus 1*> [<*marker locus 2*>]]. Performs Elston and Keats sib pair linkage analysis for codominant markers. Default is adjacent pairs of markers (ie marker 1 with marker 2, marker 2 with marker 3...). If one marker is named, then gives estimate of recombination distance to all other markers.

The following script performs a number of analyses on a dataset containing four loci.

Test.in

```
set work c:\tmp\
set weight founders
set out verbose
set impute on
set locus quant quantitative
```

```
set locus trait affection
set locus marker1 marker
set locus marker2 nam
read pedigree test.ped
run
freq
mix quant 2
assoc trait
tdt trait
ass quant
sibpair quant
recode marker1 126 128 to 999
freq marker1
tdt trait
apm trait
sibpair quant
! create a new trait
set locus new_quant qua
if (quant le 0) then new_quant= −sqrt(−2*quant) else new_quant=log(quant)
! adjust the binned allele sizes of marker
if (marker1 ne 999) then marker1=marker1+1
drop trait quant
write pedigree testout.ped
```

DATASETS

The data set contains one record (newline character delimited) per individual. Records must be sorted into pedigrees. Records take the format used by GAS:

pedigree−id person−id father−id mother−id sex−of−person locus−value−1...locus−value− N

A pedigree ID may be up to 10 alphanumeric characters, and an individual's personal ID up to 8 characters. Missing values are denoted *x* (or .), and represented internally as a trait value of −9999. Locus values for a binary trait are *y* (expresses trait), *n* (does not express trait). Sex takes the values *m* (male) and *f* (female), and may be missing. Alleles at a *marker* locus are integers between 1 and 999 or single letters. A pedigree file may contain a comment at any time, prefaced by ! or #, and may contain a locus header of the form (though this has no function and is included to allow compatibility with the GAS pedigree format):

pedigree locus <*locus−name−1*>...<*locus−name−N*>.

If only one parent of an individual is specified in the pedigree file, a dummy record and ID number for the other parent is generated by the program.

Here is the data set analysed by the script *test.in*:

Test.ped

```
! test pedigrees including one halfsib in pedigree 1000
!                                        Marker 1  Marker 2
! The seven mating types:  1000−1 x 1000−2 Type VII  Type III
!                          1000−1 x 1000−3 Type VI   Type II
!                          1001−1 x 1001−2 Type IV   Type V
!
1000 1   x   x   m   10  y   126 132   1   1
1000 2   x   x   f   10  n   128 130   1   2
1000 3   x   x   f   25  n   128 132   2   2
```

```
1000 4   1   2   f   20  y   126 128   1   1
1000 5   1   2   m   30  y   130 132   1   1
1000 6   1   2   m   40  n   128 132   1   2
1000 7   1   2   f   50  n   126 130   1   2
1000 8   1   3   f   60  n   126 128   1   2
1000 9   1   3   m   40  y   132 132   1   2
1001 1   x   x   m   20  y   124 124   1   2
1001 2   x   x   f   30  n   126 128   1   2
1001 3   1   2   f   40  n   124 128   1   1
1001 4   1   2   m   30  n   124 126   1   2
1001 5   1   2   m   40  n   124 126   2   2
1001 6   1   2   m   40  n   124 128   1   2
1002 1   x   x   m   10  y     x   x   x   x
1002 2   x   x   f   40  n     x   x   x   x
1002 3   1   2   m   30  n   126 126   1   2
1002 4   1   2   m   60  n   126 126   1   1
1003 1   x   x   m   20  n   126 126   1   2
1003 2   x   x   f   25  n     x   x   x   x
1003 3   1   2   m   40  n   126 126   1   2
1003 4   1   2   m   15  y   126 126   1   2
! end−of−pedigrees
```

The pedigree file written by Sib−pair contains the original records plus any additional dummy records for missing parents of nonfounders. It is sorted by founder/nonfounder status, generation number, and the collation order of the parental IDs and individual ID.

TIPS AND TRICKS

How do I?

- *Get more output:* Set the print level higher.

```
#
# Get full TDT tables but summary results for association analysis
#
set ple 1
tdt trait
set ple 0
ass trait
```

- *Analyse only selected traits:* Use the *drop* then *undelete* commands to specify loci or ranges of loci to be included or exclude from particular analyses:

```
# Drop out year of birth and don't show any allele frequencies
drop yob $m
describe
undelete
```

- *Write out only the markers:* Use *keep $m*, then *write pedigree*. Note that this will undelete previously deleted markers.
- *Delete marker data for particular individuals:* Use *keep $m*, then *delete* <pedigree> <person>, followed by *und*.

```
...
set loc errprob qua
keep $m
```

```
delete where errprob>0.9
und
```

- *Use the new parser:* This hopefully reduces the number of steps in data manipulation required for a complete analysis.

```
#
# Create a new variable that is a function of
# three existing quantitative variables
#
set loc b1 aff
set loc q1 qua
set loc q2 qua
set loc q3 qua
read ped ex.ped
run
set loc new_var qua
new_var=log(q1+q2)/q3^2
if (male) then new_var=new_var+10
#
# Select a subset of pedigrees where two or more probands
# meeting multiple criteria
#
select containing 2 where new_var>35 and q1 le q2 and isnon
write newped.ped
```

- *Use wildcard selection:* This allows selection on pedigree or person names in a flexible fashion:

```
#
# Select first six pedigrees in file
#
>> select famnum<=6
>> wri

! Pedigree  Person   Father   Mother  x
!
a                a        x        x x
and            and        x        x x
are            are        x        x x
as              as        x        x x
at              at        x        x x
be              be        x        x x
```

print ped * id a.

id=as−as sex=x
id=at−at sex=x

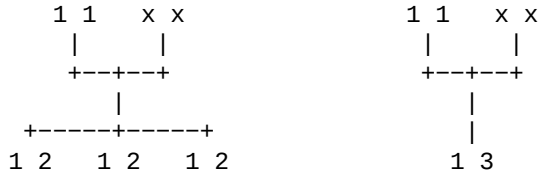
>> print ped a*e
id=are−are sex=x

>>print ped *s*
id=as−as sex=x

- *Use variable names in formulae when the variable name shares the first three letters with a command eg "trait" and "transform":* surround the variable name with brackets.
- *Analyse multiple pedigree files:* Jobs can be stacked, providing each begins with the *clear* command. The loci will have to be declared each time however.
- *Delete a marker that is giving too many mendelian inconsistencies:* Use the *drop* command on that marker before the *run* command.
- *Ignore error messages from a marker that is giving too many mendelian inconsistencies:* Drop the marker out before error checking as before, then use the *undelete* command before the hopefully robust type of analysis chosen. Alternatively *set checking off* and *set impute −1* will turn off checking for all markers.
- *Test for segregation distortion:* Do a TDT with everyone affected, for example, *set sex on* then *tdt sex over −1*.
- *Log transform a quantitative trait:* Use *tra trait 1 0 0* or *trait=log(trait)* (slower).
- *Get a histogram for a quantitative trait:* Use the *hist* command (this is a synonym for *mixture* trait 1). Setting the print level higher for *mixture* also prints out posterior probabilities of membership of the different distributions −− useful for choosing thresholds.
- *Print genotype frequencies:* Set *plevel* to 1 so that the full table is printed when the *hwe* command is used.
- *Remove unrelated individuals or nuclear families that have become disconnected from the main pedigree, although they have the same pedigree ID:* Use the *subped* command, followed by *select $n gt 20*, or some other suitable number that will keep only the main pedigree.
- *Do a multipoint ASP linkage analysis:* Write the appropriate format pedigree and locus file, and call another program like SIB_PHASE:

```
# Write a locus file
write locus aspex batch.tcl
# Write out the pedigrees as nuclear families (if multigenerational)
nuclear
write aspex batch.ped
# The resulting output will be included in the Sib−pair output
$ sib_phase −f batch.tcl batch.ped
$ rm batch.tcl batch.ped
```

- *Which TDT result should I trust?* The genotypic TDT is currently a more experimental test. In the case of nuclear families with typed parents, it reduces to a simulation based CPG GRR test. In larger pedigrees with multiple generations of affecteds, matings must have both parents genotyped before they are included in the simulation, but this is done over the complete pedigree.
- *What does the "founder" option for the allele frequencies give (updated)?* Since a simple count of alleles in typed founders can miss alleles segregating in the pedigree (and thus inherited from untyped founders), I have provided a method that enumerates all alleles in each pedigree, but weights the contribution of the pedigree by the number of founders it carries (that is, the number of representatives from the population whose allele frequencies one is trying to estimate). So, in a simple example,



the contributions from each family would be weighed equally, as each contains two founders. For allele 1 for example,

```
Family 1      Family 2
  2 * 5/8  +  2 * 3/4
 --------------------  =   11/16
   4 (total founders)
```

	Allele 1	**Allele 2**	**Allele 3**
The naive estimate is:	8/12 (.67)	3/12 (.25)	1/12 (.08)
The weighted estimate:	11/16 (.69)	3/16 (.19)	2/16 (.13)
The imputation estimate:	6/8 (.75)	1/8 (.12)	1/8 (.13)
The MLEs (MENDEL USERM13):	.6254	.2182	.1564
The MCEM MLEs (Sib−pair mcf):	.6250	.2201	.1549

In this example, the frequencies of the 2 and 3 alleles are better estimated by the weighted method than the naive method. In the imputation estimation approach, the untyped parents were imputed as 1/2 and 1/3 respectively. The MCEM estimate gives the MLEs within the limits of stochastic error. In general, providing there are enough pedigrees, the naive estimate is as good as any.

- *What do I do if I have covariates or liability classes? (updated)* Several quantitative trait analyses allow the Eventually the various binary trait analyses Sib−pair does will allow for covariates. At the moment, the best thing you can do is create multiple phenotypes eg male diabetes and female diabetes, with the phenotype set to missing appropriately. Then one can do the analysis within each stratum, and pool test statistics in various ways used in meta−analysis. In the case of quantitative traits, analysis of residuals formed by adjusting for covariates will carry you a long way.
- *Why can't I have variables called d1 or e1 etc?* Following Fortran conventions, d1 and e1 are read to mean 0d1 and 0e1 and evaluate to a real number: 0. You can have a1, f1, 1d, dd1 etc as names.
- *Does the fpm command actually work?* It does seem to in examples, but multiple runs should be performed, and those with high coefficient of variation of log likelihood looked at with a jaundiced eye. I have applied it now to a number of nongenetic generalized linear mixed model problem datasets, where it seems to give answers that roughly agree with those from other programs ;).

DOCUMENTATION OF ROUTINES

Regression (multiple) is performed by AS (Applied Statistics) 164, which uses modified Givens rotations to perform weighted least squares regression including linear constraints. It is also used to give generalised linear models (poisson and binomial regression) via IRLS. The random number generator is the well known AS 183. The approximate randomization routine is styled after general templates described by Noreen [1989]. Mixtures of distributions are fitted using AS 203. Various standard distributions are evaluated using AS 3, 66, 111. Likelihood maximization is performed using AS319 ("varmet") which seems to do a very good job of it.

LIMITATIONS

Sib−pair currently is limited to a maximum of 600 (or 1000 or 2000) individuals per pedigree, 120 (or 1000) locus values or fields (eg 60 codominant genotypes, 120 binary traits etc), and 40 (or 60 or 100) possible alleles per codominant locus. The "low memory" version of the program stores phenoset for each individual in a pedigree is stored in a direct access file, with resulting overheads. Write to me if this sounds of use. The standard version performs imputation in memory. The Monte−Carlo based routines are computationally intensive. Generally speaking 200−300 iterations of such a routine are sufficient to give a good estimate of a mean or variance (as in the apm routine), but 1000 iterations or more are advised for an accurate P−value.

Using "set iter 0" will provide only the parametric estimators, e.g. for screening purposes.

Sib−pair does not know about mitochondrial or Y DNA.

There are only a few multipoint procedures (*twopoint*, *multihomoz* and *haplotype*).

The program is (fairly) standard Fortran 77, and runs successfully on PCs (under DOS, NT or Linux), SUN Sparcstations, DEC Alpha and HP9000 workstations. The current DOS version is compiled via f2c and the DJGPP port of the GNU C compiler, and is compressed using djp, an LZ0 executable compressor.

ACKNOWLEDGEMENTS

This program was developed while the author was an Australian National Health and Medical Research Council Neil Hamilton Fairley Postdoctoral Fellow and later a Research Fellow. This included a period working in the Genetic Epidemiology Division of the Johns Hopkins University School of Public Health and in the Epidemiology Unit at the Queensland Institute of Medical Research.

REFERENCES

- Aitkin MA, Clayton D (1980): The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Appl Statist* **29**: 156−163.
- Andersen PK, Borgan O, Gill RD, Keiding N (1993): Statistical models based on counting processes. *New York: Springer Verlag*.
- Besag J, Clifford P (1991): Sequential Monte Carlo p−values. *Biometrika* **78**: 301−304.
- Bishop DT, Williamson JA (1990): The power of identity−by−state methods for linkage analysis. *Am J Human Genet* **46**: 254−265.
- Blangero J, Samollow PB, Rocha MB, Hixson JE, Rogers J (1995): The IGF1 locus is a major determinant of serum osteocalcin levels in Mexican Americans. *Fourth Annual Meeting of the International Genetic Epidemiology Society, Snowbird, Utah, June 20− 22, 1995*.
- David F, Johnson NL (1956): Some tests of significance with ordered variables. *J R Statist Soc B* **18**: 1−20.
- Davie AM (1979): The 'singles' method for segregation analysis under incomplete ascertainment. *Ann Hum Genet* **42**: 507−10.
- Davis S, Schroeder M, Goldin LR, Weeks DE (1996): Nonparametric simulation−based statistics for detecting linkage in general pedigrees. *Am J Hum Genet* **58**: 867−80.
- Excoffier L (2001): Analysis of population subdivision. *In:* Balding DJ et al. Handbook of Statistical Genetics. *London: Wiley and Sons*. 271−307.
- Fain PR (1977): Characteristics of simple sibship variance tests for the detection of major loci and application to height, weight and spatial performance. *Am J Hum Genet* **42**: 109−20.
- Falk CT, Rubinstein P (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* **51**: 227−233.
- Filliben J (1975): The probability plot correlation coefficient test for normality. *Technometrics* **17**: 111−117.
- Guo SW, Thompson EA (1994): Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics* **50**: 417−432.
- Haberman SJ (1979): Analysis of quantitative data. Volume 2. New developments. *New York: Academic Press*.
- Haseman JK, Elston RC (1972): The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**: 3−19.
- Hedges LV, Olkin I (1985): Statistical methods for meta− analysis. San Diego: Academic Press.

- Jones GL, Haran M, Caffo BS, Neath R (2005): Fixed−width output analysis for Markov Chain Monte Carlo [Preprint]. *http://www.stat.umn.edu/~galin/mcse_rev.pdf*
- Kaplan NL, Martin ER, Weir BS (1997): Power studies for the transmission/disequilibrium tests with multiple alleles. *Am J Human Genet* **60**: 691−702.
- Keats BJ, Elston RC (1986): Determination of the order of loci on the short arm of chromosome 11 using two and three locus linkage analyses of pedigree and sib pair data. Genet Epidemiol Suppl **1**:147−52.
- Knapp M, Seuchter SA, Baur MP (1993). The haplotype−relative−risk (HRR) method for analysis of association in nuclear families. *Am J Hum Genet* **52**: 1085−1093.
- Knapp M, Wassmer G, Baur MP (1995): The relative efficiency of the Hardy−Weinberg Equilibrium−likelihood and the Conditional on Parental Genotype−likelihood methods for candidate−gene association studies. *Am J Hum Genet* **57**: 1476−1485.
- Knapp M (1999): The transmission/disequilibrium test and parental−genotype reconstruction: the reconstruction−combined transmission/disequilibrium test. *Am J Hum Genet* **64**: 861−870.
- Kruglyak L, Daly MJ, Reeve−Daly MP, and Lander ES (1996): Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* **58**: 1347−1363.
- Laird N, Horvath S, and Xu X (2000): Implementing a unified approach to family based tests of association. *Genetic Epi* **19**(Suppl 1): S36−S42.
- Lange K (1986a): The affected sib−pair method using identity by state relations. *Am J Hum Genet* **39**: 148−150.
- Lange K (1986b): A test statistic for the affected−sib−set method. *Ann Hum Genet* **50**: 283−290.
- Lange K, Goradia T (1987): An algorithm for automatic genotype elimination. *Am J Hum Genet* **40**: 250−256.
- Lange K, Matthysse S (1989): Simulation of pedigree genotypes by random walks. *Am J Hum Genet* **45**: 959−970.
- Lange K (1997): Mathematical and statistical methods for genetic analysis. New York: Springer−Verlag.
- Noreen EW (1989): Computer−intensive methods for testing hypotheses: an introduction. *New York: Wiley*.
- Olson J (1995): Robust multipoint linkage analysis. An extension of the Haseman−Elston approach. *Genet Epidemiol* **12**: 177−194.
- Olson J, Rao S, Jacobs K, Elston RC (1998): Linkage of chromosome 1 markers to alcoholism−related phenotypes by sib−pair linkage analysis of principal components. Genetic Analysis Workshop 11. September 8−10, Arcachon, France.
- Pons O, Chaouche K (1995): Estimation, variance and optimal sampling of gene diversity. II. Diploid locus. *Theor Appl Genetics* **90**: 122−130.
- Ripley BD (1987): Stochastic Simulation. *New York: Wiley*.
- Resek RW (1974): Alternative tests of skewness: Efficiency comparisons under realistic alternative hypothesis. *Proc Bus Econ Statist Section Am Statist Assoc* **1974**: 546−551.
- Royston P (1993): A pocket−calculator algorithm for the Shapiro−Francia test for non−normality: An application to medicine. *Statist Med* **12**: 181−184.
- Schaid DJ, Sommer SS (1993): Genotype relative risks: methods for design and analysis of candidate−gene association studies. *Am J Hum Genet* **53**: 1114−1126.
- Sham PC, Purcell S (2001): Equivalence between Haseman−Elston and Variance−Components Linkage Analyses for Sib Pairs. *Am J Hum Genet* **68**:1527−1532.
- Spielman RS, McGinnis RE, Ewens WJ (1993): Transmission test for linkage disequilibrium: the insulin gene region and insulin−dependent diabetes mellitus. *Am J Hum Genet* **52**: 506−516.
- Spielman RS, Ewens WJ (1996): The TDT and other family−based tests for linkage disequilibrium and association [editorial]. *Am J Hum Genet* **59**: 983−989.
- Steele F, Diamond I, Amin S (1996): Immunization uptake in rural Bangladesh: a multilevel analysis. *Journal of the Royal Statistical Society, Series A* **159**: 289−299.

- Therneau TM, Grambsch PM, Fleming TR (1990): Martingale−based residuals for survival models. *Biometrika* **77**: 147−160.
- Visscher PM, Hopper JL (2001): Power of regression and maximum likelihood methods to map QTL from sib−pair and DZ twin data. *Ann Human Genet* **65**: 583−601.
- Weeks DE, Lange K (1988): The affected−pedigree−member method of linkage analysis. *Am J Hum Genet* **42**: 315−326.
- Ward PJ (1993): Some developments on the affected−pedigree−member method of linkage analysis. *Am J Hum Genet* **52**: 1200−1215.
- Ward PJ, Bonaiti−Pellie C (1995): Measuring gene−disease association using a general pair method. *Genet Epidemiol* **12**: 681−686.
- Whittemore AS, Halpern J (1994a): Probability of identity by descent: computation and applications. *Biometrics* **50**: 113−117.
- Whittemore AS, Halpern J (1994b): A class of tests for linkage using affected pedigree members. *Biometrics* **50**: 118−127.
- Yates F (1948): The analysis of contingency tables with groupings based on quantitative characters. *Biometrika* **38**: 176−181.
- Young A (1995): Genetic Analysis System, version 2.0 [Computer program]. *Oxford: Oxford University*.

PROGRAM HISTORY

21−Jun−2006 (1.00a17)

Added "m" modifier as in "$mm" to give markers in map order. This affects "ls", "lis", and "order". The "recode <marker> fre" command renames that markers' alleles from *1..N* ordered by allele frequency; "recode" also accepts a single wild card or class eg $m. The "marcom" function now works correctly. Merlin data file declared "Zygosity" indicator automatically sets the twin indicator variable if "read locus merlin" is used.

20−Jun−2006 (1.00a17)

Corrected sex shapes for "wri dot". Added in "marcom" function to count up maximum number of typed markers shared between ego and his relatives.

15−Jun−2006 (1.00a17)

Fixed bug in "tab" tabulation of SNP genotypes by levels of a trait (segfaulting if badly behaved SNP eg no heterozygotes). Twinning indicator now works after a "pack" or "reorder". If the string "MZ" is encountered while reading a quantitative trait, this is assumed to indicate an MZ twin, and is converted to a "1". The behaviour of "set twin" has been altered, so that positive values of the indicator variable are taken as belonging to an MZ twin pair.

13−Jun−2006 (1.00a16)

Output from "test" includes sex and mean heterozygosity. Output from "ls" ends with count of active traits and active markers. The "drop"/"keep" command now allows selection of every Nth locus. The locus file for eclipse2 and eclipse3 can be written. Parser limitation on number of loci upped to 100000.

07−Jun−2006 (1.00a16)

The "set nde" number of decimal places is now respected when quantitative traits are written in Mendel and Fisher pedigree format files (as f8.d). If alleles are already consecutively numbered (1..*numall*), then the "wri lin <fil> num" will be faster (especially for many markers). The "mzt" command now also checks if putative MZ twins are same−sex.

01−Jun−2006 (1.00a16)

Command lines now stripped of nonprinting characters, so DOS files under Unix don't choke the reading of data.

31−May−2006 (1.00a16)

Pointer to first child in post−Makeped Linkage−format pedigree file was not always to first child (missing brackets in if statement). And binary trait loci not being included. Locus annotations not carried along by "ord" reordering of loci.

26−May−2006 (1.00a16)

Finally got around to smart truncation of locus names for MENDEL's eight character limit.

25−May−2006 (1.00a16)

Simulation of a quantitative trait now respects a request it be linked to a marker. If sex−linked markers are present, they are now used to test the designated sexes as the pedigree is read in. The "test <ped> <id>" command tabulates multilocus IBS similarity of an index individual with other pedigree members and the most similar individual from the rest of the dataset (allowing sample duplications and possibly mixups to be found). The set of active markers can be thinned so they are all at least a minimum distance apart (eg if using a dense set of SNPs for linkage) using "keep dis <gap>".

19−May−2006 (1.00a15)

Minor prettification of output. Many commands now allow the trait number to be given instead of the name. Fixed "pack" −− did not correctly deal with annotations. Reorganised work arrays (ord, wloc), so SNP version of Sib−pair (32000 columns of data) works reliably when keep/drop loci based on annotations ("keep|drop where <search string>").

17−May−2006 (1.00a15)

Bug fixed in list of IDs printed out by connect() −− this is produced as the pedigree is being read in when the *plevel*>1. This did not affect "gener" or "subped". Thanks to Audrey Grant for pointing this out.

15−May−2006 (1.00a14)

Fixed bug in test for Mendelian inconsistenceies due to genotypes arising from evaluation of an expression −− array not declared. Fixed bug in "write linkage" due to increased allowable length of ID strings. Table row names from "tab" now respect the number of decimal places set via "set ndecimal".

08−May−2006 (1.00a13)

Exact biallelic locus HWE test for unrelateds added (accessed via "hwe 2", "hwe founders" (when *numal*=2) or via "tab <trait> <biallelic_marker>. The command "tab <trait> <biallelic_marker> gives a table of genotypic counts, allelic proportions and exact HWE P−values for each stratum of the trait (for convenient analysis of case−control SNP association studies). The "tabulate" command also now prints the "Mantel−Haenszel" trend test (Yates 1948) for RxC contingency tables (ie assuming an ordinal by ordinal model holds true). Replacement genotypes generated by expressions are now tested as to whether they give rise to Mendelian inconsistencies in each pedigree, the action taken depending on the value of "error_drop".

11−Apr−2006 (1.00a12)

Individual IDs can now be up to 10 characters in length. Fixed bug in evaluation of expressions involving genotypes −− "untyp" was not working correctly (always false). If simple operation involving a missing genotype (eg addition of a constant), the result is now a missing genotype. Bug in "wri ppd" fixed.

17−Mar−2006 (1.00a11)

The "keep" and "drop" commands cleaned up slightly −− the "where" condition can be a search string for the marker annotations (eg select all markers with "chr 6" in description). The "dis" command no longer segfaults if there are no marker loci in the file.

16−Mar−2006 (1.00a10)

Write post−Makeped linkage files with "wri ppd". Nicer output from "edit", which *does* allow wild card searches eg "edit * 0001 val to x". To obtain numerical sequential IDs for all individuals (instead of 1000*<ped>+<pos_in_family>), "uni seq".

15−Mar−2006 (1.00a10)

Fixed bug in "tab": change from single to double precision meant some categories were not equal do to precision problems. Metropolis slice sampling merged in.

13−Mar−2006 (1.00a9)

Fixed bug in fpm(): genotype chain inaccurate when only one family (metropolis criterion is versus last global update of likelihood rather than last local update).

28−Feb−2006 (1.00a8)

Fixed newly introduced bug in select() −− never excluded any pedigrees.

27−Feb−2006 (1.00a7)

Genotypes can now be included in expressions, if quoted (eg if(apoe=="3/4")). Added "ishom" , "ishet", and "alla" ("allb") to access the first (and second) alleles of marker genotypes. At the moment, unfortunately, alla returns the numeric value for a letter allele (A=10065 etc, so that "y/y" eq alla "y/y" does work!).

20−Feb−2006 (1.00a7)

Weibull added to "fpm".

13−Feb−2006 (1.00a6)

Weibull regression added to "reg".

6−Feb−2006 (1.00a5)

Fixed newly introduced oneseg() ("fpm") bug where likelihood for additive polygenes sometimes incorrectly calculated (when likelihood ratio for a proposal only evaluated for changed individuals, rather than recalculated for entire pedigree).

3−Feb−2006 (1.00a4)

Fixed segsim() ("fpm") bug that gave individuals with missing covariates the wrong imputed (covariate mean value) values. Moved all data from single precision to double precision, so that large integer data is represented correctly (notably dates encoded as YYYYMMDD). Added in "date" (and "set epoch") command for moving between Gregorian and Julian dates. The "last" command shows the command history and allows replaying a selected command. The command history is saved to a file "sib−pair.log".

30−Jan−2006 (1.00a3)

Fixed bug setting fixed effects bounds too narrow, and neatened intermediate MCMC parameter output.

30−Jan−2006 (1.00a2)

MCMC batch size now defaults to a theoretical optimum, the square root of the total number of (non−burnin) iterations [Jones et al 2005]. The variance components are now estimated by two different methods: from the variance of the simulated individual and group random effects; and a "direct" MLE via MCMC (the originally implemented method). Comparison of the two estimates can be used as a convergence diagnostic. Averaging over multiple MCMC chains for the individual random effects is implemented by duplicating records and appropriately adjusting the likelihood contributions. The number of chains is controlled by "set chain".

20−Jan−2006 (1.00a1)

Added a random effect shared by offspring of the same mother (S). Poisson and binomial GLMMs working for "fpm". The "pri" modifier to "fpm" prints out replicates of the simulated random effects for the pedigrees. The "reg" command now allows poisson regression and the specification of an offset. The first marker included in a "reg" analysis now automatically receives dummy allelic encoding. Tweaking of the drop() MCMC genotype routine seems to have made it more robust.

24−Dec−2005 (0.99.9)

Improved starting values for "var" −− occasionally Q stuck at zero. Fixed half−sib IBD for "sib" (occasionally was missing ie treated as −9999).

23−Dec−2005 (0.99.9)

Added CE and ACE variance components models to "var" (C is a familial environment random effect).

22−Dec−2005 (0.99.9)

The "var" and "qtl full" commands now allow fixed effects. The first marker locus in the covariate list is encoded as N−1 allelic effects (other markers are encoded as the mean allele size, which is fine for diallelic markers and certain types of repeats). The "lrt" command compares the last two VC models fitted (allowing tests of fixed effects).

16−Dec−2005 (0.99.9)

The drop() MCMC genotype routine now includes a local (Gibbs) conditional update as one of the alternated proposal−acceptance methods. This improves efficiency in large pedigrees where there are many untyped individuals.

09−Dec−2005 (0.99.9)

The "hbd" command estimates single−locus homozygosity−by−descent. The "mcf" command produces MLEs for marker allele frequencies using an MCEM algorithm ("set emi" to alter the number of EM iterations). Locus annotations (text following map position in the locus declaration) are saved and displayed where appropriate (currently "inf", "sho map"). Documented the "head" command.

14−Nov−2005 (0.99.9)

Further revision and testing of "fpm". Folded in old code to generate BLUPs ("blu"). The "set tune" command adjusts the single tuning parameter for the MCMC proposal distribution variance for quantitative variables. The "help" keyword search is now case insensitive.

13−Oct−2005 (0.99.9)

Revised "fpm" and its interface.

5−Oct−2005 (0.99.9)

The "dro" command can now drop based on a condition, either where markers are monomorphic ("whe mon"), nearly monomorphic ("whe max <frq>"), or the number of individuals typed is below a threshold ("whe num <ntyp>").

29−Sep−2005 (0.99.9)

The "mcm" command lists the series of genotypes for selected individuals in the MCMC chain. Currently, this is to allow diagnose mixing problems etc. Tidied up output from "schaid" test. The "hrr" routine now skips monomorphic markers. The "dis" and "ld" commands now print out r−squared when *plevel*=0. The number of attempts to generate starting genotypes (for MCMC routines) can now be increased ("set start <num>) above 5000 −− this is sometimes needed for big pedigrees.

21−Sep−2005 (0.99.9)

The "mul" command does a form of multipoint (IBS) homozygosity mapping.

12−Sep−2005 (0.99.9)

The "dis all" command now can evaluate arbitrarily large sets of markers (was stopping after only 1000 pairs).

31−Aug−2005 (0.99.9)

"set sex on" didn't set imputed sexes for unincluded parents correctly (Andrew Birley found this). Fixed. The "wri var" command writes a list of quantitative trait names to a file for MENDEL. Finally document the "rel" command to print out relatives of an individual. By "set ski", one can skip *N* lines at the beginning of a pedigree file.

27−Jul−2005 (0.99.9)

"kin <trait>" gives a numerator relationship matrix for cases, or the average relatedness and inbreeding of cases along with a count of "sporadic" cases ie cases unrelated to any other affected pedigree members.

18−Jul−2005 (0.99.9)

Average inbreeding within pedigrees is now for all nonfounders (was previously average of nonzero coefficients, as is not uncommonly seen). The "ancestry" command calculates average inbreeding for affected pedigree members only. "hrr" analysis now gives simulated P−values. The "wri csv" command can write out files suitable for immediately reading into statistical programs or spreadsheets; "wri sol" extends this to write out the various format comma delimited files that SOLAR requires. The "fpm" command runs a MCMC finite polygenic model −− set to one QTL, it performs classical segregation analysis, but needs to be run over a grid of QTL allele frequencies at present. The "read map" command reads marker map positions from a file, guessing the format based on the first two lines (recognizes MERLIN and MENDEL formats at least). The "read loc merlin" command reads a MERLIN format ".dat" file.

28−Feb−2005 (0.99.9)

Feng−Shen Kuo pointed out an error in the abbreviated TDT output (when compared to the output P−values for the Ewens test when plevel=1). The incorrect degrees of freedom were being used in the abbreviated output.

25−Feb−2005 (0.99.9)

Fixed "residuals" command (not recognising missing values). For binary trait analyses (eg tdt, apm), quantitative traits can now be tested for equality or nonequality with a constant.

06−Jan−2005 (0.99.9)

Fixed imputation in pedigrees containing loops (imputation level increased in the pedigrees following in the file).

20−Dec−2004 (0.99.9)

The "edit" command now knows about letter alleles. The "hrr" command performs a basic haplotype relative risk analysis.

08−Sep−2004 (0.99.9)

Andrew Birley has pointed out two bugs in the Schaid test: chi−square was double the correct one (!); problems with the HWE based test were due to the offset vector not being fully initialized.

21−Jun−2004 (0.99.9)

With "set wei imp" on, X chromosome marker frequencies were not excluding a male dummy second allele. Thanks to Latchezar Dimitrov for pointing this out.

3−Jun−2004 (0.99.9)

Fix: name of file to be written to can be up to 80 characters long (same as input).

24−May−2004 (0.99.9)

Fixes: letter alleles in haplotypes not correctly printing for "dis"; table of paternal v. maternal genotypes incorrect for autosomal loci.

20−May−2004 (0.99.9)

Fixed problem with handling of half−sibs in Haseman−Elston and related approaches: half−sib pairs not contributing correctly to t−statistic calculation (thanks to Ziad Taib for pointing this out).

19−May−2004 (0.99.9)

The genotypic association table now has genotypes rather than indices. Letter alleles written to pedigree formats that support them eg MENDEL.

7−May−2004 (0.99.9)

Letter codes for alleles now read and displayed in results transparently. The "mztwin" command enumerates discordant genotypes for sib pairs indicated to be monozygotic twins. The "wri loc lin" now prints more than 100 (increased to 1000) locus positions on the "locus order" line.

23−Apr−2004 (0.99.9)

The "flip" command recodes SNP alleles to their complement if they are nucleotide codes (ACGT). Adding "r" to a class eg "$mr" gives that class in reverse order, with its main use being inverting a linkage map. Logistic regression now calculates empirical P−values but is hideously slow (to minimize memory, it reads/writes lots of scratch files). The "recode" command, if applied to a marker without a "to" statement recodes the alleles to 1..*N*. Familial correlations versus sex now produced eg father−son (at Manuel F's request). The "dis all" command gives LD measures for all pairs of markers.

18−Feb−2004 (0.99.9)

The "inf" command now summarizes the number of selected pedigrees and number of available values for each variable. The "hwe" command now tabulates husband versus wife genotypes, if the print level is 1 or more.

12−Feb−2004 (0.99.9)

Sex−specific parent−offspring and sibling correlations now included in output from "des". The Genehunter locus file code for covariates ("4 0 # name") gives a quantitative trait.

21−Jan−2004 (0.99.9)

Log−linear LD models for X−linked markers working correctly. Note that "ld" gives results from the old phased−only approach. The logistic regression allelic association model now admits covariates (it still doesn't gene−drop). Assorted cleanups for output (eg no more excessive assignment outputs to missing from evaluations).

8−Jan−2004 (0.99.9)

Fixed log−linear LD models. The "qtl" VC linkage analysis can now use data from all members of a pedigree (the single marker ibd sharing is estimated via MCMC as elsewhere). Help now allows wild card searching. Wild card searching correctly deals with partial matches preceded by a wild card (would skip rest of that word). Permutation P−values for RxC contingency tables are now calculated (and tables may now be entered from the command line "chi"). Quantiles printed for "his". Loci may be renamed via the "ren" command.

18−Dec−2003 (0.99.9)

Added log−linear models for modelling LD for unphased genotypes or mixtures of phased and unphased data. Both this and HWE testing will also accept a table entered at the command line. The "ls" command now accepts locus names, types, ranges and wildcards, as do any commands using loadnam(). Timings for each command are produced by "set timer on".

24−Nov−2003 (0.99.9)

Fixed mean and SD displaced for Kruskall−Wallis test (these assumed each value was observed only once −− calling moment() instead of dssp()).

11−Nov−2003 (0.99.9)

Fixed assignment from a deleted (dropped) marker locus: these evaluated to the value of the first allele only.

24−Oct−2003 (0.99.9)

Minor fixes, eg stops leaving behind a binass() work file. SIGINT (^C) now stops operation of most commands and returns control to the main loop. Conditional statements now test both alleles of a marker, but note operations are still on first allele at all markers in expression, then all second alleles.

13−Oct−2003 (0.99.9)

The "recode" command now deals cleanly with the case of recoding alleles at a marker to missing. With the introduction of a "order" command, loci can reordered for analysis and output, and the "read loc lin" command now respects the locus order line. Errors in command usage now elicit a more helpful message. The beginnings of a logistic regression based association approach are included, but this does not give gene−drop empirical P−values yet.

9−Sep−2003 (0.99.9)

Little cleanups in output: "print" now respects the set width and number of decimal places; the "fre snp" gives N rather than 2N; maximum D' allows negative values.

21−Aug−2003 (0.99.9)

Fixed X−linked TDT for mother to son transmissions (was dropping those where the paternal genotype was inconsistent with autosomal transmission). This was already fixed for the "dis" command, but not propagated through to "tdt". Nader Deeb helped sort this out.

18−Aug−2003 (0.99.9)

X−linked TDT default includes males where father untyped (previously had to "set tdt 1"). Fixed the HWE test: had mixed up the iteration through the lower triangle again! but am sure this used to give correct results at one time.

07−Aug−2003 (0.99.9)

Fixed genof3() for X−linked markers.

30−Jul−2003 (0.99.9)

Output file misspecification now trapped. Added selection on individual as well as on pedigree ID: "select id in <list>". Documented the "fre snp" option to summarize information about SNPs.

29−Jul−2003 (0.99.9)

The "select ped" and "print ped" now support wildcard identifiers.

18−Jul−2003 (0.99.9)

Fixed bug in selecting classes of locus type to undelete i.e. "`und $m`" now works correctly.

17−Jul−2003 (0.99.9)

The "sim" command simulates an autosomal marker, which may be completely linked to an existing marker locus. In the latter case, the new marker may be perfectly informative.

11−Jul−2003 (0.99.9)

Where sexes of parents misspecified (eg male x male mating), the correct sibship is now identified. The "unique_id" command generates new pedigrees and IDs (1..nped).

26−Jun−2003 (0.99.9)

HWE test now handles X−linked markers.

26−May−2003 (0.99.9)

Fixed infinite loop occasioned by last pedigree in file containing a pedigree error.

24−April−2003 (0.99.9)

The "del" command now accepts a logical expression defining those individuals whose data are to be deleted. The loci to be deleted can be given as a list. The "famnum" and "index" variables now contain the position of the family and that individual in the dataset.

17−April−2003 (0.99.9)

Prevented calculation of LD between X and autosomal markers.

10−April−2003 (0.99.9)

The "reg" command applied to a binary trait really does give the logistic regession. The "combine" command combines rare alleles at a marker into a single new allele.

09−April−2003 (0.99.9)

The "reg" command applied to a binary trait gives the logistic regression.

28−March−2003 (0.99.9)

Added "read locus linkage" to read locus information from a Linkage style .dat file.

20−March−2003 (0.99.9)

Recognise "/" as separating alleles in a pedigree file.

28−February−2003 (0.99.9)

Write "mainparams" and data file for Jonathan Pritchard's *structure* program ("wri str <datfil>", "wri loc str <locfil> <datfil>"). X chromosome TDT and allele frequencies, export locus files.

12−February−2003 (0.99.9)

Fixed bug in ordering of loci for "wri lin" (caused by introduction of X−linked marker class).

31−January−2003 (0.99.9)

Various changes to allow Mendelian error checking for X−linked markers. Defined "xmarker" as a class of variable.

23−January−2003 (0.99.9)

Fixed "tab", "dav", and "des" (for binary traits) so that they respect the new method of selection. Pedigrees are now marked as active or inactive, but not deleted −− need to "pack" (or write) for that. Quantitative variable printing in "pri" repaired.

22−January−2003 (0.99.9)

Fixed "edit" command fallout from addition of X−linked markers. Added "cas" command to divide pedigrees into unrelated cases and controls eg founders with information, or one child of parents with no information etc. The command "var <trait> ae" fits only the AE (and E) models.

17−January−2003 (0.99.9)

Genotypic association analysis option: "ass <trait> gen" gets a table of genotype rather than allele counts. The "unselect" command returns all pedigrees previously excluded by one or more "select" commands back into the analysis. A new type of marker "xmarker" added. Still sorting out the Mendelian error checking for this.

04−October−2002 (0.99.9)

Minimum intermarker map distance for Genehunter locus files fixed −− set to 0.01 cM.

18−October−2002 (0.99.9)

One too many recombination distances being written to LINKAGE locus files when a dummy locus specified. Thanks to David Evans for pointing that out.

17−October−2002 (0.99.9)

Improved nuclear family mendelian checker (back to former level!). Documented the "edit", "set errordrop", "set checking" and "delete" commands (as well as "prop" and "pchisq"). These shortcomings pointed out by David Evans. Added "wri map merlin" command. A "xlinked" flag added to the "write locus linkage" command.

4−October−2002 (0.99.9)

If "set err_drop 2", then long−distance errors cause the entire pedigree to be set to missing for that marker.

25−September−2002 (0.99.9)

After the "nuc gra" command, a parent with missing parents could come after a nonfounder parent in the work file, causing a segfault on subsequent analysis.

18−September−2002 (0.99.9)

The "select pedigree" command allows inclusion or exclusion of specific named pedigrees from further analysis. The "gener" command now only lists summary information at the default print level. The "write pap" was writing MCMC start marker genotypes to the PAP phen.dat as if they were observed genotypes −− Sandra Hasstedt helped sort this out. Added a "write map" command, currently only for MENDEL map files. Changed default tdt to "both" −− both parents must be present.

29−July−2002 (0.99.9)

The "prune" command reduces a pedigree to the probands and a minimum number of connecting relatives. The "ancestor" with increased print level gives number of affected descendants for all individuals. Some code reorganisation.

01−July−2002 (0.99.9)

The "rank" command writes the ranks for a variable.

26−June−2002 (0.99.9)

Results of "and" and "or" operations with missing values are now more consistent (eg *F && x = F*). Bug in "dis <marker>" form of call to "dis" fixed (alleles at that marker were scrambled by the exact test). A "factor <marker> <trait>;" command allows genotypes to be coded as an quantitative variable (value 1...Ngenotypes).

20−June−2002 (0.99.9)

Can force calculation of Kruskal−Wallis test for RxC table using "kru <quantitative trait> <factor>".

19−June−2002 (0.99.9)

Can now print out values for selected pedigree members tested for by a conditional expression. The "tab" command extended to tables of arbitrary dimension.

12−June−2002 (0.99.9)

MCMC based "exact" test for LD added. Visscher and Hopper test repaired −− the squared trait sums were not centred (obviously, only affected unstandardised data). Peter Visscher pointed out and helped fix this.

28−May−2002 (0.99.9)

Added "numtyp", "alltyp" and "anytyp" automatic variables that report how many markers an individual is typed at. Cleaned up syntax for "des", "reg" to allow ranges of loci.

17−May−2002 (0.99.9)

Added Visscher & Hopper's version of Haseman−Elston ("vis"). The "gener" command can now write the generation number to a quantitative variable. Check and repair troublesome locus names, eg duplicates or reserved words.

17−Apr−2002 (0.99.9)

If "0" was used as a missing value for a binary trait, this was recognised as such: except for algebra. Now automatically recoded when read in. Thanks to Jacki Wicks for pointing this out. The "read linkage" command did not have this problem.

05−Apr−2002 (0.99.9)

Locus names as heading when write pedigree to screen. Documented and improved symmetry (skewness) test.

13−Mar−2002 (0.99.9)

A divide−by−zero error and internal read error (on some compilers) fixed. Thanks to Alexa Sorant for those reports.

12−Mar−2002 (0.99.9)

One last parser problem evaluating conditional expressions where the switch statement contains a missing value. When a variable is missing and is part of an expression other than equal or not equal ("==" and "^="), the expression evaluates to missing. The "if" statement was incorrectly carrying out the "then" branch rather than returning an error.

06−Mar−2002 (0.99.9)

Added F statistics to "assoc" command (assumes the binary trait indicates membership of a subpopulation). Cleaned up output of "assoc " command when no eligible individuals genotyped.

19−Feb−2002 (0.99.9)

Fixed bug is "istyp" and "untyp" command: these did not evaluate correctly when starting values for the genotypes were not imputed (ie "set imp −1" was used).

18−Feb−2002 (0.99.9)

Stopped segfault in RC−TDT when trait missing for any siblings. Genotypic TDT simulation not done for completely missing data.

15−Feb−2002 (0.99.9)

GH and linkage locus file map distances now written to 2 and 4 decimal places respectively.

31−Jan−2002 (0.99.9)

Sham and Purcell Haseman−Elston now has slope scaled and intercept fixed so that slope is estimated QTL genetic variance (as proportion of total). Also added keywords to "sib" to allow specification of the population trait mean, variance and sib correlation, so that selected samples can be analysed. Thanks to Anastasia Iliadou for prompting these changes.

23−Jan−2002 (0.99.9)

Write out control file for Loki's *prep* (pedigree preparation) program.

18−Jan−2002 (0.99.9)

Allowed estimation of dominance variance component in "var" command. Added QTLs to "sml" command.

11−Jan−2002 (0.99.9)

Added variance components analysis including QTL linkage analysis for full sibs.

3−Jan−2002 (0.99.9)

Fixed bug for RC−TDT under Windows , thanks to Lyle Palmer (program attempted to print contents of an unassigned variable).

29−Nov−2001 (0.99.9)

Added MC P−value for global RC−TDT.

27−Nov−2001 (0.99.9)

Fixed parent−of−origin TDT where parents and child all same genotype −− thanks to Emiko Noguchi for pointing this out. RC−TDT "allowed" to work when no unaffected children in dataset.

26−Nov−2001 (0.99.9)

Implemented the Reconstructed parents−Combined Transmission Disequilibrium Test (RC−TDT) of Knapp [1999], available through the *assoc* command, where it replaces the old sibship permutation test. This is essentially identical to the default test for binary data provided by the *FBAT* program of Xu, Horvath and Laird [2001].

16−Nov−2001 (0.99.9)

Uninformative sibships (all children same genotype) now stopped from diluting the sibship permutation test. Calculated genotypes are ordered by allele size.

15−Nov−2001 (0.99.9)

Reallow mother to precede father in list of parents of an individual. Pedigree errors now cause that family to be deleted (and an error message generated), rather than halting the program. Users should check a new line of the summary output eg:

```
Number of pedigree errors  =     1
Number of deleted records  =     4
```

02−Nov−2001 (0.99.9)

Several versions of how to handle missing data in expressions, especially where these recode data. At present, can test equality or inequality of a variable with missing (test missingness), but other operations evaluate to missing. Therefore, the expression "not male" is *true* when sex is missing, but "sex<2" is *missing*. Arithmetic expressions resulting in −9999.0 (the internal missing value code) now give −9999.0001.

11−Sep−2001 (0.99.9)

Fixed bug in evaluation of "<=", ">=" "==" "^=": precedence in complex expressions was not always correct. Added CPG chi−square to *schaid* output and loglinear modelling to the service subroutines (uses IRLS). Zeroing of genotypes when error dropping on changed so always deletes all genotypes for that nuclear family rather than attempting to guess where the error is (deleting child that gave rise to inconsistency).

04−Sep−2001 (0.99.9)

Fixed minor bug in qtdt() when zero subjects in a group: zero−trapped log routine use extended.

21−Aug−2001 (0.99.9)

Added "<=", ">=" "==" "^=" as synonyms for the logical comparisons.

17−Aug−2001 (0.99.9)

Better headings for "sib" output when *plevel* equal to 1. Matching on locus names now explicitly only on first 10 letters (an annoying feature was that declaring a locus name longer than 10 characters led to the name being truncated, but a "keep" command would use the full string). Result from "tab" fixed for case where only one level of a binary trait is present (eg affected only). Results of logical equality or inequality with missing now give true or false rather than missing.

14−Aug−2001 (0.99.9)

Added "count" and "linkage" commands. The latter follows Elston and Keats' approach (as seen in Sibpal!) to sib pair linkage of codominant markers, and is *interesting* (and likely to be replaced with something else). Fixed bug affecting "ne" keyword (code implementing had been deleted somehow!).

18−Jul−2001 (0.99.9)

Fixed bug in code implementing new Sham and Purcell Haseman−Elston.

17−Jul−2001 (0.99.9)

Made "else if" work. NB: you may not nest if statements, as in:

```
if b eq 1 then (if c eq 1 then d=1 else d=2) else d=3
```

The legal equivalent is:

```
if (b eq 1 and c eq 1) then d=1 else if (b eq 1 and c ne 1) then d=2 else
d=3
```

or

```
if b ne 1 then d=3 else if c eq 1 then d=1 else d=2.
```

14−Jul−2001 (0.99.9)

Quoting blanks in some algebraic expressions no longer crashes the program eg "+" "".

10−Jul−2001 (0.99.9)

Nicer output from "tab", including results for a single variable and correct printing of genotypes.

6−Jul−2001 (0.99.9)

Added extra front ends to regress() so can access predicted values (for imputation) and multiple regression residuals: "predict" "residuals" and "impute". Fixed segfault occurring when "nuclear" met families of size 1.

5−Jul−2001 (0.99.9)

The output from "mix" now includes the Filliben correlation as a test for normality. The command "his" is a synonym for "mix 1".

28−Jun−2001 (0.99.9)

Adds "untyp()" and "round()" functions. Expressions involving genotypes evaluated for both alleles where appropriate, for example:
If genotype missing, replace with new genotype allowing for different
binning
if untyp D1S124 then D1S124=D1S124_2−1

20−Jun−2001 (0.99.9)

Fixed bug in famcor() where families of size 1 contributed to the count of matings. Quoting refined −− a quotation mark starts a new word, regardless of location. Added attributable risk to *schaid* procedure.

20−Jun−2001 (0.99.9)

Implemented improved combined Haseman−Elston regression of Sham & Purcell [2001]. Added quoting so special characters can be part of a variable name in mathematical and logical expressions.

19−Jun−2001 (0.99.9)

Rejinked selection criteria for TDT probands (1/2 x 1/2 −> 1/2 trios now contribute again) so that genotypic TDT mimics CPG GRR test (same expectations, uses unconditional chi−square as test statistic, but simulation reproduces conditional likelihood). Reintroduced "set wei imp" which counts the imputed and observed founder genotypes −− this is less accurate in small datasets (eg made up of nuclear families) than the unweighted or weighted versions.

18−Jun−2001 (0.99.9)

Fixed minor bug in writing MENDEL files: if zero had been used as a missing value for a binary trait locus in the original pedigree file, but the "read linkage" command *not* used, then this would be replaced by the last locus for the previous person when writing the MENDEL pedigree file. Added the Schaid and Sommer genotypic risk ratio test (HWE but not CPG, as latter overlaps the TDT).

14−Jun−2001 (0.99.9)

Fixed minor bug parsing negative arguments to commands (eg "set ple −1" was read as "set ple − 1"). Precedence of exponentiation lowered so "int(4.2)^2" gives 16 and not 17! And the size of evaluable expressions increased.

8−Jun−2001 (0.99.9)

Adds "inht" (inverse hyperbolic tan a.k.a. Fisher−Z transformation).

6−Jun−2001 (0.99.9)

The operator "<" was acting as ">" − corrected. Distributed DOS, Windows32/NT and Linux binaries are now compressed using UPX.

24−May−2001 (0.99.9)

Added more functionality to parser so allows (one level of) if−then−else construct, and creation and calculation of new variables. This extended to the "select" statement to allow arbitrary selection criteria to be specified. The "write" command with no arguments now writes to the screen. Quantitative variables can now be written up to 20 columns wide.

26−Apr−2001 (0.99.9)

If plevel is set to −1, Mendelian errors cause a list of possibly involved genotypes to be printed, rather than a pedigree drawing.

12−Mar−2001 (0.99.9)

Added "write dot" to produce drawing files for the *dot* graph drawing package (this does nice marriage node pedigree drawings). The "davie" command now adds the overall sibling recurrence risk. Default output from "apm", "asp", "ass", "hwe", "sib" and "tdt" is now a summary table with one line of output per test. The output print level must be increased by one to get the old output (slightly rearranged). Genehunter type pedigree files can be produced by "wri gh *file* [dummy]".

03−Oct−2000 (0.99.9)

The option "write arl *file* par" writes haplotypes from two genotyped parents per pedigree for Arlequin.

26−Sep−2000 (0.99.9)

Added "write arl" to produce haplotype files for Arlequin. Haplotypes from one child with genotyped parents per pedigree are output. The "hwe" table now includes genotype frequencies as well as counts.

20−Sep−2000 (0.99.9)

For PAP, families containing one member are now correctly recorded in *trip.dat* (as a parent of a dummy child, along with a a dummy spouse). The "grr" command calculates recurrence risks for a single major locus model parameterized in terms of prevalence and penetrance ratios.

26−Jul−2000 (0.99.9)

Fixed stupid error in "sib" command where half−sibs with one missing trait value were sometimes included in the analysis with trait value "−9999". Did not affect full−sib H−E regression. The "ibs" command writes out mean IBS sharing for all pairs, as "ibd" writes mean IBDs.

14−Jul−2000 (0.99.9)

Cleaned up bug in writing *popln.dat*, which expects only 5 allele frequencies per line. Now prints coefficient of fraternity when "kin pairwise" is used. Command "dis" estimates two−locus haplotype frequencies using independent or nearly independent informative matings.

23−Jun−2000 (0.99.9)

Fixed the obligatory equivalence problem, this time affecting "xta". Changed name to "tab". Added "sib−pair" as a locus file type.

15−Jun−2000 (0.99.9)

Added "all" as possible person ID for "edit" and "delete" command, allowing entire pedigree to be zeroed at a particular marker or phenotype.

13−Jun−2000 (0.99.9)

Corrected the P−values for the "he1 " command. This was giving the lower tail probability, rather than the upper tail probability.

08−Jun−2000 (0.99.9)

Pedigree and locus files produced for Genehunter 2 now have the loci automatically sorted. Added "xta" for RxC contingency tables for traits.

28−Feb−2000 (0.99.9)

Fixed stupid error introduced when fixing last one: caused segmentation fault in qtdt().

25−Feb−2000 (0.99.9)

Fixed stupid error in the quantitative trait "TDT" (conditional on parental genotypes allelic ANOVA) −− the founder genotypes and phenotypes were not transferred to the work arrays. Linkage locus file now always has N−1 thetas (occasionally would write N, where N is the number of loci).

15−Feb−2000 (0.99.9)

Another equivalence problem fixed, in assoc(), was overwriting allele counts for table (ANOVA results were OK).

14−Feb−2000 (0.99.9)

Minor tweaks to code for allele frequencies. Where all subjects in the dataset are untyped at a marker, they are given a starting genotype of "1/1".

09−Feb−2000 (0.99.9)

Minor changes: if print level is set to 2, *tdt* also prints out the transmitted and nontransmitted alleles for each informative proband. Use of the *he1* command gets the old squared−difference Haseman−Elston regression. A quantitative trait "TDT" added to the output from *assoc*.

20−Jan−2000 (0.99.9)

Replaced marginal genotypic TDT with simpler one conditioning on parental genotypes. Simulated P−value is now based on typed ancestors of probands (at least both parents). A memory error affecting the IBD matrix for the entire pedigree (*wribd*) is repaired.

08−Dec−1999 (0.99.8)

Minor bug fixes: greatly increased speed of MC generation of starting genotypes by replacing an inefficient loop.

26−Nov−1999 (0.99.7)

Minor bug fixes − was printing 2*N for the number of genotypes in the pedigree file.

12−Nov−1999 (0.99.7)

Added "gh2" option to *write linkage*: this adds a dummy trait locus. and writes missing quantitative traits as "−". When added to *write locus linkage*, it adds the dummy locus and writes the inter−locus distances as centimorgans rather than as recombination fractions. If *set tdt first* is used, only one proband (both parents typed) in every family is used.

29−Sep−1999 (0.99.7)

Added sibship permutation test for binary trait allelic association analysis. Fixed small bug in ibd estimation for nuclear families.

16−Sep−1999 (0.99.7)

Multilocus IBS sharing calculated for all relative pairs by the *share* command. This is an unweighted statistic. The expectation and sampling variance are generated by gene−dropping, allowing for the (sex−averaged) linkage map. The results of this may be difficult to interpret −− for example, I surmise they will detect marry−ins from a different ethnic background.

03−Sep−1999 (0.99.7)

Output from Mendelian error checking enhanced slightly −− lists the odd−allele−out in phenoset of an untyped parent causing an inconsistency.

26−Aug−1999 (0.99.6)

Fixed bugs in domix() −− reading wrong data column, and dohist() −− histogram had spurious extra bars.

23−Aug−1999 (0.99.5)

Finally altered sib pair ibd estimation algorithm to use information from all (full−sib) offspring when one or both parents untyped. Half−sib algorithm unchanged. Need to similarly swap over *twopoint* commands. Fixed bugs in nuclear() −− writing last sibship in each pedigree twice.

12−Aug−1999 (0.99.4)

Minor bugfixes in writing pedigree files (no whitespace between quantitative traits. Added "grandparents" option to *nuclear* command, so can add if phase information wanted.

05−Jul−1999 (0.99.3)

Adds experimental haplotyping routine (recmin based) and procedure to pinpoint the "lowest common ancestor" of multiple affecteds in a pedigree. The mean marital IBS sharing is now tested versus expectation in the *hwe* command, and the homozygosity test can also now be applied easily to all typed individuals. Mixtures of normals etc can be fitted to quantitative traits, as can multiple regression analysis. Various Gconvert pedigree and locus file writing routines moved to Sib−pair.

27−Jan−1999 (0.98.9)

Further small bugs removed (mainly printing IDs). The *asp* command output includes mean IBD sharing for full−sibs, along with the "mean" test P−value (exact binomial two−tailed).

25−Jan−1999 (0.98.8)

Removes two bugs introduced in 0.98.7: one lost every first member of a pedigree (save the first pedigree).

21−Jan−1999 (0.98.7)

This version contains a number of minor improvements, mainly in output from the genotyping error checking routines. Dummy records (and if necessary ID) are now generated for missing parents of nonfounders (that is, where only one parent is specified, or a parent ID is specified but a record for that person was not included).

This was previously performed by the auxiliary awk program *addpar.awk*. Individual IDs can now be alphanumeric. Errors occurring when *MAXSIZ*, the maximum pedigree size, was increased above 800 have been fixed (these arose from an overflow in the sort key).

28−Aug−1998 (0.98.3)

Fixed bug in algorithm for producing generation numbers: this was adding extra generations when loops were encountered (the pedigree was still correctly sorted in that parents always preceded children. Further tinkering with the MCMC algorithm. This is still not fully correct, as it needs a correct specification of the proposal distribution for the new fsimped() algorithm.

14−Aug−1998 (0.98.2)

Further refinement of MCMC algorithm −− much better handling of multiple marriages by simply replacing the algorithm of genof4() with that used by genof3() −− the latter is used to generate starting genotypes.

21−Jul−1998 (0.97.9)

Implemented "include" command, so commands can be read from an external file. Added ability to select on pedigree size to "select", and to trim nuclear pedigrees to a set size.

2−Jul−1998 (0.97.7)

In Gconvert, fixed problem with "wri loc tcl", which was writing cM, not M. There is an undocumented (save here!) "write locus mim", which writes a header in the appropriate format which can be concatenated onto a Linkage format file. The "keep", "drop" and "undelete" commands now accept ranges eg "drop D1S1 to D1S10".

26−Jun−1998 (0.97.6)

Cleaned up the gener() algorithm so that it correctly assigns generation number when a nominal single pedigree is in fact made up of multiple disjoint pedigrees. The "gener" command now prints out each such subpedigree in turn (if present), and the output is now more compact and easier to read.

16−Jun−1998 (0.97.5)

This adds routines that print out the mean ibd sharing at a marker locus for all pairs of relatives in a pedigree ("ibd"), as well as ("kin") the expected sharing given the degree of relationship (coefficient of relationship), or the inbreeding coefficient. There is also a utility for calculating recurrence risks under SML models ("sml"), and a command ("select") for selecting pedigrees for later analysis based on a trait value of one of its members (very basic). The "write pedigree" command now writes quantitative traits as F9.4, and the martingale residuals output by the "kaplan" command have been changed to those of Therneau et al [1990].

25−May−1998 (0.97.3)

Mainly bug repairs. Both Gconvert and Sib−pair were incorrectly recoding sexes after the "nuclear" procedure (due to change of sex from logical to integer in most but not this routine!). A backslash "\" as the last word on a line now means the next line is a continuation line. The connect() routine in Gconvert now lists all families contained within a disjoint pedigree (ie unrelated individuals with that pedigree ID are present in the pedigree

file). Now all printed thetas are positive.

01−Apr−1998 (0.97.2)

Minor bug repairs. Labelling of paternal and maternal genotypes in output from wrgtp() −− list of nuclear family genotypes when inconsistency detected −− was reversed. Stacking of jobs made easier by retaining work and data paths after "clear" issued. Gconvert upgraded to include functions such as "clear", "set data".

13−Mar−1998 (0.97.0)

This version has further changes made to the MCMC algorithm for simulating missing genotypes. The proposals made using fsimped() were not in fact being tested via the Metropolis criterion, as they were being stored in array set() and not array set2(). Again, it seems these changes make little difference to the estimates of *ibd* in the test pedigrees, although they are noticeable in the joint missing genotype distributions.

03−Mar−1998

Changes the residuals output by the "kap" procedure to the Nelson−Aalen "martingale residuals" described by Commenge from "survivor residuals" based on the product−limit estimator.

02−Mar−1998 (0.96.5)

This adds association analysis for a quantitative trait, implemented as a permutation test ANOVA. A new command "kap" gives the Kaplan−Meier estimate of the survivor function for a binary trait with variable age of onset. The residuals from this analysis can be saved for Haseman−Elston linkage and association analysis.

The Monte−Carlo Markov Chain algorithm for simulating missing genotypes has been further tinkered with, following problems encountered in multigeneration pedigrees where multiple consecutive generations are untyped. I have reintroduced a "switch of origins" operation for heterozygote children of untyped x untyped matings, as the existing "mutation" operation will shuffle such parental genotypes only when the child has no offspring (a fact that had eluded me until now). This seems to fix the trouble. Reanalysis of the example datasets included with Sib−pair found little difference in *ibd* based statistics compared to the old algorithms. Using the Lange−Goradia algorithm for producing starting genotypes for MCMC can still fail, as described in the original 1987 paper, when loops are present. If this occurs, Sib−pair now switches to its alternative gene−dropping based algorithm.

11−Feb−1998 (0.96.0)

The internal (and outputted) sorting of pedigree members is now by (founder/nonfounder) (generation) (father ID) (mother ID) (ID). This fixes a problem in some pedigrees (thanks to Hank Juo) where starting genotypes for the MCMC algorithms could not be generated via the Lange−Goradia approach. It did not affect the other MC−based approach used when imputation was "off". Since the sort key is currently integer*4, the family size in this new program is limited to approximately 1290.

A "set burn−in" option has been added to allow specification of the number of iterations of the MCMC algorithms to be run prior to the iterations used to calculate statistics. Previously, this was set to zero. In the empirical tests I had done, the Lange−Goradia algorithm generated starting genotypes seemed to give unbiased results, and so I had removed the burn−in. More recently, I have found example families (with missing genotypes) where estimated *ibd* is biased upwards. This bias is reduced by a suitable number of burn−in iterations.

28−Jan−1998 (0.95.6)

Generate workfile names using time as seed, to avoid clashes on multitasking systems. The "sta" command added. Further work on allpair: "all <tra> wpc" calculates an experimental variant of the randomization weighted−pairwise−statistic of Commenges (see Genet Epidemiol 1997;14:971−4). Further prettification of Gconvert output.

23−Jan−1998 (0.95.5)

Some prettification of Gconvert output.

13−Jan−1998

Added a "set map" command to Gconvert so that the ASPEX and LINKAGE locus files can have a sex−averaged linkage map specified. Maps can also be specified by "set dist" to give intermarker map distances, and by adding a map position at the end of the "set loc" line for a marker.

12−Jan−1998 (0.95.4)

Added ASPEX locus file and GDA pedigree file output to Gconvert. ASPEX reads command files in TCL containing the marker names and intermarker recombination distances (set to 0.50 Morgans by Gconvert). Analysis is set to be of a Linkage format pedigree file containing one binary trait and all the available marker loci. GDA reads an extension of the Nexus format. Gconvert will only write out founders unless told otherwise, as GDA is designed for population genetic work. The likelihood ratios for the IBD based ASP analysis are now corrected back (!) to be twice their former values. The HWE chi−square has been changed to a LR chi−square, to make it more robust in sparse tables. Individual IDs are now written without leading zeroes in most cases, so person 112−00000001 is now written 112−1. A "time" command will give the time elapsed since the program started, or since "time" was last called.

11−Dec−1997

The error in the loop accumulating grandparent−grandchild pairs was present for the parent−offspring pair test as well.

10−Dec−1997 (0.95.3)

Family correlations and recurrence risks now include all pairs regardless of id numbering (the old version would miss grandparent−grandchild pairs where the grandchild ID number was higher than that of the grandparent AND the grandparent was a nonfounder). The "fre[quency]" command now has a synonym "des[criptive]", and can be applied to a single quantitative trait. Previously, the correlations, sibship variance test etc were only available via a global "fre" command. The "und[elete]" command now has a default of returning all deleted loci back to the analysis.

02−Dec−1997

Reformatted apm output (list of affecteds and unaffecteds now after family summary statistics when plevel=1). Ascertainment corrected segregation ratios and standard errors default to those for complete ascertainment (reducing to those of Li & Mantel) if no variable corresponding to proband status is defined (same as "davie trait trait").

26−Nov−1997

Now trims long locus names (>10 characters) when parsing "drop", "keep" and "undelete" commands to correctly match. Does not check for names different at greater than 10th character.

25−Nov−1997 (0.95.2)

Repaired a bug in the "trans" command (boxcox()). If genotypes preceded a quantitative phenotype in a pedigree record, the wrong column would be transformed.

19−Nov−1997 (0.95.1)

Repaired annoying bug in Haseman−Elston regression analysis. Function for calculating ibd−sharing was treating the starting genotypes used for MCMC as if these were observed (therefore does not affect versions prior to 0.93). Never a problem for Gconvert's wrsib command.

Elapsed time now measured by time(), rather than secnds() for DOS version. A refinement only of interest for overnight runs.

Program now warns if there are fewer fields for a pedigree record than expected −− previously these were silently padded out as missing. It still silently ignores extra fields.

Memory utilisation decreased by equivalencing several work arrays.

13−Nov−1997 (0.95.0)

Allowable whitespace in pedigree and control files now includes tabs. Shell commands are echoed as a comment (self documenting calls to other programs).

06−Nov−1997

Fixed bug where multiple drop/keep/undelete statements clashed. Removed punctuation stripping from parser "();". These were originally present so that GAS type control files could be used as a template without editing.

05−Nov−1997

Gconvert now writes an additional line to the dummy description in an outputted LINKAGE style locus file (the variance multiplier), and sets quantitative trait zero values to 0.0001, so they are not treated as missing by LINKAGE.

24−Oct−1997

Summary of MC P−values for APM is now done using an inverse Z transform of the P−value for each pedigree. fval() now reads "y" and "n" as 2.0 and 1.0, making transformation/recoding more transparent.

25−Sep−1997

Repaired bug which made ibs affected half−sib pair chi−square incorrect for sparse tables (few pairs), by changing from Pearson to LR chi−square. Improved gener() algorithm so marry−in generation numbers

correct in the presence of loops. DOS executable now a compressed version (using DJP), reducing the size of the file by half.

21−Aug−1997

Repaired bug which meant that the "generation" command only worked if it was the first command after "run". Updated on−line help.

20−Aug−1997

Implemented routine to compare marker homozygosity in probands to that expected based on the sample allele frequencies.

11−Aug−1997

Added keywords "mat" and "pat" to perform TDT analyses only on the contributions of the mother or father of the proband.

06−Aug−1997

Updated Gconvert by adding help, setting all keywords to 3 letter minimum.

03−Aug−1997

Implemented exact two−tailed probabilities for single−allele TDT.

16−Jul−1997

Added routine for correcting segregation ratios for ascertainment.

15−Jul−1997

Added routine for testing HWE at all markers.

28−Jun−1997

Added "set tdt bot[h parents]|one [parent]" limiting TDT if requested to cases where both parents typed. Using cases where one parent is untyped can lead to bias, esp for diallelic markers with unequal allele frequencies. Alpha version of allpairs() routine set working.

26−May−1997

Added IBS based ASP analysis to Sib−pair, and an IBS based test for checking if full−sib pairs are not half−sib pairs (or unrelated!) to Gconvert.

06−Mar−1997

Prettified output of describe().

03−Mar−1997 (0.94)

Sib−pair writing out all simulated genotypes regardless of imputation level to pedigree file. Fixed so now only writes all genotypes if imputation level 3. Added "read linkage" to read Linkage−style pedigree files without having to recode zeroes to missing for quantitative traits. DJGPP V2 now used to produce DOS executable.

05−Feb−1997

Fixed up problem with imputation level 3 in Gconvert. This problem long since fixed in Sib−pair. Involved failure to correctly update genotype array in sequential imputation if genotype became fixed as side effect of another target.

Feb−1997

Trialled djgpp v2. Code seems to be slightly slower than v1 equivalent for some examples, but faster in others. Prepared source code for release.

18−Dec−1996 (0.93)

Released Sib−pair with successful MCMC simulation of missing genotypes. Added to the freq command to gives familial correlations and a version of the Fain sibship variance test for a quantitative trait. Included generation command to lists pedigree members by sibship and generation number. Included transform command to transform a quantitative trait.

18−Oct−1996 (0.92)

Cleaned up several bugs involving recoding/downcoding of alleles [only met if multiple recode statements involving the same marker], calculation of P−values for very large values of chi−square statistics [evaluated incorrectly as 1.0 or NaN in some cases], and error checking [would delete pedigree files with errors in the parent field].

Sep−1996

Versions of Sib−pair for ASP analysis following Faraway (1992). GPM ibs algorithms weight improved as Patrick Ward's program allowed calculation of exact result for comparison.

Jul−1996

Changed MC/randomization routines to sequential approach of Besag.

Mar−1996

Various unreleased versions using different MCMC algorithms.

Aug−1995

Moved Lange−Goradia algorithm from using random−access file to entirely in memory. Required moving from MS−Fortran 4.1 to f2c/djgpp for DOS version.

May−1995

Added estimation of sib−pair IBD sharing where one or two untyped parents.

Apr−1995

First code for multiallelic TDT tests. Imputation done only within nuclear families. Input and output code based on earlier PHI program. GAS pedigree file structure chosen as more readable than LINKAGE style (essentially identical).