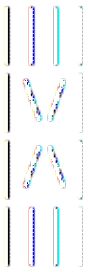


SIB-PAIR
A Program for Simple
Genetic Analysis
Version 1.0.0

USER MANUAL

David L Duffy MBBS PhD

SIB-PAIR manual

Table of Contents

SIB-PAIR 1.00b (10 May 2020)

by

David L. Duffy

(1995-2020)

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.
- 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 "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 to the standard output (the screen if you are using it interactively). Output can be saved to a file by diverting it to a file using the "out" command (in which case you will no longer see it coming to the screen). Alternatively, you may use another program that can copy from the standard output, such as *tee*, *emacs* (the powerful editor usually found on Unix systems), *syn* (a powerful Windows editor) 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 1.00.beta (12-Sep-2008)
|/\| Author : David L Duffy (c) 1995-2008
|||| Job run : Tue Sep 16 12:25:34 2008 (moonboom)

Type "help" for help, "quit" to quit, "ctrl-C" to interrupt.


>>
```

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 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
>> undrop

a* b* (m1) (m2)

The "drop" commands drops loci from the scope of subsequent commands, while the "undrop" 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:

>> undrop
>> 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 sim

```
-----------------------------------------------------
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 MERLIN type pedigree file, so we can further examine our data using a multipoint program:

```
>> b=log(b+1)
>> keep b -- m2
>> write locus merlin merlin.dat
>> write merlin merlin.ped
>> write map merlin merlin.map
```

If we like, Merlin can be run from within Sib-pair:

```
>> $ merlin --vc
```

The "$" command shells out to run another program. When we exit from Merlin, 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
 Tet Corr         0.007         0.003         0.003         0.002
 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
```

Finally, most commands can be interrupted by typing "ctrl-C" (that is holding the control and C keys down simultaneously). This returns you to the Sib-pair prompt, after finishing the current command as quickly as possible. To interrupt the program completely, you need to strike "ctrl-C" six times.

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, or the MLE of the founder allele frequencies calculated by MCMC (see below), or the BLUE of the founder allele frequencies via the approach of McPeek et al [2004].

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 diallelic markers, the exact test of Hardy-Weinberg equilibrium (assuming all individuals is unrelated) is also calculated.

Binary trait correlation. The ANOVA type estimate of the within-pedigree intraclass correlation is calculated, along with Tarone's score test [1979] for extra-binomial variation (testing whether $r>0$).

Quantitative and categorical trait correlation. The pairwise (all-possible-pairs) type estimates of the intraclass and interclass correlations are calculated for MZ twins, parent-offspring, full-sib, half-sib, grandparent, avuncular, cousin and marital pairs. In the case of categorical traits, this correlation is the (unweighted) Cohen kappa. The polychoric correlation (a faster two-stage estimator) is available for ordinal traits. Standard errors are estimated using a random subset delete-*d* jackknife method [Shao and Tu 1995]. The draw size is $max(1, min(10, Nobs/10))$, and *iter* pseudosamples are evaluated. The standard error of the test statistic is:

$$JSE=((n-d)/dm * \mathrm{Sum}(r_i-mean(r_i, i=1, m))^2)^{1/2}$$

where d is the number of observations dropped for each pseudosample, r_i is the test statistic value for the ith pseudosample, m is the number of pseudosamples, and n is the number of observations [Shao and Tu 1995].

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 routine is not currently available in the Fortran 95 version of Sib-pair. 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 for a quantitative trait:

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-jth 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) or BOBYQA (bound optimization by quadratic approximation). 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.

Standard errors for the fixed effects coefficients are the usual GLS estimates,

$(X' S^{-1} X)^{-1}$.

For binary traits, one can fit the above likelihood, which often gives quite adequate parameter estimates but dubious hypothesis testing, or move to the probit-normal mixed model, using the Multifactorial Threshold Model type formulation.

The log-likelihood to be maximized is then:

$LL = log[cdf_{MVN}(u,S)]$

where u is now the threshold value, and the probability is integrated from u to $+Inf$ when y=1, and to $-Inf$ when y=0. Integration of the multivariate normal is carried out using either the code provided by Alan Genz [Genz 1992] or the Mendell-Elston approximation code in TOMS717 [Bunch et al 1993]. The former uses quasi-Monte Carlo integration, and is limited to less than 500 dimensions (individuals per pedigree). Because this is a Monte Carlo algorithm, the maximizer can have problems settling on an exact solution. The Mendell-Elston approach is deterministic and faster, but less accurate (biased).

Categorical 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 different trait values (RxC 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. The gene-dropping can also be further conditioned on identity by descent at another marker (which may in fact be the marker being tested for association). This therefore gives a "model-free" test of association conditional on linkage.

This can be compared to results from the quasi-likelihood score methods (WQLS and MQLS) of Bourgain et al [2003], and Thornton and McPeek [2007]. These use the kinship matrix to adjust for the resulting score for

familial correlations in marker genotype (and trait phenotype by using this as the predictor). These methods are implemented as per those papers for analysis of binary traits:

$T = (p_a - p_e)\ (SD_{diff})^{-1}$

$p = \ (w_i g_i) / \ (w_i)$

g_i is half the allele count for the *i*th individual.

For the WQLS,

$w_a = \quad \boldsymbol{A}^{-1} a$
$w_e = \quad \boldsymbol{A}^{-1} \boldsymbol{1}$
$Var_{diff} = \ 0.5p(1-p)\ (\boldsymbol{aA^{-1}a})\ (\boldsymbol{aA^{-1}1})^{-2} - (\boldsymbol{1A^{-1}1})^{-1}$
where $\boldsymbol{A}$ is the numerator relationship matrix, $\boldsymbol{a}$ the phenotype vector, and $\boldsymbol{1}$ a unit vector.

It is relatively straightforward to extend the WQLS approach to multiple categories and to quantitative traits. In the former case, the score contains terms for the multiple indicators for different levels of the trait variable.

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 (FBAT) 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. For binary traits, the sibship permutation test in still available in the form of a conditional logistic regression stratifying on sibship. Finally, it is also possible to construct permutation based tests by using the **permute** or **get** commands.

In the case of multiple generation pedigrees, Sib-pair scans the pedigree upwards nuclear family by nuclear family, saving imputed genotypes, so that these contribute as *children* to subsequent nuclear families. I don't think this will be biased, but have yet to check. This behaviour can be turned off.

Population genetic F statistics (F_{IS}, F_{IT} and F_{ST}) are also calculated for each marker assuming individuals with different indicator trait values come from separate related demes. These are estimated following the approach of Pons and Chaouche [1995], as described by Excoffier [2001]. Results for each locus are combined to give multilocus summaries for the sample as:

$F^*_{IS} = (\ \ H_S - \ \ H_O)/ \ \ H_S$
$F^*_{IT} = (\ \ H_T - \ \ H_O)/ \ \ H_T$
$F^*_{ST} = (\ \ H_T - \ \ H_S)/ \ \ H_T$
Estimation of empirical (marker identity by state based) kinship uses an EM algorithm described by Choi et al [2009].

Multidimensional scaling (MDS) analysis of interindividual genetic distances is performed on the mean IBS sharing matrix for all pairs of individuals. The algorithm is Classical MDS, and so uses EISPACK routines [Smith et al 1976] to extract the first *D* eigenvectors from the doubly-centred distance matrix (1-mean(IBS)).

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.

Sequence Kernel Association Test (SKAT). This is the original formulation of the SKAT [Wu et al 2011]. The default test uses the unweighted linear kernel, but weighting to favour uncommon alleles, either using the default *Beta*(1,25) allele frequency or the inverse variance weighting of Madsen and Browning [2009] are available. The saddlepoint approximation of Kuonen [1999] for the quadratic form in normal variables is used to generate P-values. The tail probabilities from this routine are also accessible directly from Scheme (`pchisqsum`).

Two-locus and N-locus linkage disequilibrium estimation. One 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. This is the default for markers where there are more than 6 alleles per marker. Both *D* and *D'* measures are calculated.

For two locus analyses, the second algorithm combines phased and unphased genotype data in an EM fitted log-linear model. This can handle X-linked loci. If the markers are diallelic, then the model is solved directly as a cubic equation (note that for speed, the resulting chi-square is the Pearson chi-square, based on the r^2). For more than two markers, and for testing haplotype association to a trait, phase information is discarded (ie all individuals are treated as unrelated), and only autosomal markers can be used.

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.

The runs-of-homozygosity estimate of the inbreeding coefficient can also be calculated. The sum of the lengths (bp) of all runs exceeding a minimum threshold (default 1.5 Mbp) is divided by the total genome map length.

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, Weibull and Extreme Value Distribution regressions are implemented as Poisson regressions (with log time as offset) as per Aitken and Clayton [1980]. Conditional logistic regression calls the original `logccs` subroutine of Breslow and coworkers [Smith et al 1981].

Bivariate survival analysis. This is a ranks-based approach, using Kendall's tau as the measure of association. Both the original estimator [Oakes 1982] and the "normalized" estimator suggested by Oakes [2008] to get around high censoring are used. The within pair correlations in age at onset are calulated for twin and sib pairs. A bootstrap is used to estimate the standard error for tau, and the zygosity labels for the pairs are permuted to give a test of the MZ:DZ difference in correlation. This currently assume pairs are independent.

Jonckheere-Terpstra trend test. This is currently for a marker versus a quantitative trait, where the genotypes are assumed to be naturally ordered eg diallelic loci. An empiric P-value can either be calculated by permutation of the quantitative phenotype, or by gene dropping.

Stratified proportional odds test of Whitehead and Whitehead. This is a test [Whitehead and Whitehead 1991] in the spirit of the Cochran-Armitage-Berry approach, and comes in fixed and random effects flavours, the latter following Der Simonian and Laird (1987). It can be applied to binary and ordinal data.

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.

Deterministic Estimation of Pedigree Likelihoods

Rather than the original Elston-Stewart algorithm [Elston and Stewart 1971] augmented by pedigree traversal, Sib-pair uses the iterative peeling approach described by Wang et al [1996] (following Janss et al [1992]). This has the advantage of "automatically" dealing with loops, although the resulting likelihood in that case is only approximate: this approximation can be improved by several methods, but these are not currently implemented. The algorithm follows the detailed description in Chapter 2.1 of Schelling [2004], though at present only allows evaluation at one or two codominant loci.

Briefly, the iterative approach peels up and down simultaneously by calculating *anterior* and *posterior values* for each individual, where the anterior and posterior values represent the scaled likelihood contributions for ancestors and siblings, and mates and descendants respectively. The values for any individual rely on those for the other relatives, so these are reciprocally updated over multiple iterations until they converge. Usually a maximum of 10-20 iterations suffices, and if an ideal ordering of evaluations is used, a single iteration in unlooped pedigrees is sufficient ie it is equivalent to the usual pedigree traversal algorithms eg Lange and Boehnke [1983]; Wang et al [1996], Fernandez and Fernando [2001].

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.

Multipoint ibd estimation. This is a simplification of the Cardon-Fulker approach as extended by Almasy and Blangero [1998]. The *ibd* estimates from sets of markers in complete linkage are combined by calculating the variance-weighted mean *ibd* for each pair of relatives over the set. The estimates of the *ibd* variances for each marker arise as a byproduct of the MCMC algorithm.

Gene-dropping conditional on ibd. The previous algorithms can be combined by gene dropping a marker conditional on "*ibd* indicator" genotypes that represent segregation at either that same locus, or a neighbouring more informative marker. This allows us to simulate the null distribution assuming linkage but no association.

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.

In the case of the logistic-normal GLMM, the heritability is estimated as:

$$h^2 = V_A/(V_A+pi^2/3).$$

The realizations of the parameters are summarized as means, and approximate standard errors produced by batching (default B=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 BUGS Examples manual.

	Parameter Estimate (SE)				
Method	Sib-pair	AGQ	PQL1	PQL2	BUGS
SD of Plate Effect	0.28 (0.17)	0.24 (0.09)	0.23	0.24	0.29 (0.15)
Intercept	-0.51 (0.13)	-0.54 (0.17)	-0.54 (0.17)	-0.54 (0.16)	-0.51
Seed	0.06 (0.32)	0.10 (0.28)	0.09 (0.27)	-0.09 (0.28)	
Root	1.31 (0.26)	1.33 (0.24)	1.32 (0.23)	1.32 (0.24)	
Seed x Root	-0.79 (0.44)	-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 (1.23)	1.70 (1.05)	1.98	3.27	1.66
Intercept	3.70 (0.76)	2.86 (0.48)	2.74 (0.38)	2.75 (0.48)	2.81 (0.48)
Low dose	-1.46 (0.74)	-1.36 (0.82)	-1.25 (0.64)	-1.25 (0.82)	-1.35 (0.82)
High dose	0.60 (0.74)	0.58 (0.85)	0.49 (0.67)	0.49 (0.85)	0.58 (0.85)
Week>2	-1.66 (0.51)	-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 (0.101)	0.252	0.197	0.101
Intercept	1.834 (0.112)	1.833 (0.074)	1.870 (0.106)	1.870 (0.074)
Progabide	-0.346 (0.152)	-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 (0.330)	0.481 (0.244)	0.534 (0.346)	0.533 (0.244)
Period 4	-0.159 (0.054)	-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 (0.037)	0.170 (-)	0.161	0.125	0.102 (-)
Intercept	-0.151 (0.058)	-0.139 (0.043)	-0.129 (0.049)	-0.129 (0.043)	-0.138 (0.017)
UVB insolation	-0.035 (0.011)	-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.

Multiple genotype imputation. This is one approach to dealing with association in families where trait data is available for some individuals without marker genotype data. The genotype sampler is used to simulate missing marker genotypes conditional only on observed marker genotypes in the pedigree and the population allele frequencies for that marker (which can be fixed to a specified value). It is not, unfortunately, simulated conditional on values at a given trait, so it assumes data is missing-completely-at-random (MCAR). Multiple cycles of imputation followed by association analysis of the augmented data are carried out, then averaged results are produced. The estimate of the mean effect (with standard error) for an allele is produced following Rubin [1987].

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. For cases where none of the simulated test statistics exceeds the observed statistic, I provide an estimate of the P-value based on extreme value theory following Davis & Resnick [1984]. If the observed statistic is sufficiently large, it is in the tail of the distribution function which will tend to one of the classical extreme value distributions, the Pareto if the *tail index* is finite. The tail index a can be estimated following Hill [1975] as:

$$a = \mathrm{m}^{-1} \quad (log(X_{(i)})-log(X_{(m+1)}),$$

where $X_{(i)}$ is the m'th order statistic. The estimate of the tail is then:

$$P = (m/n)\ (x/X_{(m+1)})^{1/a}.$$

This estimator seems to be conservative, but not as conservative as the previous estimate $1/(n+1)$.

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].

Shortest paths. A shortest path through a pedigree between two individuals is generated using Djikstra's algorithm. The presence of multiple paths is not flagged, and the path is not purely genetic relationships.

Pedigree loops. This is a depth-first search with backtracking. Currently prints only one loop per pedigree.

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 directly allow complex programming. It does offer simple macro facilities including a loop construct. The *eval* command accesses a Scheme interpreter that can carry out more sophisticated computing.

There are command line options that may save a few keystrokes. Running "**`sib-pair -i test.in`**" starts Sib-pair and "includes" the contents of *test.in*. The command "**`sib-pair -l test.in`**" starts Sib-pair and runs the "locus" command on the contents of *test.in*. The command "**`sib-pair -h`**" (or **`--help`**) starts Sib-pair and runs the help command. If a file with suffix ".in" ".bin", or ".bin.gz" is the first argument, then it will automatically be *included*, while the ".bin", or ".bin.gz" is taken to indicate the file is a Sib-pair binary format pedigree file, and Sib-pair will attempt to read this into memory using the "read binary" command. This behaviour can be used to associate Sib-pair with that file type in a graphical directory browser, such as the Microsoft Windows Explorer.

If the program has been compiled with *g95*, the **`--g95`** flag prints useful information about environmental variables that can be tweaked to alter performance, and run time error messages.

Global commands

1. **!|#**. The rest of the line is a comment, and is echoed to standard output.
2. **echo**. Prints the rest of the line to standard output.
3. **$** <*operating system command*>. The rest of the line (up to position 80) is a command, and is passed to the shell for execution.
4. **dir** [<*OS specific args*>]. Prints list of files in current directory.
5. **[set] pwd** [<*new directory*>]. Prints name of current directory, or changes directory to that specified.
6. **file rename** <*name*> [<*new_name*>]. Renames a file.
7. **file delete|query|cat|head** <*file1*> [...<*fileN*>]. Deletes, tests for existence, prints contents or first 10 lines of a file or files.
8. **file print** [/<*search string*>/] [(<*fmt*>)] [**+**] [**NR**] [<*col1*> ...<*colN*>] <*filename*>. Prints or searches contents of a text file. Can select subset of fields to print from each record, and can use a Fortran format statement to format the output. A **+** allows one to print lines below a line matching the search string. Search recognizes wild cards "*" and ".".
9. **file transpose** <*file1*> [...<*fileN*>]. Transposes rows and columns forming contents of a file or files. If a row is short, then the corresponding column will have missing value tokens making up the total.
10. **file inverse** <*file1*> [...<*fileN*>] [<*ridge_constant*>] Inverts a numerical matrix read in sparse form (each line of file give row and column indices, followed by the element value.
11. **file vcf** <*filename*> ([<*start_position*> [<*end_position*>]]) | ([<*locus_name*> [... <*locus_nameN*>]]). Prints locus names and positions within a VCF files. Can select subset of loci to print by physical position or locus name.
12. **file vcf order|liftover** <*VCF_infile*> <*VCF_outfile*>. Subsets and orders records of a VCF file to match the order of the active locus map. Writes a new VCF file. Sorting a VCF file thus requires three steps: read the loci into Sib-pair, use the **order** command, then **file vcf order**. The **liftover** modifier leads to the VCF locus genomic position being altered to that on the Sib-pair map.
13. **file tbi** <*filename*> [(<*position*> | <*locus_name*>)] [ann] . Prints summary of tabix index file (list of indexed sequences and range of map positions) or prints the record from the indexed VCF file that matches those coordinates (directly specified or current map position for a given locus). If the **annotation** modifier is added, the values of the INFO variables for that record are printed one line per variable-value pair.
14. **file fasta** <*filename*> ([<*start_position*> [<*span*>]]) | ([<*index*>. Prints summary for each sequence in the FASTA file, a specified subsequence, or produces an index file of the (samtools) faidx format.
15. **clear**. Restarts the program, closing all workfiles and zeroing all arrays.
16. **help** [**All|Examples|Globals|Operators|Data| Analysis|**<*search_string*>]. Prints a brief description of the commands -- either all, a subset, or all matching the search string.
17. **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.
18. **list|ls|which** [<loc1> [[**to**] <locN>]] [**$(a|q|m|h|x|A|D)[m|r]**]. List of loci in current analysis. The "$<letter>" keyword denotes the list of of one class of locus, which can be reordered according to genetic map position (**m**) or in reverse (**r**) to the dataset ordering. The **which** command gives the position/rank of each locus argument in the total list of loci. If the *plevel* is set below -1, an unadorned list of locus names is printed (suitable for directly being read by other programs).
19. **list|ls|which where** <search_string>...|(**position** <pos1> [--] [<pos2>] ...) List of loci in current analysis subsetted by annotation matching a given search string or for a range of map positions. Positions take the form <chr>:<coordinate> or chrXX <coordinate>. If a chromosome is not specified for a given position, then the chromosome for the previous position is used.
20. **show chromosomes**. List of active marker loci listed by chromosome.
21. **show loci**. List of active loci along with table of numbers available (as per **info**).
22. **show pedigrees**. Same as **gener**, with print level 0.
23. **show ids** [<search-string1>]. List of indidivual IDs tabulated versus pedigree(s), or list of IDs meeting a search criterion.
24. **show map** (**position** <pos1> [**--**] [<pos2>]...)| [<loc1>...]|(where (chromosome

<chr>...)|<search_string>...). Shows the current marker map, or a subset of the map selected by map position or range of map positions, or by locus name or annotation string. Positions are given <chr>:<pos>, where position is in the current map units.

25. **show macros**. Shows the current macro definitions.
26. **show missing**. Same as **show loci** but with number of missing values per locus.
27. **show sex**. Table of counts of each sex
28. **show spectra**. Table of nucleotide allele sprectra of active loci.
29. **time**. Print time elapsed since start of the program.
30. **set timer** [**on**|**off**]. Show time taken by each command.
31. **include** <*command_file*>. Read in Sib-pair commands from a file. If the *command file* is not specified, a directory browser will start up.
32. **output** [<*output_file*>]. Divert Sib-pair output from standard output (the screen) to a file. Issuing the *output* command a second time closes the output file. If the *command file* is not specified on the first call, a directory browser will start up.
33. **macro** <*macro_name*> [(**=** <*macro value*>) | (**<- all**|**fre** <*marker*> | **bur** | **che** | **epo** | **imp** | **ite** | **las** | **lik** | **ls** | **min** | **ple** | **pri**| **pva** | **pwd**| **sex** | **twi**)]. Create a macro variable or function. If an equals sign is present, the rest of the line is taken as the macro variable body. If the <- (setting) sign is present, the macro variable takes the current value of the named program setting (or list of alleles or markers).

    Otherwise, multiple lines of the body of a macro function are expected to follow on subsequent lines, terminating with an empty line or ";;;;". The function body contains Sib-pair commands and parameters for substitution, represented as *%1*, *%2* etc. The macro is called either as a function or a variable.

    To call a macro as a function, the entire macro name must be the first word on the command line, followed by any required parameters (which can be any legal string). The function body is evaluated, and the resulting Sib-pair command(s) is then acted upon. For example:

    ```
    # Draw a pedigree (need to have dot and gv):
    # First set up macro
    >> macro draw
    draw> select pedigree %1
    draw> write dot %%.dot %2
    draw> $ dot -Tps -o %%.ps %%.dot
    draw> $ gv --scale=-4 %%.ps
    draw> $ rm %%.dot %%.ps
    draw> unselect
    draw>
    # Call macro
    >> draw ped1 trait
    ```

    The *%%* keyword is replaced by the macro call ID, a unique 5 character string. The *%0* keyword is replaced by all the macro parameters, while *%+2*, *%+3*... is replaced by that parameter and all subsequent parameters.

    To call the macro as a variable, the entire macro name is prepended by the "%" character. It can then appear anywhere in a command. The macro variable is replaced by the macro body, but there is no evaluation of macro function parameters or %%. The variable name can be delimited by brackets, so that it can be embedded in another string. The resulting command is then parsed:

    ```
    >> macro a=1
    >> macro b=+
    >> macro b1=2
    >> %a%b%a
    ```

```
=> 2.
>> 1%b1
=> 12.
>> 1%(b)1
=> 2.
>> macro a=D14S52 D14S43
>> tab %a
------------------------------------------------
Cross-tabulation of "D14S52" ... "D14S43"
------------------------------------------------
[...]
```

34. **eval** [<*Scheme expression*>]. Accesses the Scheme read-eval-print-loop. If called without arguments, presents the Scheme prompt "%%", otherwise evaluates the rest of the line as if it were a Scheme expression. Sib-pair macro variables and functions are stored within the Scheme environment, and so can read and written to. The Sib-pair specific extensions to Scheme are "`(nloci)`", "`(ls)`", "`(loc)`", "`(loctyp)`", "`(locord)`", "`(locnotes)`", "`(locnotes-set!)`", "`(read-line)`", "`(run)`" and "`(help)`".

(ls ["m" | "x" | "h" | "a" | "q" | "d[mxhqa]"])	List locus names
(nloci)	returns total number of loci
(loc <index>)	returns locus at that position in the locus list
(locord <loc>)	returns position of a locus in the locus list
(locnotes <loc>)	returns notes for a locus
(locnotes-set! <loc> <string>)	rewrites notes for a locus
(loctyp <loc>)	evaluates type of a locus ("adhmqx")
(read-line <port>)	Read next line from port as a string
(run "<Sib-pair command>")	Run a Sib-pair command
(string-split <string> [<sep>]	Convert from string to list of words split on white-space (or a specified character separator)
(substring? <sub> <str>)	returns start of substring in string.
(system "<operating system command>")	Run an OS command
(help)	Information about this Scheme implementation

See the appendix for more details.

35. **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".
36. **quit**|**bye**. Halts the program.
37. **set prompt** [**on**|**off**]. Displays a prompt, and activates/resets the command line history.
38. **set gui** [**on**|**off**]. Activates or stops the GUI. The GUI is either Java based (using the japi library to interface with the Java AWT library), or GTK2.0 based (using the pilib library to interface), depending on the compile time choice.
39. **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*.
40. **set epoch** [**iso**|**jul**|<*epoch*>]. Set or show the epoch used for julian dates. Defaults to "iso" epoch of 1970-01-01.
41. **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.

42. **set printstyle** [**rectangular**|**pairwise**]. Sets the format used by the *print* command. The default type is rectangular, so that the selected records are printed on one line with values at each variable in regular columns. If the *pairs* style has been set, then each value is printed as a *variable_name*=*value* pair.
43. **set categorical_trait levels**|**labels**. Sets the output representation of values of a categorical trait to either the level (1..N) or the label (as seen when first read in, or from the annotation.
44. **set table_separator** <*separator_character*>. Sets the character used to separate columns in summary tables of test results. Defaults to " " (a space). This affects the output from the *summary*, *hwe*, *frequency*, *assoc* commands.
45. **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.
46. **set analysis** [**imputed | observed**]. Includes imputed genotypes in generalized linear models fitted using the *regress* command. Genotypes are automatically reimputed each time *regress* is called. The idea of this is to allow multiple imputation association analysis. Imputation is performed using Sib-pair's single locus genotype MCMC sampler. The *regress* command with imputed genotypes respects the *set frequencies* command, so that the population allele frequencies can be specified in advance.
47. **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).
48. **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.
49. **set starting_trials** <*number of iterations*>. Controls the number of Monte-Carlo trials used to generate a starting genotype configuration for the MCMC algorithms. Defaults to 5000. Sib-pair uses either this gene dropping approach or a sequential imputation approach using Lange-Goradia elimination to produce a legal set for ungenotyped individuals in a pedigree.
50. **set emit** <*number of iterations*>. Controls the number of (Monte-Carlo) Expectation Maximization iterations used by the *mcf* algorithm.
51. **set batch** <*number of batches*>. Controls the number of batches used by the *fpm* algorithm for the estimation of parameter standard errors.
52. **set chain** <*number of MCMC chains*>. Controls the number of chains used by the *fpm* algorithm.
53. **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.
54. **set jackknife** <*number of jackknife draws*>|**off**. Controls the number of deleted observations used by the *describe* command's algorithm for the estimation of standard errors for familial correlations for a quantitative trait. If **off** modifier chosen, reverts to default behaviour.
55. **set mft** (**mendell-elston**|**genz** <*number of function evaluations*>. [<*absolute error*> [<*relative error*>]]). Controls the multivariate normal CDF algorithm (TOMS717 Mendell-Elston or Genz). The number of function evaluations and absolute or relative tolerances used for integration by the Genz *mft* algorithm can also be specified.
56. **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.
57. **set order_statistics** <*number of order statistic*>. Controls the number of Monte Carlo simulated statistics used to extrapolate P-values for even more extreme observed statistic values using the approach of Davis & Resnick [1984]. Defaults to highest 10.
58. **set chi-square** [**gibbs**|**pearson**]. Set whether the Pearson or Gibbs chi-square is used for categorical trait association. Defaults to Gibbs.
59. **set seeds** <*seed1*> <*seed2*>< *seed3*>. Initializes random number generator seeds to given values,

rather than via system time.

60. **set optimizer varmet|bobyqa**. Choose BOBYQA or VARMET optimizer for variance components model fitter.
61. **set fbatimpute on|off**. Enables (default) or disables imputation of missing child alleles based on their own offspring.
62. **set vcf** [<*Alternate_allele_count_variable*> <*Total_allele_count_variable*>]. Defines names of the VCF INFO variables from which reference counts of alternate and total alleles will be read by the "ass vcf" command. Defaults to "AC" and "AN".
63. **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.
64. **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.
65. **set roh** [<minimum_length>]. Sets the minimum length of a run of homozygosity to be used to contribute to the F_{roh} estimator of inbreeding. Defaults to 1.5 Mbp (or 1.5 cM).
66. **set model** [**allelic**|**genotypic**]. Select allelic or genotypic encoding for marker loci in regression models. Default is allelic encoding.
67. **set map function kosambi|haldane**. Set the mapping function used by multipoint analytic and locus file output routines.
68. **set map units** [**cM**|**M**|**Mbp**|**kbp**|**bp**]. Set the map units for reading scripts and writing to the screen.
69. **set sml** <*Frequency of **A** allele*> <*Penetrance of **AA** genotype*> <*Penetrance of **AB** genotype*> <*Penetrance of **BB** genotype*>. Set default single major locus model written to locus files (usually p=0.05, pen=(0.5, 0.5, 0.05)).
70. **set liability** <*binary trait*> <*liability class indicator*> <*number of classes*>. Declare a quantitative trait to indicate liability class for the named binary trait. Used to generate Linkage format locus and pedigree files.

Utilities

71. **pchisq** <*chi-square*> <*degrees of freedom*> [(**ncp** <*ncp*>) | <*df2*>]. Calculate P-value for central or noncentral (if the *ncp* keyword is present) Chi-square distribution (or F-distribution).
72. **qchisq** <*P-value*> <*degrees of freedom*> Calculate quantile corresponding to the given P-value for the central Chi-square distribution.
73. **chisq** <*nrows*> <*ncols*>. Calculate contingency chi-square and permutation P for flat table entered via keyboard.
74. **polychoric** <*nrows*> <*ncols*>. Calculate polychoric correlation and association P for flat table entered via keyboard.
75. **proportion** <*numerator*> <*denominator*> <*confidence interval width*>. Calculate accurate confidence interval following Wilson (as described by Agresti and Coull) for a proportion.
76. **tetrachoric** <*prevalence*> <*recurrence risk ratio*> Calculate tetrachoric correlation equivalent to given recurrence risk ratio and trait prevalence.
77. **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.
78. **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.
79. **grr** <*trait prevalence*> (<*Frequency of **A** allele*> <*genotypic risk ratio*> [**add**|**dom**|**rec**]) | (<*Frequency of **A** allele in cases*> <*Frequency of **A** allele in controls*> **case-control** [**population-controls**). 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). If the *case-control* modifier is present, instead it expects the prevalence, risk allele frequency in cases, and risk allele frequency in controls who can either be unaffected, or unknown status (*population*).
80. **ito** <*Frequency of **A** allele*> [<*Penetrance of **AA** genotype*> <*Penetrance of **AB** genotype*> <*Penetrance of **BB** genotype*>]. Calculates conditional genotype frequencies for a relative of a proband of given genotype (if only allele frequency provided) or affection status under the specified diallelic generalized single major locus model (if penetrances were also specified). ITO refers to the matrices used to perform the calculation for pairs of relatives [Li & Sacks 1954].

Algebraic operators and functions

81. **"**<*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.
82. (<*value*>|<*locus*>) ***|/|mod|+|-|^** (<*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.
83. (<*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.
84. **if** <*logical expression*> **then** <*action*> [**else if** <*logical expression*> **then** <*action*>]... [**else** <*action*>]. Conditional evaluation of expressions. The **if** statements can be nested.
85. **log|log10|sqrt|exp|sin|cos|tan|asin|acos|atan|abs|int|round** (<*variable*>|<*locus*>). Functions acting on numerical constants and/or trait values.
86. **rand|rnorm**. Produce a random value from U(0..1) or N(0,1).
87. **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).
88. **ishom|ishet** <*marker*>. Test if homozygous (or heterozygous) at given marker.
89. **alla|allb** <*marker*>. Return the first or second allele for each individual at the given marker.
90. **marcom**. Show the maximum of the number of markers an individual and any of his relatives are both genotyped at.
91. **numtyp|anytyp|alltyp|protyp**. Show number of markers an individual is genotyped at, or indicate whether genotyped at any one or all marker loci, or the proportion typed out of all marker loci .
92. **male|female|isfou|isnon**. Test sex and founder status of individual.
93. **num|nfound**. Number of members and number of founders of the pedigree containing an individual.
94. **famnum|index**. Position of pedigree and of individual in the *active* dataset.
95. **chosen**. Indicates individuals who were affected by or contributed to the last operation eg hash, merge, table.
96. <*expr*> **:** <*expr*> [**:**...]. The colon separates the members of a sequence of algebraic expressions. Its main function is to allow multiple expressions in the branches of an if-then-else statement, but it will minimize overheads in repetitive calculations. If multiple bracket-enclosed expressions are encountered, then a **:** is automatically inserted:

    >> (a=rnorm) (b=a^2)

Command iteration

97. **... {** (<*val1*> <*val2*> ...<*valN*>)|{<*val1*> : <*val2*>} | ($[mxqa]) **} ...**. Repeat the associated command, placing each value of the iterator list into the position the list currently occupies. There may be multiple iterators, and iterator lists may be nested. This macro extension allows much of the functionality of proper computer languages:

Iteration over a range of numerical values of a function	sml {0.1 0.3 0.5} 1 0 0
Iteration over a range of loci of a function	tab a1 m1 m2
Iteration over a range of functions	{tdt ass} a1
Combinatorial generation of strings eg locus names	set loc a{1 : 3} aff

Compound statements	if (male) then m{1 2}=x

Data Declaration commands

98. **set datadirectory** <*pathname*>. Sets directory to be searched for pedigree files.
99. **set workdirectory** <*pathname*>. Sets directory to which temporary files are written.
100. **set impute off|on|full|sequential|lange-goradia|nil|**<*level*>. Toggles imputation routine.
    - Imputation level *0* (**off**) does not impute genotypes. It does generate legal genotypes via gene-dropping for all untyped individuals, that are used to start the MCMC genotype sampler. These genotypes are "hidden" to other routines. The gene-dropping approach is slowed by, and can fail (stochastically) in large pedigrees. In this case, the imputation level can be set to *-1* or **nil** (*completely off*!).
    - Imputation level *1* (**on**) imputes single individual's genotypes if unambiguous using results of the Lange-Goradia imputation algorithm. All other missing genotypes are silently imputed via gene-dropping.
    - Imputation level *2* (**full**) carries out the same imputation as level *1*, but the imputed values for all missing genotypes are printable and saveable using **write**. If imputation has already been carried out, changing the imputation level to *2* affects just the printing of "hidden" imputed values of unobserved genotypes.
    - Imputation level *3* (**lange-goradia** or **sequential**) imputes values for all missing genotypes via sequential application of the Lange-Goradia genotype elimination algorithm. That is, the missing genotype is imputed to be the most likely genotype conditional on the typed members of the pedigree, and those genotypes imputed prior to that iteration. The imputed genotypes are *not* visible to non-MCMC commands (eg **write**) unless the imputation level is changed to *2*.
101. **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).
102. **set checking off|on**. Toggles the first level testing for Mendelian inconsistencies within nuclear families.
103. **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
haploid	Y or mitochondrial codominant marker
quantitative	quantitative (or interval or ordinal) trait
affection	binary trait
categorical	categorical 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 fifth 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**). If you wish to annotate, but do not have a map position, a "." for the fifth column is unobtrusive and accepted by Sib-pair.

    The annotation for a categorical variable can also include labels for the levels of the trait. These are of the form "level=label". If the categorical variable is string-valued in the input dataset, then Sib-pair will convert to levels and write an appropriate annotation.
104. **declare loci** <*number*>(**m**|**x**|**q**|**a**) [<*number*>(**m**|**x**|**q**|**a**)...]. Declare a batch of loci, automatically generating names: "trait1", "trait2"..."traitN", "mar1", "mar2"..."marN" as appropriate.

105. **rename** (<*locus name*> [**to**] <*new name*>)|[**bim** | **vcf**] *map_file_name* [*INFO_variable*]. Change name of previously declared loci. Either the new name is specified on the command line, or it is obtained from a map file where the appropriate replacement name is associated with a locus at the same map position as the target locus. If an INFO variable in a VCF file gives a locus name, this can be specified as an additional field, and used instead of the VCF file `ID` variable.
106. **loci** <*command file*>. Read in Sib-pair locus and pedigree file declarations from a file. If the *command file* is not specified, a directory browser will start up.
107. **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
108. **read locus merlin** <*locus file name*> [**xli**] [**snp**]. Read locus names, types from a Merlin-format locus (*.dat*) file. Markers are treated as sex-linked when the *xli* modifier is used. Markers are stored as SNPs when the *snp* modifier is used, with reduced memory consumption.
109. **read locus plink** <*locus file name*> [**append**] [**human**]. Read locus names, types and map positions from a PLINK locus (*.map*) file. The *append* modifier stops a dummy binary trait being declared as the first locus, and appends to rather than overwrites any existing locus information. The *human* modifier causes chromosomes 23...26 to be interpreted as X, Y, XY, MIT.
110. **read locus file** <*locus file name*> <*locus_name*> <*chromosome*> <*map_position*> <*reference_allele*> <*alternate_allele*> [**human**] . Read locus names, types and map positions from a rectangular locus (*.map*) file where the first line contains names for each column. The column names corresponding to the locus name, chromosome, map position (bp), and referance and alternate alleles are specified as arguments. These can be numeric if names are not provided.
111. **read locus vcf** <*VCF file name*> [<*start*> [<*end*>]] [**human**]. Read locus names, types from a VCF. Reading can be retricted to a subset of loci within an interval (specified in base pairs, prefixed optionally by a chromosome identifier). The *human* modifier recognizes chromosomes 23-26 as sex and mitochrondrial chromosomes.
112. **set append** [**skip**|**version**]. Controls how repeat declarations of a locus are treated. Default behaviour is to modify the name of the repeated locus ("*name*_v2", "*name*_v3"...), but this can be changed so that such declarations can be skipped.
113. **read pedigree** <*pedigree file name*>|**inline** [**skip** <*lines_at_beginning*>] [nopedid] [nosex]. 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 ";;;;". The **skip** keyword leads to the skipping past that number of lines at the beginning of the file. The **nopedid** and **nosex** modifiers allow the pedigree file to be missing a pedigree ID column (in which case all records are taken to belong to a single kindred), or a sex column (where sex defaults to missing for all records, and is usually imputed in a mating-consistent fashion)
114. **read linkage** <*pedigree file name*>|**inline**. Reads a pre-MAKEPED LINKAGE type pedigree file.
115. **read ppd** <*pedigree file name*>|**inline**. Reads a post-MAKEPED LINKAGE type pedigree file.
116. **read cases** <*data file name*>|**inline** [**noid** [**sex**|**nosex**] [**sep** <field_separator>] [**skip** <*lines_at_beginning*>]. Reads a data file containing unrelated individuals. The individual ID is the first column of data, which is followed by all the phenotypes. If the **noid** keyword is present, then an ID field is not expected, and row number will be used as ID. If the **sex** keyword is present then the second column of data is expected to be the sex. The **skip** keyword leads to the skipping past that number of lines at the beginning of the file.
117. **read csv** <*data file name*> [**noheader**] [<field_separator>]. Reads a CSV format file containing data for unrelated individuals. The default is to expect a header row giving the column names. The type of each column ("a", "q", "m", "c") is inferred based on their contents, and used to declare the variables to Sib-pair, and to read the dataset.
118. **read vcf** <*VCF file name*> [*...VCF file N*] [**ped_id**] [<*lines_length*>]. Reads only the IDs from a VCF file header and sets up the corresponding pedigree in memory. If the **ped_id** modifier is present, the ID is split to give a pedigree and individual ID (not applicable if multiple VCF file names are give). Neither the marker names nor genotype data are read from the VCF file.
119. **read annotation** <*VCF file name*> [<INFO_var1>[:<field>]...<INFO_varN>[:<field>] [**sep**] [**dump**]. Reads INFO variable values from a VCF file. If a variable name is not specified, lists the descriptions

of all the variables from the header. Otherwise, finds the matching locus in the current dataset (by name then by position if the former fails), and prepends the value(s) to the Sib-pair locus annotation (`locnotes`), or dumps these to output. If an INFO variable is an array (with values separated by "|"), the field name or index can be specified.

120. **show annotation** <*VCF file name*> [<INFO_var1>[:<field>]...<INFO_varN>[:<field>] Prints cross-tabulations of INFO variable values from a VCF file. If a variable name is not specified, lists the descriptions.
121. **read bin** <*data file name*>. Reads a Sib-pair "binary" pedigree file. The "run" statement is bypassed, so this command reads in locus and pedigree data immediately. No checking is done for pedigree and Mendelian errors and reordering of pedigree members is not performed. These can be large files, but on systems where gzip is available, can be automatically compressed (see *write bin*) and decompressed. The default format for the file arises from a Fortran unformatted write of the locus and pedigree arrays, and so will be compiler and platform specific.
122. **read hapmap** <*data file name*>. Reads a HapMap style genotype data file. All individuals are assumed unrelated.
123. **read plink** <*file prefix*> [**compress**]. Reads a PLINK `.bed` and ancillary files: <*prefix*>.*bim*, <*prefix*>.*fam* and <*prefix*>.*bed*. The "run" statement is bypassed, so this command reads in locus and pedigree data immediately. No checking is done for pedigree and Mendelian errors or reordering of pedigree members -- this is assumed to have been performed by the program that prepared the files. If the **compress** modifier is present, then the genotype data is stored in a 4 bits per genotype format internally.
124. **set sex marker** <*locus name*>. Declares a sex-informative marker, such as Amelogenin.
125. **set sex** <*report_threshold*> [<*hom_to_het_rate*>]. Sets reporting threshold and assumed X-linked genotype heterozygote to homozygote miscall rate for diagnosis of sex from X-marker information.
126. **set twin** <*quantitative trait*> [**merlin**]. Declares a variable to identify twin zygosity, or specifically just monozygotic (MZ) twins. All individuals within a pedigree with the same value of this variable are taken to be part of the same twin sibship (twin pair or higher order multiple). Different values indicate different pairs within the same family. If the **merlin** coding has been specified, odd numbers indicate MZ twins, and even numbers DZ twins. For the default coding, nonzero values indicate MZ twins, and DZ twins can be indicated using a zero value for the indicator). This information is used to write out MZ twin indicators to the pedigree files used by MENDEL, MERLIN and SOLAR, and to test for marker inconsistencies.
127. **set twin error** <*threshold*> <*minimum markers*>.

    The **error** modifier shows and allows setting of the quick genotype discordance test (**test dup**). The first argument is the allowed mistyping (ie discordance) rate for genotypes compared between putative genetically identical pairs of samples. This defaults to 0.005, which is acceptable for SNP or sequence markers. The second argument determines how many markers must be present before the comparison is made, defaulting to 100 markers.
128. **set skiplines** <*slines*>. Skip first *slines* lines in pedigree file) when reading in.
129. **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).
130. **set map** <pos1>...<posN>. Set map positions for the marker loci. This will overwrite any original map positions.
131. **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.
132. **set chromosome** <chr1> [...<chrN>]. Assign each marker locus to a specific chromosome. If there are more markers than specified chromosomes, the last specified chromosome is reused.
133. **read map** <map file name> [**[k]bp**]. 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 GVF, Merlin, Mendel, Solar. Tests number of columns and whether column contents are numeric or

alphabetical, skipping first row as possible header, or looks for identifying strings in the header. Default units are cM or Mbp. The *bp* (*kbp*) modifier tells Sib-pair to divide the positions by 10^6 (10^3); thus the map distances become Mbp, and remain readable.

134. **read chain** <chain file name>. Read in the specified UCSC chain file and update the map positions of currently active loci using that information.
135. **set frequencies** <marker> [[<allele_name1>] <allele_freq1>... [<allele_nameN>] <allele_freqN>)]. Sets the "population" allele frequencies for a marker to be used by MCMC procedures that simulate missing genotypes for that marker. This currently only affects the **set analysis** and **gpe** commands. Only one marker at a time can have specified frequencies. The **obs** option can fix the frequency to that observed for the currently active dataset. To free the frequencies (allow calculation from the dataset), call the command without specifying any frequencies after the marker name.
136. **run**. Reads in pedigree file and creates working pedigree file. Imputes genotypes if requested.

Data manipulation commands

137. **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.
138. **undrop** [<*loc1*>...[<*locB*> **to** <*locC*>]... [**$(m|x|q|a)**] ...<*locN*>] Return previously dropped loci to analysis. Default is to undrop all dropped loci. Loci can be selected on name (including wild cards), class. This is not the reverse of the *delete* command.
139. **keep|drop|undrop where** (**monomorphic**
    | diallelic
    | snp
    | allele_number [<c_op>] <*numal*>
    | spectrum <*allele1-allele2-...alleleN*>
    | max <*frequency*> [<c_op>] <allele_frequency>
    | number_typed [<c_op>] <*ntyp*>
    | missing [<c_op>] <*ntyp*>
    | chromosome <*chr1*> [...<*chrN*>] [**trait**]
    | distance <*smallest_gap*>
    | r2 <*smallest_r2*>
    | every *number_skipped*
    | position <*pos1*> [--] [<*end_position*>;] ...
    | near <*loc*> [...<*locN*>] [<*N*>]
    | hwe_p [<c_op>] [<*critical_P*>]
    | homoz [<c_op>] [<*threshold*>]
    | test_p [<c_op>] [<*critical_P*>]
    | in [<file>]
    | covers <*trait*> [<*n_uncovered_cats*>] | <*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 or diallelic or a specified number of alleles, that the commonest marker allele frequency exceeds or falls below a threshold, the number (or proportion) of individuals typed or missing exceeds or falls below a threshold, the marker is closer than a set amount to the last included marker (map distance), marker is in too high linkage disequilibrium (r2) with the last included marker, every Nth marker in list, on a given chromosome or chromosomes (may be a trait eg expression level), the N nearest loci to a target locus, within an interval or intervals on the genetic map, the HWE test or most recently applied test P-value is smaller (or larger) than a critical level (defaults to 0.05/Nmarkers), the marker homozygosity, the number of categories of a specified trait that are completely ungenotyped (defaults to zero), locus name is listed in a text file, or the marker annotation matches the search string.
140. **select** [**containing|exactly** <*nprobands*>] [**where**] <*a logical expression*>. Select pedigrees containing one or more individuals with a trait value meeting the criterion.
141. **select pedigree|id** [[**not**] **in**] (<*ped1*>...<*pedN*>)|**file** <file>. Select pedigrees included or excluded from a list of pedigree or individual names, which can be in a file. 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).
142. **unselect** [<Nth>]. Returns all pedigrees excluded by a select command back to the analysis. If an integer argument is given, this gives how many **select** statement subsettings to roll back. The argument can be negative (reversing the effect of a previous targetted unselect).
143. **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.
144. **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> *".

145. **copy** <*pedigree1*> <*person1*> <*pedigree2*> <*person2*> [**insert | merge**] Copies all *active* trait values and genotypes from first record to second record. If the **merge** modifier is present, copying only occurs where the existing value of the target is missing at the variable.
146. **update|merge** [<*file_type*>] <*file_name*> :

merge

[**genotypes**] [**illumina|locus_first**] [**qua** <c_op> <thr>] [**skip** *lines_to_skip*] |

[**probabilities** (**key** <*mergekey*>)|**pedid|id** [**mach** | (**impute2 file** <*locus_file*>)]]

[**plink|bed** <*root_prefix*>] [**join|compress**] [**id**] [**pos**]

[**vcf** <*VCF_file_name*>] [**ped_id**] | [**quality** <QCstat> [<c_op> <thr>]].

[**dose** <*PLINK_dosage_file_name*> [<*thresh*>]] |

fimpute <*locus_file_name*> <*genotype_file_name*> [**compare**]

[**mac** <*datfile*> <*pedfile*> [<*thresh*>] |

[<*locus1*> ...<*locusN*>]
<*phenotype_file_name*> [*lines_to_skip*].

Updates phenotype/genotype data in the current dataset using values read from a file. The default file format requires the first line of the file to give the names of the variables that are included in the subsequent lines. Usually, the first column is the *pedigree_ID* and is named *ped*[*igree*]; the second column is the *individual_ID*, and is named *id*. Alternatively, the pedigree ID can be omitted, so the merge requires the individual IDs to be all unique. The remaining columns should have names that match locus names in the current dataset (data for nonmatching names are skipped).

```
ped id locus_name1 locus_name2 locus_name3 ...
1    1  A/A         12.434      y ...
1    2  A/B           x         n ...
...
```

Where the pedigree and individual IDs for a record in the update file match that of an active individual in the current dataset, the corresponding phenotype and genotype values for that individual are updated using the values read from the file. If **merge** was called, only missing values in the current dataset are updated, but all values are overwritten if the call was to **update**. By specifying locus names on the command line, you can further control the loci that are updated from the new file.

The **merge genotypes** option reads files containing one record per individual locus genotype (ie fields are individual_ID, locus name, genotype):

```
1    locus1  A/A
1    locus2  1/2
...
```

or if the **locus_first** modifier is present:

```
locus1  1 A A 0.85 ...
locus2  1 1 2 0.99 ...
...
```

or if the **illumina** modifier is present:

```
[Header]
....
[Data]
```

```
SNP Name  Sample ID GC Score  Allele1 - Forward Allele2 - Forward
locus1  1 0.85 A A
locus2  1 0.91 2 9
...
```

If a quality score is present, this can be used to filter out poor quality genotypes. This is automatically enabled for the Illumina report format, with a threshold of 0.6. This can be altered using the **quality_score** option. For other formats, the quality score is assumed to be in column 5.

The **merge plink** option reads the supplied PLINK `.bed`, `.bim`, `.fam` files and merges the new SNP genotypes to the existing pedigree data. The loci merged are those active in the current SIb-pair dataset and present in the `.bim` data under the same locus name, unless the **pos** modifier is present, when matching is by map position. If the **join** modifier is present, it also automatically reads the accompanying .bim file and declares all these loci before merging -- this is silently called by the "read plink" command. The **id** keyword sets the merge key to be only the individual ID. The **compress** option stores genotypes as 4 bits per genotype, but is less flexible than the "snp" storage mode.

The **merge probabilities** option reads SNP genotypic probability files from Beagle, Impute2, and Mach, converting to the most likely genotype. One can specify the name of a mergekey variable using the *key* modifier (giving the column number that corresponds to each individual) or the **pedid** or **id** keywords to specify that the probabilities file column headers contain ids of the form <pedigree_ID>-<individual_ID> or <individual_ID> respectively. If headers are absent, as in the case of Impute2, these can be read from a locus file, as specified by the *file* modifier.

The **merge dose** option reads SNP allelic dosages from a dosage file prepared for PLINK. A "hard" threshold is used to convert dosages to genotypes, which can be set by appending a <threshold> value to the command. The default threshold value is 0.5, giving cutpoints of 0.5 and 1.5 (the dosage scores take values 0-2).

The **merge mach** option reads SNP allelic dosages from a pedigree prepared using MaCH or minimac. The names of the pedigree file and datfile specifying the locus names need to be specified. A "hard" threshold is used to convert dosages to genotypes, which can be set by appending a <threshold> value to the command. The default threshold value is 0.5, giving cutpoints of 0.5 and 1.5 (the MaCH dosage scores take values 0-2).

The **merge vcf** option reads SNP genotypes from a VCF file. The **ped_id** modifier tells Sib-pair to match on pedigree and individual ID, where the VCF IDs are expected to take the form <*pedigreeID*>_<*individualID*>. The name of a genotype quality score present in the VCF file can be given, and optionally a comparison that will be used to filter the genotypes to be tested. The mean quality score for each locus is saved, and is accessible to the **summary** command.

The **merge fimpute** option merges in genotypes from an FImpute genotype file (both FImpute locus and genotype files need to be specified). The **compare** option compares genotypes in the specified genotype file to those in the current dataset matching on individual and locus, including enumerating the updateable genotypes from FImpute.

147. **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.
148. **delete** [<*locus1*>...<*locusN*>] **[whe|where]** <*a logical expression*>. Sets specified data to missing for all individuals meeting particular criteria.
149. **get** <*relationship*> <*statistic*> <*trait*> [<*newtrait*>]

get	**all**	**mean**	<*trait*>	[<*newtrait*>]
	offspring|children	**minimum**		
	sons	**maximum**		

daughters	**sum**
parents	**count**
father	**sample**
mother	
spouse	
husband	
wife	
siblings	
brothers	
sisters	
mztwins	

Summarizes trait values of all relatives of the specified class of all individuals, saving the result to a trait locus if requested. The **sample** option samples with replacement unless the **all** option is used, when the sampling is without replacement. Statistics for siblings include *ego* ie give sibship mean etc; sampling for siblings excludes *ego*.

150. **recode** (<*marker*>|$(m|x)) [**frequencies**|**letter**|**number**|**nucleotides**]. Recodes alleles at that marker or set of markers to 1..N, where the ordering defaults to the allele size (or collation order for letter alleles). If the **freq** modifier is present, the numbering is by ascending allele frequency. If the **let** modifier is present, the numbering is "1..4" for "ACGT", and the reverse for **num**. The **nuc** option recodes for genotypes using "A/B" allele coding to the appropriate nucleotides provided in the locus annotation, where it should be in the format "[<A1>/<A2>]".
151. **recode** (<*marker*>|$(m|x)) [**reference** <filter_trait>]. Recodes missing genotypes at that marker or set of markers to the wild type homozygote. An indicator trait can be used filter this so this restricted to a subset.
152. **recode** <*marker*> <*all1*|*value1*>...<*allN*|*valueN*> **to** <*new allele*|*new value*> [...<*newN*>]. Allows pooling and/or recoding of marker alleles prior to subsequent analysis. If there are fewer new values than old values, the last new value is recycled.
153. **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.
154. **swap** <*marker*> [**to** <*marker2*>] [...<*markerN*>]. Swap diallelic marker alleles names eg "[12]" -> "[21]".
155. **swap** <*trait*> [**lod** <*swap_lod_threshhold*>]. Tests for and repairs likely allele swaps with particular levels of a categorical trait. Calculates likelihood for table of association between marker and the target trait, and then for each table where the allele labels are swapped for a single level of the trait. If the ratio of likelihoods for each ordering expressed as a lod exceeds the threshhold - default 2 - then the allele labels will be swapped for individuals where the value of the target trait matches the most deviant level.
156. **flip** <*marker*> [**to** <*marker2*>] [...<*markerN*>]. Recode SNP marker alleles to the complementary strand coding ie "[ACGT]" -> "[TGCA]".
157. **flip** [map|fasta <*file*>] Compares observed alleles for a SNP to reference alleles, either from the locus annotation (where they will be written [A/B] with A the consensus or reference allele) or those in an external VCF, GFF file, and flips or swaps the alleles to match. It also alters the locus annotation reference alleles to match the external file. If allele frequencies are provided in the external file (eg VCF INFO AF variable), this is used to test for strand flips for ambiguous SNP ie A/T and G/C. The *flip fasta* option changes only the locus annotation. The FASTA file must be indexed, ie have a .fai file as produced by the samtools faidx program or Sib-pair, and *flip fasta* changes only the locus annotation. Results are sent to locstat, so one can keep/drop/undrop on the value of the comparison for each SNP.

158. **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).
159. **date** [(<*yyyymmdd*> **julian**)| (<*juldate*> **gregorian**)]. Convert a single date from Julian to Gregorian or Gregorian to Julian.
160. **test age** <*age_variable*> [<*threshold*>]. Test that parent and offspring ages are consistent. The *threshold* controls the minimum age of parents at birth of offspring, and defaults to zero, since the units are unknown.
161. **test dob** <*DOB_variable*> [**gregorian**] [<*threshold*>]. Test that parent and offspring DOBs are consistent. The *threshold* controls the minimum age of parents at birth of offspring, and defaults to 4380 days, if **gregorian** has been used to declare the DOB variable to represent Gregorian dates.
162. **test age** <*age_variable*> [<*threshold*>]. Test that parent and offspring ages are consistent. The *threshold* controls the minimum age of parents at birth of offspring, and defaults to zero, since the units are unknown.
163. **test sex**. Uses X-chromosome and/or Amelogenin to test the putative sexes of genotyped individuals. If Y-chromosome markers are declared, then presence of data for an individual is taken as further evidence of male sex; absence is ambiguous, as may simply be missing.
164. **test haploid** [**mitochondrial**]. Uses Y-chromosome or mitochondrial markers to test the putative relationships of genotyped individuals.
165. **test map** [(**merge** [<thresh>]) | <map_file>]. Test for multiple loci dclared to have the same map position. If the *merge* modifier has been specified, Sib-pair will merge the genotype data for the two loci if the allelic spectrum is compatible ie same alleles, **or** compatible with a strand mixup. If too many incompatibilities between genotypes at the same individual are detected (default *threshold* 0.005), then the merge will not proceed. If a file name is specified, the current map is compared to that contained in that file, which is read as if by "read map".
166. **test ids** <*file*>. Compare IDs in current dataset to those in a specified file, which may be either a simple list of pedigree and individual IDs, or a Sib-pair binary or VCF file.
167. **test vcf** <*VCF_file*> [**ped_id**] [qua <QCstat> [<c_op> <thr>]]. Compare genotypes for the matching IDs and loci in current dataset to those in the specified VCF file. Flips strand as necessary. The name of a genotype quality score present in the VCF file can be given, and optionally a comparison that will be used to filter the genotypes to be tested. The mean quality score for each locus is saved, and is accessible to the **summary** command.
168. **test strand** [<*marker 1*> [...<*marker N*>]]. Test if mix of genotypes for a locus are compatible with a strand mixup.
169. **test flips** map|fasta <*file*>|<source-indicator> [**lod**]. Compare reference alleles given in locus annotations to those in an external VCF, GFF or FASTA file. The FASTA file must be indexed, ie have a .fai file such as that produced by the samtools faidx program or Sib-pair. Results are sent to locstat, so one can keep/drop/undrop on the value of the comparison for each SNP. If the name of a stratifying variable is given, then differences in allele frequency between strata are tested to see if they are compatible with an allele swap in one stratum, and the identity of the most deviant level and associate lod printed.
170. **test duplicate** [**ids** [**merge**]] [<*indicator*>]. Test if genetically identical individuals are present in the dataset. Write an indicator value for each individual tested to a phenotypic variable if specified. If the *ids* modifier is specified, then testing is limited to persons with the same individual ID in different pedigrees. If found, and the *merge* modifier has been specified, then Sib-pair will reciprocally fill in missing data in one record using data from the other record. A duplication indicator can be written to a phenotype variable if this is specified.
171. **test locus** [<*marker 1*> [...<*marker N*>]]. Carry out the usual Sib-pair Mendelian error screen on the specified locus and only on active pedigrees. Useful if checking was turned off when the pedigree was first read in.
172. **test lange-goradia** [<*marker 1*> [...<*marker N*>]]. Carry out the usual Sib-pair Lange-Goradia elimination algorithm test for Mendelian errors on the specified locus and only on active pedigrees.

Useful if checking was turned off when the pedigree was first read in.

173. **test** <*pedigree*> <*individual* [<*ped2*> <*id2*>]. Default is to test for the most highly related individual in the entire dataset. If a second ID is given, prints the genotypes at loci where they are discordant, along with the proportion discordance - good for detection of somatic mutations or genotyping errors

Analysis commands

174. **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.

175. **normality_test** [<*trait*> [...<*traitN*>]]. Calculate Filliben correlation (correlation between value and expected quantile under assumption of normality) for each quantitative trait. The test statistics are accessible to subsequent summary, keep, and drop commands.
176. **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 will be performed using the individual's family mean and standard deviation, if the **fam** keyword is included.
177. **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. Has been superceded by the more powerful residuals command.
178. **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.
179. **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
180. **impute** <*ytrait*>. Replace missing quantitative trait values with the predicted values from a regression on the spouse, sibling and offspring observed values. Designed mainly for imputing missing age or date of birth.
181. **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.
182. **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.

183. **lifetable** (<*start*>|0) <*end*> <*censor*> [<*cohort stratum width*> [<*period stratum width*>]] [**time**] [**covariate** <*covariate*>]. Prints the life table for the quantitative trait *end*, where *censor* is the binary outcome trait, which is *affected* when *end* represents the age at which the individual first expressed the trait. Start represents the time of entry into the study of the individual. If the *start* trait is "0", the start of observation is taken to be 0 for all individuals. The <*cohort stratum width*> is the bin width used to divide up the entry times into the study. The <*period stratum width*> is the bin width used to divide up person time of followup (and defaults to years). The <*time*> modifier causes the start and end values to be treated as a continous measure, rather than a Julian date (the default). The **covariate** keyword declares the following named variable to be a categorical covariate which will be used to stratify the dataset.
184. **surv** <*age-at-onset*> <*censor*> [<*covariate1*> [..<*covariateN*>]]. Carries out the logrank nonparametric test for equality of survival curves across the different strata formed by crossclassifying one or more covariates. The variable *censor* is the binary outcome trait, which is *affected* when *age-at-onset* represents the time at which the individual first expressed the trait. If no covariate is specified, tests all active markers in turn, carrying out a gene-dropping test of significance.
185. **rank** <*trait*> <*rank*>. Write the ranks of a quantitative trait to the quantitative variable *rank*.
186. **blom** <*trait*> <*blom_score*>. Write the approximate inverse normal scores for a quantitative trait to the quantitative variable *blom_score*.
187. **quantile_normalization** <*trait1*>...<*traitN*>. Carry out quantile normalization over a set of quantitative traits, replacing each observed value with the mean value at the corresponding order statistics of all the chosen traits. Missing values are dealt with by reusing the nearest value rather than linear interpolation.
188. **simulate** <*trait*> [<*h2*> [<*linked extant marker*>]]. The data for the named trait is replaced by simulated data for a trait under the control of a QTL (or polygenes) with a total heritability of *h2* (defaulting to 50%). If a second marker name is given, the controlling QTL is simulated as being completely linked to the second marker.
189. **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.
190. **simulate qtl** <*trait*> <*marker*> [<*h2*>]. The data for the named autosomal marker is replaced by simulated genotype data that is generated conditional on the trait values and the genetic model for the trait. The model for the QTL is taken from the results of the *set sml* command, and the residual heritability is specified as the last argument to the command. Binary traits are modelled under the multifactorial threshold model, and quantitative traits are not currently supported.
191. **simulate pedigrees** [<*nped*> [<*ngen*> [<*min_number_of offspring*> [<*max_number_of_offspring*>] [<pedigreeID_prefix>]]]]. Generate a set of *nped* (default 100) random pedigrees, each of *ngen* (default 2) generations. The component nuclear families each contain between *min_number_of offspring* and *max_number_of_offpsring* offspring (defaulting to a range 0-2). The generated pedigrees are each descended from a single founder couple (with marry-ins). Multiple calls to *simulate pedigree* can be made in an incremental fashion. A string to be prefixed to the pedigree ID can be given as argument 5.
192. **permute** <*trait*> The phenotype values for the named trait are permuted within pedigrees.
193. **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.

194. **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.
195. **join** <*pedname1*> [<*pedname2*>...<*pednameN*>] Join or rejoin pedigrees into a single pedigree, appropriately dealing with shared individual IDs and their associated data. Note that Sib-pair allows noncontiguous pedigree blocks to use the same pedigree name, so multiple pedigrees of the same name will be collected up.
196. **prune** [<*binary trait*> [|<*quantitative trait*> **over**|**under** <*threshold*>]]. Reduce pedigree to contain probands and minimum number of connecting relatives.
197. **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.
198. **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...
199. **hash** [<file_of_IDs> | **file** <file_of_IDs> [<col1> [<col2>]] | <pedigree_ID> <individual_ID> | **id** <individual_ID> | **show** [**locus**]| **delete** | **size** <percent_table_size>] | **vcf** <VCF_file> . Sets up, show, deletes or utilizes a hashed index for searching out IDs. Does not allow use of wild cards cf print. If the single argument is a file name, each line in the file is expected to contain a pedigree and individual ID as the first two words, or just an individual ID as the single field. These will be searched for in the current dataset. If the **file** keyword was used, then the file name can be followed by one or column names specifying which contains the individual ID, or pedigree and individual IDs. Increasing the `plevel` gives lists of unmatched and matched IDs. If two arguments are given, these are taken to be a pedigree and individual ID to be searched for, while **id** followed by a string is taken to be an individual ID to be searched for. The **show**, **delete**, and **size** are for tuning, and will not be needed for ordinary use. The **vcf** option tabulates the overlap in IDs between the present dataset and a VCF file
200. **print** [**where**] <*a logical expression*>. Print trait values for individuals, with a combination of trait values meeting the criterion.
201. **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).
202. **write** [<*pedigree file name*>]. Writes a GAS type pedigree file from the current dataset. Default is to screen.
203. **head**|**tail** [**map**|**loci**|**kinships**] [<*nrec*>| (<*skip*> <*nrec*>)]. Writes the first or last *nrec* records (default 10) of the current dataset, loci, marker map or kinship matrix to the screen. If two arguments are present, then the first represents the number of records to skip over before writing *nrec* records.
204. **more** [<*nrec*>]. Pages through the current dataset, *nrec* records (default 20) per page. Allows paging backwards and forwards by full or half pages.
205. **write pap**. Writes the required pedigree files *trip.dat* and *phen.dat* (note that you may have to sort *trip.dat*).
206. **write bin** <*pedigree file name*> [**compress**]. Writes a Sib-pair "binary" pedigree file. The file actually contains both locus descriptions and pedigree/genotype/phenotype data, and so saves the state of the program at that time. An image of Scheme's memory and the twinship and sex marker indicators are also saved. These can be large files, but on systems where `gzip` is available, will be compressed if the **compress** modifier is present. Such files are automatically decompressed by the *read bin* command (if the filename has a `.gz` suffix. The default format for the file arises from a Fortran unformatted write of the locus and pedigree arrays, and so will be **compiler and platform specific**.
207. **write** <*format*> <*pedigree file name*> <*modifier*>

write	**pedigree**|**gas**	<*pedigree file name*>	[**header**]
	arl		<population>

asp|tcl	
beagle	[**fou|tri**]
crimap|tcl	
csv	[<delimiter> [<missing_value_token>]
dot	[[<trait>] [(<marker>|<trait>) [<colour-y> [<colour-n> [<colour-x> [<marker-background-colour>]]]]]]
fimpute	[**chip** <chip_variable>] [<filter_trait>]
fisher	
gda	[**all**]
haploview	
linkage|ppd|gh	[**dummy**] [**numbered_alleles**]
mendel	
merlin	
phe	
plink	[<trait>]
sage	
sas	
sib-pair	[**header**]
roadtrips	
snp	
snap	
solar	[**phe**] [**nopedigreeID**]
structure	[**fou**]
vcf	[**ped_id**] [<trait>]

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. If the modifier *header* is present, a line containing the names of all the variables is prepended. The keyword *dot* writes a script that will draw all the active pedigrees using the dot program as a marriage node graph (a binary trait locus and one marker locus can be represented by colour and the written genotype in the node respectively); 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 genotypes from all active individuals; the data may be separated into populations, based on the categories of a specified population membership indicator trait. 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. The *haploview* option is a linkage style file with recoding of letter alleles from "ACGT" to "1234". 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; *merlin* writes a pedigree file suitable for Merlin -- actually a LINKAGE "pre" format file with zygosity included as the first trait, if the "set twin" command has been previously issued; *fisher* writes a pedigree file from the current dataset suitable for use by the program FISHER; *cri* writes the ".gen" style file required by CRI-MAP; *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 pedigree ID column can be

dropped via the *nop* option, and the SOLAR phenotype (or genotype, depending on a prior keep statement) file written by the *phe* option. The SOLAR pedigree file has two additional columns: MZ twin indicator (requiring a previous "set twin") and a household (actually pedigree) indicator. The *sas* command writes a datastep with the pedigree data as inline "cards". The *structure*, *roa*, and *beagle* commands write genotype data files for Structure, ROADTRIPS, and Beagle respectively (and can be restricted to writing just founder data using the *founders* option). The *plink* option writes PLINK format `.bed`, `.bim` and `.fam` files. If a (trailing) trait name is given, this is written as the trait in the `.fam` file. The *snp* modifier writes a PLINK `.tped` file (row-major pedigree file), while *snap* writes a similar file for WOMBAT. A VCF file containing the genotypes of the currently active markers is written when the *vcf* modifier is present. The **ped_id** modifier tells Sib-pair to write a combination pedigree and individual ID, where the VCF IDs take the form <pedigreeID>_<individualID>. If a trait name is given, then eonly individuals where this trait is non-missing will have data written to the output file. The *fimpute* writes genotype data files for the FImpute imputation program of Sargolzaei et al. The **chip** modifier gives the trait identifying which SNP array (or genotyping method) was used for that individual. A trait name at the end of the command restricts the outputted data to those where the trait is non-missing.

208. **write map mendel|merlin|loki** <*map file name*>. Writes out the map file required by MENDEL 4.0, MERLIN or LOKI.
209. **write locus pap**. Writes the required locus files *header.dat* and *popln.dat*.
210. **write locus** **aspex|tcl** <*locus file name*>

beagle	
fisher	
fimpute	[**chip**< *chip*>]
gas	
haploview	
linkage|gh	[**dummy**] [**xlinked**]
loki	
mendel	[**trait**]
merlin	
relpair	
sage	
sib-pair	[<*pedigree file name*>]
structure	<*pedigree file name*>
superlink	[**dummy**] [**xlinked**]

Use of the keyword *gas* writes a GAS type locus file from the current dataset; *haploview* writes an "info" file, giving the marker position as the first word of the annotation if it is numerical, or the map position multiplied by 10^6; *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*, while the *xlinked* option declares the markers to be all X-linked. 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 (with binary traits defaulting to a factor, but given as a diallelic locus if the *trait* modifier is present); *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; *superlink* writes a LINKAGE style locus file modified for the SUPERGH option of SUPERLINK; *relpair* writes the RELPAIR format (modified from that for MENDEL): it infers chromosome number from the map position (as multiples of 1000) or from the locus name (if it takes the form of "DxSxxx"); *beagle* writes a Beagle marker list (marker name, position in bp, allele names). *fimpute* writes a FImpute locus file. The *chip* keyword preceded the trait indicating the SNP

array or genotyping method for each marker.

211. **write var** [**mendel**] <*var file name*>. Writes out the var file (list of quantitative traits) required by MENDEL.
212. **write grm** <*GRM file name prefix*> [<filter_trait>]. Writes out the current dataset kinship matrix in the GCTA GRM binary file format (actually 3 files, one with IDs, N used to estimate each kinship, and the actual kinships). An indicator trait can be used filter this so this restricted to a subset.
213. **read grm** <*GRM file name prefix*>. Read in kinships from a GCTA binary format GRM matrix binary file format that match the IDs of the current Sib-pair dataset. GCTA actually provides 3 files, one with IDs (pedigree, individual), N used to estimate each kinship (which Sib-pair discards), and the actual kinships. These replace the current dataset "big" kinship matrix.
214. **generations** [<*quantitative trait*> [**reverse**]]. List founders/marry-ins and sibships by generation number for all pedigrees, (over)writing the generation number to a quantitative trait if requested. If the *reverse* modifier is present, the generation number counts up from the bottom of the pedigree, rather than from the top.
215. **loops** [<*binary trait*>]. Prints marital or inbreeding loops in the active pedigrees. If a binary trait is specified, the members of the loop are flagged as "y" at that trait, with nonmembers set to missing at the trait.
216. **gpe** <*codominant marker*> [**mcmc**] [<*allele dose estimate*>]. Gives iterative peeling or MCMC (Monte-Carlo Markov Chain) estimates of the genotype probability estimates for the given marker for each individual: a vector of probabilities corresponding to the possible genotypes at that marker. For an observed genotype, this is 100% for the observed value and 0% for all other possible genotype values. For an unobserved genotype, it gives the probability distribution of possible genotypes conditional on the sample allele frequencies (assuming Hardy-Weinberg Equilibrium) and the observed genotypes in the individual's relatives. If the name of an extant quantitative trait is appended to the command, the expected gene dose for the *first* (ie lowest in the collation order) allele will be written to this variable. The *gpe* command respects the *set frequencies* command, so that the population allele frequencies can be specified in advance.
217. **peel** <*codominant marker*>. Calculate the pedigree likelihood for the given marker.
218. **haplotypes** <*marker1*> <*marker2*> <*newmarker*> [<*threshold*>]. This infers phased genotypes when the two markers are in complete (or near-complete) LD. The threshold sets the maximum number of the rare haplotype that is acceptable when LD is not complete.
219. **triads** This routine lists haplotypes inferred from fully typed parent-offspring triads, along with counts of obligate recombinants.
220. **relatives** <ped> <id>. This routine lists relatives of an index individual: parents, sibs, spouses, offpsring and descendants. If the pedigree is small (<12 members) or the *plevel* is set to *1* or higher, then a list of shortest paths from each pedigree member to the index individual is also printed.
221. **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.
222. **typed** [<*binary trait*>|<*quantitative trait*>]. Prints number of individuals phenotyped at each locus. If a stratifying variable is specified, these counts are versus each level of the *trait*.
223. **frequencies|describe** [[<*codominant marker*>| <*binary trait*>| <*quantitative trait*>]...[**to**]...<*trait*>] | [**snp|polychoric**] . Print allele frequencies for marker loci, segregation ratios for a binary trait, relative-pair pairwise agreement for a categorical trait (Cohen kappa), polychoric correlation for an ordinal 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.
224. **mcfrequency** <*codominant marker*> | **$m**. Print MCEM (Monte-Carlo Expectation-Maximization) estimates of the founder allele frequencies for marker loci. A fixed number of EM iterations are carried out, usually 20. This can be set higher if desired using the *set emit* command.
225. **count** [**where**] <*a logical expression*>. Count individuals, and sibships and pedigrees containing such individuals, with a combination of trait values meeting the criterion.
226. **print** [**where**] <*a logical expression*>. Print phenotype data for individuals with a combination of trait values meeting the criterion.

227. **hist** <*quantitative trait*> [<*number_of_bins*>]. Produce Alternative interface to **mix** with one distribution. The number of bins in the lineprinter histogram can be set.
228. **plot** <*quantitative trait*> <*quantitative trait*> [<*categorical trait*>] [<*file*>]. Produce an Encapsulated Postscript scatterplot for two quantitative traits. If a third binary or quantitative trait is specified, then this controls the plot symbol used for each point (10 different symbols are available for the trait values 1 to 10). Graphic file name defaults to "sib-pair.eps".
229. **tabulate** [**ped** <*trait*>) | ([**showmiss**] [**polychoric**] [**sampleweights** <*trait*>] (<*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. If the **showmiss** keyword is present, then an additional level indicating that the trait value is missing is included in the table, while the **polychoric** option will calculate the polychoric correlation for a two-way table. One can specify for a oneway table to be printed for each pedigree using the **pedigree** option, which is supplemented by the Tarone score test for extrabinomial variation if the trait is binary. The default behaviour is to print a one-way table for each active variable.
230. **llm** [(] <*trait 1*> [[+] <*trait 2*>...])^<N>]] [[+] <*trait 1*> * <*trait 2*>...][-1] [allelic]. Carry out log-linear modelling of a multidimensional contingency table under the specified model. The **allelic** option causes an allelic model to be fitted for genotypes of the first listed marker in the model. Formula exponentiation expands to all interactions up to the level of the exponents for the selected terms. The "lrt" command can be used to compare sequentially fitted models.
231. **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.
232. **means**|**correlations** [<*trait 1*> ...[<*trait N*>]]. Calculates means, standard deviation and correlation matrix for a list of traits.
233. **pca** [<*quantitative trait 1*> <*quantitative trait 2*> [...<*quantitative trait N*>]]. Carries out principal components analysis for the listed traits. If a single integer is given, this allows entry of a covariance matrix for that number of variables via the keyboard.
234. **regress** <*ytrait*> **=** <*x1*>...[**to**]... <*xN*> [**offset** <*offset*>] [**poisson**|(**exponential**|**weibull**|**evd** [<*censoring_trait*>)] [**shape** <*shape*>] [**sim**] [**rep** <*nreplicates*>]. Performs linear or logistic or poisson or weibull or EVD 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. The **shape** keyword declares a starting value for the solution of the Weibull distribution shape parameter. The **sim** keyword gives a gene-dropped P-value for the first marker locus in the list. The **rep** keyword specifies a number of replicates for multiple imputation of the test marker locus genotypes, and is usually used when **set analysis imputed** has already been issued.
235. **clreg** <*ytrait*> **=** <*x1*>...[**to**]... <*xN*> [**ped**|**stratum** <stratum_indicator>]. Performs conditional logistic 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 default stratification is on sibship, but the addition of the **ped** keyword gives an analysis stratified on pedigree.
236. **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.
237. **kinship** [**inbreeding** [**mc**] | **pairwise** | **dominance** | **roadtrips** | <*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, along with an indicator as to whether a pair are full sibs. When

the **dominance** keyword is present, a pairwise list is printed of bilineally-related pairs (defined as *K*>0) that are not full sibs. Alternatively, if requested, print a list of individuals with a non-zero inbreeding coefficient, using a Monte Carlo estimator if **mc** modifier is present. 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, and permutation test of significance of observed mean coefficient. For use with the ROADTRIPS programs, a pairs format where the kinship, rather than coefficient of relationship, and inbreeding coefficients are printed for self-pairs (rather than 1+F) can be printed.

238. **ibd** <*codominant marker*> i [... <*codominant marker N*>][**pairwise**]. Write the estimated mean identity-by-descent sharing at a marker or set of closely linked markers for all relative pairs in each pedigree as a lower triangular matrix or a list of pairs.
239. **ibs** [**ibd** | **moment** | <*binary trait*> | <*quantitative trait*> **>|>=|<|<=|==|^=** <*threshold*>]. Prints the estimated mean identity-by-state sharing at all active markers for all pairs in the dataset as a list of pairs, or if the *ibd* modifier is present, then the MLEs for the kinship coefficients. The *moment* modifier gives the shrinkage estimate of the kinship matrix of Endelman and Jannink [2012]. If a binary trait is specified, sharing is only calculated for pairs where both are affected. Ungenotyped individuals are skipped completely, but genotyped individuals sharing no markers with any other individual will be evaluated (so the mean IBS will be printed as missing).
240. **set kinship A|C|ridge_constant** <constant>. A kinship matrix for the entire dataset (stored internally as matrix `kinmat`) can be specified by this command, or read in from an external file. While *A* and *C* set it to the pedigree-based NRM or a block diagonal family environment correlation matrix, the *ridge_constant* option can be used to modify an existing matrix.
241. **hwe** [**founders**] [<*mar1*> ..[**to**].. <*marN*>] [**$**(**m** | **x**)]. 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.
242. **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.
243. **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.
244. **mds** <*dim1*>[... <*dimN*>]. Performs classical (metric) multidimensional scaling of interindividual genetic distances. These are based on identity by state (IBS) sharing at all active marker loci across all pairs of active individuals in the dataset. Individuals have to be typed at at least 50% of the active markers to be included in the analysis. The number of dimensions output is controlled by the number of quantitative traits named in the command: each will contain the estimated coordinates for every individual on one of the dimensions.
245. **mztwin** <*monozygosity_indicator*> |(<*zygosity_score*> **even**|**odd**|(**>|>=|<|<=|==|^=** <*threshold*>)) [**clean**|**delete**|**find**]. 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. The **find** option uses the currently active markers to identify likely MZ twin pairs as pairs of relatives (in the same pedigree) sharing 99.5% or greater marker concordance. A value indicating membership of an MZ twin pair will be written to the currently active twin indicator trait or to the named variable. The **even** and **odd** zygosity score test require the score to be positive, and are

designed for use with MERLIN's coding system.

246. **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").
247. **twin** <*trait*> [<*zygosity_score*> **even | odd** | (**>|>=|<|<=|==|^=** <*threshold*>)]. Calculate, and test for equality of, MZ and DZ twin correlations or concordances. If a zygosity indicator is not specified, the indicator declared by the "set twin" command is used to divide siblings into MZ twins and non-MZ full siblings (DZ twins and "singletons"). If a zygosity indicator is specified, then this is used to define three groups: MZ twins (indicator expression TRUE or greater than zero), DZ twins (indicator FALSE or equal to zero), and nontwin siblings (indicator value is missing). The Z test for equality of the Pearson correlation coefficients is actually the likelihood ratio test [Brandner 1933]

.

248. **kendall** <*quantitative trait*> <*censoring trait*> [<*zygosity_score*> **even | odd** | (**>|>=|<|<=|==|^=** <*threshold*>)]. Bivariate survival analysis for twins. Kendall's tau is used as the measure of association, standard errors are produced by bootstrapping, and a permutation P for the MZ:DZ difference.
249. **varcomp|mft** <*quantitative trait*> [**ae**] [**ce**] [**ace**] [**ade**] [**covariate** <*covariate trait 1*> [...+ <*covariate trait N*>]]. Performs MVN or MFT 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).
250. **lrt**. Constructs likelihood ratio test comparing the last two variance components (**var**), log-linear models (**llm**), 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).
251. **blup** (<*quantitative trait*> <*h2*>) | (<codominant marker> [<binary trait> | (<quantitative trait> [(**>|>=|<|<=|==|^=** <*threshold*>)])). Calculates BLUPs (best linear unbiased predictions of the breeding value) for a quantitative trait assuming the given heritability, or if a codominant marker is specified, the BLUEs for the marker allele frequencies, either for the entire sample, or for one subgroup.
252. **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*>...]]
[**save** <*target locus*>].
Uses a Markov Chain Monte Carlo (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. "sav" modifier keyword for "fpm" can save BLUPs or genotypes to a variable.

253. **dis** [<*marker locus 1*> [<*marker locus 2*> [.. <*marker locus N*> [<*trait*>]]]. Estimates frequencies of haplotypes based on individuals with two informative parents or no informative parents. For pairs of multiallelic markers (6 or more alleles), only the former group is used. In the informative parent-offspring triads, the loci are assumed to be tightly linked, so that four parental haplotypes may be inferred, but this maybe turned off. 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. For pairs of markers with few alleles, information from both phased and unphased genotypes is combined using an EM-fitted log-linear model, which can also deal with X-linked data. If three or more markers are specified, or two or more markers along with a trait, the genotypes are all treated as if unphased (and X-linked data is not handled). The same log-linear modelling framework is used. When a trait is specified, the haplotype frequencies within each trait category (two if a binary trait, as many levels as observed for a quantitative trait) are calculated, along with a chi-square testing equality of marker haplotype frequencies across categories. If exactly two numbers are given, then these are assumed to be the number of alleles at two loci, and genotype counts may be entered at the command line.
254. **neff** [<*Map window*> [<*critical P value*>]]. Calculates the effective number of association tests using the set of active (autosomal) markers, allowing for the observed intermarker linkage disequilibrium. Implements the approach of Moskvina & Schmidt [2008]. The optional parameters are the width of the sliding map window containing markers to be tested versus the current iterated marker (currently 1 cM/Mbp), and the genome-wide critical P-value (defaults to 0.05).
255. **assoc** (<*binary trait*> |<*quantitative trait*> [(**>**|**>=**|**<**|**<=**|**==**|**^=** <*threshold*>]) | **categorical**) [**founders**] [**vcf** <*VCF_file_name*>] [**covariate** <*covariate*>] [**genotypic**] [**ibd** <*ibd_marker*>] [**snp**|**freq**|**maf**|**risk**]. 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. If the *categorical* modifier is present, then each unique value of a quantitative trait is taken to represent a category of a multinomial trait. For dichotomous traits, a second table in the output shows the results from the reconstruction-combined TDT within informative sibships. Also prints F statistics assuming trait is indicator of membership of 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 via gene-dropping, which can either be unconditional, or completely linked to that of another marker. Analysis may be restricted to founders. Genotypic rather than allelic analyses can also be specified, using the *genotype* flag, and if the trait prevalence has been specified (by the **set prev** command), the attributable risks and sibling recurrence risk ratios are also calculated (*plevel* > 0). Covariates can be added to some analyses. The *snp* modifier limits the output to an allelic odds ratio and 95% confidence intervals (along with the P-value) for a binary trait, and allelic regression coefficient and SE for quantitative traits; the *freq*, *maf* and *risk* modifiers limit output to case and control frequencies (plus test statistic and P-values), where the frequencies are that of the base allele, minor (in control) allele, and risk allele respectively. The **vcf** option, by contrast, performs a naive chi-square test comparing case and control allele frequencies to those recorded in a VCF file as INFO variables defaulting to "AC" and "AN". These can be altered by the "set vcf" command
256. **mito**|**yhaplo** <*quantitative trait*> [<*marker 1*> [...<*marker N*>]]. Performs ANOVA analysis of the association between a quantitative trait and Y chromosome or mitochondrial haplotypes, defaulting to each eligible marker in turn if markers making up a haplotype are not explicitly listed. Haplotypes are imputed where unequivocal, and Monte-Carlo empiric P-values are produced by gene-dropping.
257. **rare** (<*trait*>|<*marker*>) [<*MAF_threshold*>]. Pools rare alleles in all currently active markers below a given threshold allele frequency. Default maximum MAF for inclusion is 0.01. If a trait is specified, lists carrier individuals by trait locus status, and provide table of number of alleles carried versus trait locus status, along with a contingency chi-square test. If a marker is specified, the genotypes for this

are replaced by dummy genotypes representing the number of rare alleles carried by the individual ("1/1"=0, "1/2"=1, "2/2"=2 or more).

258. **skat** <*trait*> [**madsen_browning**|**beta_1_25_weights**) Performs SKAT combining all currently active markers. Analysis can be unweighted, or using two different allele frequency based weights.
259. **wqls** <*trait*> [**kin** <*kinship_file_name*>] [**ridge_constant** <*ridge_constant*>]. Calculates the WQLS score test of allelic association for binary, categorical (more than 2 categories), or quantitative traits.
260. **mqls** <*binary trait*> [|<*quantitative trait*> **>**|**>=**|**<**|**<=**|**==**|**^=** <*threshold*>] [**prevalence** <*prevalence*>] [**kin** <*kinship_file_name*>] [**ridge_constant** <*ridge_constant*>]. [**hwe**|**robust**]. Calculates the MQLS, WQLS and modified-chi-square quasi-likelihood score tests of allelic association where the trait is binary, or above or below the given threshold if the trait is quantitative, or categorical (corrected chi-square only). The MQLS test requires an estimate of the trait prevalence, which can be specified by appending a value to the command following the **prevalence** keyword, or preset using the **set prev** command.

    Rather than using the given pedigree, a kinship matrix can be read from a file, containing one element per line, and fields corresponding to a header line:

    ```
    ped1 id1 ped2 id2...kin
    ```

    A ridge constant can be specified to be added to the diagonal of this matrix, and either "naive" (**hwe**) or robust (**robust**) forms of the tests used.
261. **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.
262. **multihomoz** [(<*binary trait*> [|<*quantitative trait*> **>**|**>=**|**<**|**<=**|**==**|**^=** <*threshold*>]) | **save** <*quantitative trait*>]. 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. The simulated P-value assumed linkage equilibrium so the markers must be appropriately thinned. If trait is not specified, or the *save* modifer used, the observed and expected multilocus homozygosity and the runs-of-homozygosity inbreeding coefficient estimator (F_{roh}) is written for each individual. The latter can be *save* to a quantitative trait.
263. **fstats** [<*population_indicator*> [**founders**]. Prints the F statistics comparing identity-by-state sharing of alleles within loci, within populations (demes), and between populations. Population membership of an individual is indicated by the value of a specified binary or quantitative trait. There can be more than two populations.
264. **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.
265. **hrr** <*binary trait*>|(<*quantitative trait*> **>**|**>=**|**<**|**<=**|**==**|**^=** <*threshold*>). Performs the Haplotype Relative Risk test, comparing case offspring allele frequencies to those in their parents. A gene-dropping P-value is generated, so that it is applicable to general pedigrees.
266. **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 $GRR_2=GRR_1=1$.
267. **sdt** <*binary trait*> [**ped**|**stratum** <stratum_indicator>]. Prints sibship-matched case-control conditional logistic regression results for all *marker* loci versus the *trait*, where sibships contain

affected and *unaffected* individuals at the binary trait. If the *ped* or *stratum* modifier is present, then the stratification is by pedigree of level of the stratifying variable.

268. **stratified_association** <*trait*> [<*marker*>] <*stratum_indicator*> Performs fixed and random effects score tests on a specified locus or all marker loci versus the *trait*, stratifying on the provided stratum indicator.
269. **interaction_association** <*trait*> [<*marker*>] <*stratum_indicator*> Performs the fixed and random effects score tests on a specified locus or all marker loci versus the *trait*, stratifying on the provided stratum indicator, but returns the heterogeneity test result as the summary statistic.
270. **trend** <*quantitative trait*> <*marker*> [**permute**] Prints Jonckheere-Terpstra trend test result for the specified *marker* loci versus the *trait*. Most suitable for SNP markers, where genotype ordering is clear. Monte-Carlo empiric P-values are produced either via gene-dropping or permutation.
271. **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.
272. **pen** <*locus1*> <*locus2*>). Performs Penrose [1935, 1937] sib-pair linkage statistics for two loci. Quantitative traits are treated as if categorical. The output consists of a two-by-two table of the sibling concordances at the two loci.
273. **apm** <*binary trait*>|(<*quantitative trait*> **>|>=|<|<=|==|^=** <*threshold*>) [**ibd|ibs**]. Prints APM statistics for all *marker* loci versus the *trait*.
274. **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.
275. **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.
276. **qtlpair** <*quantitative trait*> [**full**] [**cqe**] [**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).
277. **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.
278. **lod** <*marker locus 1*> <*marker locus 2*> [<recombination fraction>]. Performs two-point lod score linkage analysis for codominant markers. The algorithm is fairly slow in the presence of many missing genotypes.
279. **summary** [<*N highest results*> | **plot** [**qq**] [<*EPS file*>] | **table** | **dump** [<*output file*>]| **get** <variable_name>]. List the N (defaulting to 5) most significant P-values from the last linkage or association analysis. If the **plot** keyword is issued, a Postscript plot of the $-\log_{10}$ P-values versus map position of the active marker loci tested for association or linkage in the last analysis, unless the **qq** modifier keyword is present. In that case, a quantile-quantile plot of the $-\log_{10}$ P-values is produced. This is saved to a file (default name "sib-pair.eps"). For the **table** modifier keyword, a table of binned P-values or test statistics is produced, usual bin sizes are powers of ten. The **dump** modifier prints all the P-values to a file. The **get** <varname> modifier extracts numerical values from locus annotations

taking the form *varname=value*, or from the *i*'th column.

280. **summary combine** [<*marker locus 1*> [... <*marker locus N*>]]. Combine P-values from test statistics for the specified loci using the Fisher method, and new Cauchy method of Liu and Xie [2018].

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 20 alphanumeric characters, and an individual's personal ID up to 14 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. Slashes dividing alleles of a genotype are optional. 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 data into Sib-pair:* Sib-pair can now read a few file formats other than its native formats (notably PLINK, MERLIN, pre- and post-Makeped Linkage format). Sib-pair cannot handle multiple datasets simultaneously, though you can use the "update" command to merge new genotype data into your existing pedigree for example. I have also written a number of external utilities to do that kind of work -- see the website for programs such as *mergeped*. Often I do preprocessing in **R**.
- *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 *undrop* 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
undrop
```

- *Test for duplicate individuals* Several different commands are useful. The *test duplicates* option is the most general, but one can limit the individuals to be compared by either *test ped1 id1 ped2 id2* or *mztwin find*.
- *Write out only the markers:* Use *keep $m*, then *write pedigree*. Note that this will not return previously dropped markers to the analysis (to this, you would need to first *undrop*.
- *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 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, or precede the command with *let*.
- *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 *undrop* 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,

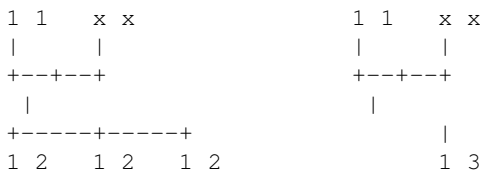
```
set loc dummy aff
dummy=y
tdt dummy
```

- *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 num > 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
The BLUEs (Sib-pair blu):	.6818	.2045	.1136

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 addition of covariates to the analysis. 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. Using the "set liab" allows you to write liability classes to programs such as MLINK.
- *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 [Stirling 1981], 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 [Wichmann and Hill 1982]. 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, retired], 241, 275. Matrix inversion for likelihood evaluation of the variance components model is performed using the LINPACK [Dongarra et al 1979] routines DGEDI and DGEFA (which replace AS6 and AS7 as of April 2008). Extraction of eigenvectors and eigenvalues is done using the appropriate EISPACK routine (RS). Likelihood maximization is performed using either AS319 ("varmet") [Koval 1997] or BOBYQA [Powell 2009], both of which seems to do a good job of it. Conditional logistic regression is performed using AS162 [Smith et al 1981]. Evaluation of multivariate normal cumulative probabilities uses routines from MVNDSTPACK [Genz 1992], or alternatively from TOMS717 [Bunch et al 1993]

The Scheme interpreter is a port of the Minischeme and Tinyscheme interpreters. Minischeme is the work of Atsushi Moriwaki and Akira Kida, following Matsuda and Saigo (Programming of LISP, archive No 5 1987, 6-42). Tinyscheme is based on Minischeme and is more fully featured. It was written by Dimitrios Souflis (dsouflis@acm.org). The Sib-pair Scheme (like Minischeme) does not support vectors, bignums and complex numbers. It does offers double precision floating point and 64 bit integer arithmetic, and provides most of the standard arithmetic and string handling functions.

LIMITATIONS

The Fortran 77 old versions of Sib-pair had limits on pedigree size, numbers of loci and allele numbers per loci. The new version uses allocatable arrays, and so is limited only by memory. Note that rewrites in 2012 allow work data to be stored as files, so one can analyse datasets that do not fit into available memory.

Locus names, and pedigree and individual identifiers are restricted to a maximum length (20, 20, 14 characters respectively). Locus annotations are also restricted to 40 characters each. These can be increased by recompiling after changing the constants in module idstring_widths or locstring_widths (this may make old binary images unreadable).

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. The MCMC algorithms need even longer runs to be ensure that they have reached stationarity.

There are only a few multipoint procedures: *diseq*, *qtl*, *multihomoz* and *haplotype*. The *twopoint* command has not been ported (yet) to the Fortran 95 Sib-pair.

COMPILATION

The program is (fairly) standard Fortran 95, and has run successfully on PCs (under DOS, NT or Linux), SUN Sparcstations, DEC Alpha, Macintoshs running OSX, and HP9000 workstations. The code compiles using **g95**, **ifort**, **gfortran**, and **sun f95**, and currently I distribute Windows and Linux binaries. The code has sometimes encountered problems compiling using the Linux **ifort** with optimization (-O) on.

I have done some comparisons of different Fortran compilers.

Job 0:
read locus plink hapmap3_r1_b36_fwd.CEU.qc.poly.recode.mapfile

read pedigree hapmap3_r1_b36_fwd.CEU.qc.poly.recode.ped
set imp -1; set che off
run

Job 1:
read locus plink hapmap3_r1_b36_fwd.CEU.qc.poly.recode.mapfile
read pedigree hapmap3_r1_b36_fwd.CEU.qc.poly.recode.ped
run

Job 2:
read locus plink hapmap3_r1_b36_fwd.CEU.qc.poly.recode.mapfile.gz
read pedigree hapmap3_r1_b36_fwd.CEU.qc.poly.recode.ped.gz
run

Program	Job 0	Job 1	Job 2
i7			
Sib-pair gfortran -O2	6.4m	6m24s	6m38s/6m45s
Sib-pair g95 -O2	11m39s	-	11m57s
Sib-pair sun f95 -O2	3m15s	-	3m36s/3.7m
gunzip	19.9s	-	-
plink	2m15s	-	-
merlin	-	3m50s	3m46s

Running the standard testsuite (testsuite.in) gives:

64bit	
g95	8.52 s
gfortran 4.4.3	6.64 s
Pathscale 4.0.10 gcc version	6.64 s
32bit	
g95	11.20 s
gfortran 4.6.0	7.38 s
gfortran 4.4.3	8.56 s
open64	9.32 s
sunf95	9.62 s

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.
- Almasy L, Blangero J (1998): Multipoint quantitative trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**: 1198-1211.
- 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*.
- Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS (2003): Novel case- control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *Am J Hum Genet* **73**: 612-626.
- Brandner FA (1933): A Test of the Significance of the Difference of the Correlation Coefficients in Normal Bivariate Samples. *Biometrika* **25**: 102-109.
- Bunch DS, Gay DM, Welsch RE (1993): Algorithm 717: Subroutines for maximum likelihood and quasi-likelihood estimation of parameters in nonlinear regression models *ACM Transactions on Mathematical Software (TOMS)* **19**: 109-130.
- Canning C, Thompson EA, Skolnick E (1978): Probability functions on complex pedigrees. *Adv Appl Prob* **10**: 26-61.
- Choi Y, Wijsman EM, Weir BS (2009): Case-Control Association Testing in the Presence of Unknown Relationships. *Genetic Epidemiology* **33**: 668-678.
- 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 R, Resnick S (1984): Tail Estimates Motivated by Extreme Value Theory. *Ann Stat* **12**: 1467-87.
- 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.
- Dongarra JJ, Moler CB, Bunch JR, Stewart GW (1979): LINPACK Users' Guide. *Philadelphia: SIAM*.
- Elston RC, Stewart J (1971): A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**: 523-542.
- Endelman JB, Jannink J-L (2012): Shrinkage Estimation of the Realized Relationship Matrix. *G3* **2**: 1405-1413.
- 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.
- Fernandez SA, Fernando R (2002): Technical Note: Determining Peeling Order Using Sparse Matrix Algorithms. *J Dairy Sci* **85**: 1623-1629.
- Filliben J (1975): The probability plot correlation coefficient test for normality. *Technometrics* **17**: 111-117.
- Gauderman WJ (2003). Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genet Epidemiol* **25**: 327-338.
- Genz A (1992). Numerical computation of multivariate normal probabilities. *J Comp Graph Stat* **1**: 141-149.

- 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.
- Hastings WK [1970] Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**:97-109.
- Hedges LV, Olkin I (1985): Statistical methods for metaanalysis. *San Diego: Academic Press*.
- Hill HM (1975): A simple general approach to inference about the tail of a distribution. *Ann Statist* **3**: 1163-1174.
- Janss LLG, Van der Werf JHJ, Van Arendonk JAM (1992): Detection of a major gene using segregation analysis in data from several generations, Proceedings of the 43rd Annual Meeting of the European Association of Animal Production, Madrid (Spain), Vol. 1 of Session 5a Free Communications, p. 144.
- 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.
- Koval JJ (1997): Algorithm AS 319: Variable Metric Function Minimization. *Applied Statistics* **46**: 515-521.
- 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.
- Kuonen D (1999): Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**: 929¿935
- 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 (1997): Mathematical and statistical methods for genetic analysis. New York: Springer-Verlag.
- Lange K, Boehnke M (1983): Extensions to pedigree analysis. V. Optimal calculation of Mendelian likelihoods. *Hum Hered* **33**: 291-301.
- 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.
- Li CC, Sacks L (1954): The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10**:347-360.
- Madsen BE, and Browning SR (2009): A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*. **5**:e1000384.
- McPeek MS, Wu X, Ober C (2004): Best Linear Unbiased Allele-Frequency Estimation in Complex Pedigrees *Biometrics* **60**: 359-367.

- Meyer MC (2001): An alternative unimodal density estimator with a consistent estimate of the mode. *Statistica Sinica* **11**: 1159-1174.
- Moskvina V, Schmidt KM (2008). On multiple-testing correction in genome-wide association studies. *Genet Epidemiol* **32**: 567-573.
- Noreen EW (1989): Computer-intensive methods for testing hypotheses: an introduction. *New York: Wiley*.
- Oakes D (1982): A concordance test for independence in the presence of censoring. *Biometrics* **38**:451-5.
- Oakes D (2008): On consistency of Kendall's tau under censoring. *Biometrika* **95**: 997−1001.
- 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.
- Penrose LN (1935): The detection of autosomal linkage in data that consists of pairs of brothers or sisters of unspecified parentage. *Ann Eugen* **8**: 133-138.
- Penrose LN (1937): Genetic linkage in graded human characters. *Ann Eugen* **8**: 233-237.
- Plummer M, Best N, Cowles K, Vines K [2006]: Coda: Output analysis and diagnostics for MCMC. R package version 0.10-7.
- Pons O, Chaouche K (1995): Estimation, variance and optimal sampling of gene diversity. II. Diploid locus. *Theor Appl Genetics* **90**: 122-130.
- Powell MJD (2009): The BOBYQA algorithm for bound constrained optimization without derivatives. *Department of Applied Mathematics and Theoretical Physics Technical Report* **DAMTP 2009/NA06**
- 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.
- Rubin DB (1987): Multiple imputation for nonresponse in surveys. *New York: John Wiley*.
- Shao J, Tu D (1995): The jackknife and bootstrap. *New York: Springer*
- 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.
- Schelling M (2004): Deterministic calculation and stochastic simulation in multi-point linkage analysis. Doctor of Technical Science Dissertation, *Swiss Institute of Technology, Zurich*.
- Sham PC, Purcell S (2001): Equivalence between Haseman-Elston and Variance-Components Linkage Analyses for Sib Pairs. *Am J Hum Genet* **68**:1527-1532.
- Sheehan N [1990]. Genetic reconstruction on pedigrees. PhD thesis, *University of Washington, Seattle*.
- Smith B, Boyle J, Dongarra J, Garbow B, Y Ikebe, V Klema, C Moler [1976]. Matrix Eigensystem Routines, EISPACK Guide, Lecture Notes in Computer Science, Volume 6, Springer Verlag, 1976.
- Smith PG, Pike MC, Hill AP, Breslow NE and Day NE (1981): Algorithm AS 162: Multivariate Conditional Logistic Analysis of Stratum-Matched Case-Control Studies. *Appl Statist* **30**: 190-197.
- 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.
- Stirling WD (1981): Algorithm AS 164: Least Squares Subject to Linear Constraints. *Applied Statistics* **30**: 204-212.
- Tarone RE (1979). Testing the goodness of fit of the binomial distribution. *Biometrika* 66: 585-590.
- Therneau TM, Grambsch PM, Fleming TR (1990): Martingale-based residuals for survival models. *Biometrika* **77**: 147-160.

- Thompson EA (1991): Probabilities on complex pedigrees: the Gibbs sampler approach. In: Computer science and statistics: 23d Symposium on the interface, pp 371-378.
- Thornton T, McPeek MS (2007): Case-Control Association Testing with Related Individuals A More Powerful Quasi-Likelihood Score Test. *American Journal of Human Genetics* **81**: 321-337.
- Thornton T, McPeek MS (2010): ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *American Journal of Human Genetics* **86**: 172-184.
- 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.
- Wang T, Fernando RL, Stricker C and Elston R C (1996): An approximation to the likelihood for a pedigree with loops, *Theor Appl Genet* **93**: 1299-1309.
- 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.
- Wei GCG, Tanner MA [1990]. A Monte Carlo implementation of the EM algorithm and the Poor Mans's Data Augmentation algorithms. *J Am Statist Assoc* **85**: 699-704.
- Whitehead A, Whitehead J (1991: A general parametric approach to the meta-analysis of randomized clinical trials *Statist Med* **10**: 1665-1677.
- 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.
- Wichmann BA, Hill ID (1982): Algorithm AS 183: An Efficient and Portable Pseudo-Random Number Generator. *Applied Statistics* **31**: 188-190.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011): Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics* **89**: 82-93.
- Yates F (1948): The analysis of contingency tables with groupings based on quantitative characters. *Biometrika* **38**: 176-181.
- Yazdi MH, Visscher PM, Ducrocq V, Thompson R (2002): Heritability, reliability of genetic evaluations and response to selection in proportional hazard models *J Dairy Sci* **85**: 1563-1577.
- Young A (1995): Genetic Analysis System, version 2.0 [Computer program]. *Oxford: Oxford University*.

LICENCE

I am happy for others to download, use, redistribute, and adapt my own code as they please, with appropriate acknowledgment of the original authorship. Users should note that I make no warranty as to fitness of the software for any purpose ;).

A number of subroutines included in Sib-pair come from other sources (see above), and are identified as such in the comments in the source code. Some of the statistical routines are from the journal Applied Statistics, and were, as I understand it, made available subject to the restriction that no fee is charged for redistribution.

Contributed code is of course welcome.

PROGRAM HISTORY

24-Apr-2020 (1.00b)

Write input files for Findhap program ("write [loc] fin").

22-Apr-2020 (1.00b)

Refactoring (eg replacing shiftr with ishft, lgamma with alngam) so successfully compiles with LLVM Flang 7.0.1. Fixed "read csv" to handle genotypes properly.

10-Apr-2020 (1.00b)

Fixed `findstr()` behaviour when wild cards and metacharacters in search string - "wri ped" etc would skip some eligible targets. The HWE P-value from the "table" command skips males for X-linked SNPs.

31-Mar-2020 (1.00b)

The "test map" and "rename map" commands can read from a Sib-pair binary data file.

23-Mar-2020 (1.00b)

The memoization of large inverse kinship matrices was not working when the matrix was for one pedigree in a set - hashing fixed.

19-Mar-2020 (1.00b)

Change of 1-D minimizer for `pchisqsum()` has made results more robust. Using "read csv" is equivalent to "read cases" but infers the type of each variable and automatically declares them. The "write [map] findhap" commands write map and genotype files for the *findhap* 4.0 genotype imputation program.

14-Mar-2020 (1.00b)

Fixed regression in "llm" - this utilized the idx array, but this was only initialized for use in subroutines like `sort_table()` and `print_table()`. Note that the detailed "llm" output sorts the cells by count.

13-Mar-2020 (1.00b)

Fixed regression in reading of compressed VCF files with DOS line endings - was due to refactoring. Fiddled with "skat" p value approximation. This was segfaulting when compiled using Oracle Fortran due to floating point errors and a failure of the 1-D minimizer, but only on that platform. Restricting the number of eigenvalues contributing to the estimate has fixed this.

10-Mar-2020 (1.00b)

Implemented simplest versions of SKAT ("skat"). In process, added `(dbeta)` and `(pchisqsum)` functions to Scheme. Pairwise results from "des" now include all "familial" pairs. The "show ann" reads VCF INFO variables and allows summaries and cross-tabulations. One can also transfer INFO variable values to the Sib-pair locus names ("rename vcf <VCF_file> <var>"), to locstat ("read stats vcf <VCF_file> <var>) for manipulation using "sum" and other commands. These can also be dumped to output. One can write ID-major SNP dosage files for the BLUPF90 program. The "file bin" command summarizes the contents of a Sib-pair datafile without reading the whole thing into memory. The Schaid and Summer TDT can be iterated over all loci (in the usual fashion).

19-Feb-2020 (1.00b)

Nasty bug in "merge bed" if merging to a *marker* as opposed to a *snp* - typo in incrementing read position. I had missed since usually use the latter locus type for large datasets. Fixed.

18-Feb-2020 (1.00b)

The "test ids" can read a Sib-pair binary file. The relative pair correlation tables for "describe" now include an entry for all members of the same pedigree. The `(active-individual? <idx>)` Scheme command indicates if that individual is in an active pedigree. One can now write the ID-major SNP genotype dose files used by BLUPF90. One can (finally) gene-drop from the command line (currently requires all founders to be genotyped).

11-Nov-2019 (1.00b)

The "simulate pedigree" command can be run multiple times incrementally, and takes an optional string to prefix the pedigree ID for that set.

05-Nov-2019 (1.00b)

The "rename" function will take advantage of a tabix index.

25-Oct-2019 (1.00b)

The "merge bed" command now allows matching loci by position instead of by name. Prettification of "sum tab" output.

26-Sep-2019 (1.00b)

The `(insert-record! <pos>|<id>)` routine did not update parental pointers correctly. Added `(set-father!)`, `(set-mother!)` and `(set-imztwin!)` to allow the low-level manipulation of these - it is important that generational ordering remains intact after such operations!

20-Sep-2019 (1.00b)

The low level `(insert-record! <pos>|<id>)` command inserts a new data record before the specified position. The "data" modifier to the "read bin" command (that skips reading in the included Scheme image) was broken - fixed. The "head pedlidl<var> command also failed to produce output - fixed.

12-Sep-2019 (1.00b)

The "test ids <file> command is a renaming of "hash file" that can now read IDs from VCF files. Recall that the *chosen* (automatic) variable records which ID matches the external file. The "merge genotype csv" command reads genotypes from an Illumina type CSV file (includes "[Header]" and "[Data]" sections, and variables: "SNP Name", "Sample Name", "GC Score"', "Allele1 - Forward", "Allele2 - Forward").

02-Sep-2019 (1.00b)

Found further errors in "read chain": unmasked when loci not in map order (always just for smaller subsets of markers) - hopefully all now fixed. The "file vcf liftover" command updates the VCF file map positions using the Sib-pair map. The `readline()` subroutine now opens files as streams so seeking is supported - previously defaulted to an ordinary Fortran read.

02-Aug-2019 (1.00b)

The "file vcf order" command can be used to sort, reorder and subset a VCF file, writing directly to a new VCF file.

31-Jul-2019 (1.00b)

The algorithm for "read chain" assumed the loci to be in map order - fixed to work on unsorted maps.

24-Jul-2019 (1.00b)

The "read chain" command uses a UCSC chain file to change map coordinates from one build to another. The "write vcf" command now includes a list of contigs ie chromosomes in the resulting header (some programs cannot process an unsorted VCF unless this is present).

09-Jul-2019 (1.00b)

Fixed regression in "read loc vcf" when tabix index files present. For large VCF files, "file vcf" estimates the number of loci by dividing file size by a locus compression ratio based on a subset of records. New Scheme commands are `(file-delete)`, `(file-list)`, `(regexp)`. Added Cauchy combination of P-values due to Liu and Xie [2018].

09-May-2019 (1.00b)

Fixed "clr" where strata are large - was because size of one work array was hard coded as size of the largest family.

30-Apr-2019 (1.00b)

Fixed same problem (regression) from refactoring, now to "file vcf".

27-Apr-2019 (1.00b)

Fixed further problems (regression) from refactoring, now to "list chrom".

26-Apr-2019 (1.00b)

Fixed further problems (regression) with refactored table command.

03-Apr-2019 (1.00b)

Fixed `asciiplot` for unordered loci.

01-Apr-2019 (1.00b)

Altered "read loc merlin" to append to existing loci - original behaviour was to expect it to be the only source of locus information. Added ASCII Manhattan plot to "sum plo" ;).

23-Mar-2019 (1.00b)

VCF related commands now take advantage of tabix index files, if present. The "blup" command works more nicely for traits. The "file tbi" command prints the matching record (or INFO variable values). The "relative" command can output kinship with propositus (already did that for trait values). Prettification of output.

01-Mar-2019 (1.00b)

When writing numeric alleles to a VCF, these are replaced by dummy polyA alleles, with length being the allele rank eg 192,194,196 to A,AA,AAA. Some programs are strict regarding the contents of a VCF alleles field. Added test for NaN in "clr"/"sdt" likelihood.

24-Feb-2019 (1.00b)

The "relative" command can include "kinship" in the resulting tabulation. Documented "set twin error" in "help" output.

22-Feb-2019 (1.00b)

Tabulations of some VCF INFO variables were not correctly sorted - fixed. Repaired spurious warning when comparing to VCF maps (changed `int()` to `nint()` - Sib-pair map is double precision). Prettied up table printing by `print_table()`.

05-Feb-2019 (1.00b)

The "show annotate" will produce a histogram for a quantitative INFO variable.

01-Feb-2019 (1.00b)

Fix to isodate() to avoid overflows. The "read annotate" command can extract particular variable values by name from VCF file INFO strings eg "ANNO:Consequence" would copy this to the front of the Sib-pair locus annotation string where it can be searched on. The "show ann" will tabulate values of such strings. Categorical variables pretty transparently can be entered and manipulated as string values eg "edit ped1 person1 colour to red". The category levels are stored in the annotation string in the form "1=first 2=second...". The "llm" log-linear model formula will accept the exponentiation operator to mean all interactions to that set of variables:
(A + B)^2 => A B A*B (no "-" operator yet).

12-Dec-2018 (1.00b)

Fixed `merge vcf reference` - if allele counts (AC etc) present, will select most frequent allele as reference, but this was failing if these variables were absent.

13-Nov-2018 (1.00b)

Now can "show map where <string> -- finally extended search for this and for `list` to allow multiple search strings. Incorporated routines to read and index BGZF files. Can (finally) use tildes in file paths for all commands (some would not expand). Can declare j2000 as epoch using "set epoch". Bumped up pedigree and individual ID string length to 25. New version can read old binary pedigree files, but not vice versa.

29-Aug-2018 (1.00b)

Commands to read and write from FImpute. The "reorder" command was badly behaved when $t was specified - fixed the code dealing with automatic variables.

07-Jul-2018 (1.00b)

Refactoring of code to process VCF files, especially in "ass...vcf" where output more informative. Added "recode...vcf..major" to supplement "recode...ref".

20-Jun-2018 (1.00b)

Fixed behaviour of "ass vcf" for multiallelic markers.

16-Jun-2018 (1.00b)

Fixed reading of missing data for categorical variables - these were being assigned as a level of the factor. Levels of binary variables (if not annotated) in a table had regressed to "1" and "2" - back to "n" and "y". Widened column with for table factor levels to 20, when number of factors < 3. Fixed reading of long macro names in macro loops. The "print where" command respects "set cat". The "merge bed" is faster, as random accesses bed file.

22-May-2018 (1.00b)

Bugfixes "fil Met" - not flipping alleles or strand correctly.

21-May-2018 (1.00b)

The "read locus file" command allows reading and declaration of loci from a rectangular data file with columns header. The "file Metasoft" command reads multiple files, extracts the statistics (currently BETA and SE) for each matching locus and writes these in "wide" format in a single file. The macro variable syntax now allows access to lists: *%NAM*[<num>] is the *num*th member of the list *NAM*. Internally, increased size of maximum wordsize for `args()` - this was limiting utility of macro loops iterating over long file names.

08-May-2018 (1.00b)

Reading in of VCF files was getting new error messages trying to parse lists of values in the INFO fields: fixed. The "file vcf" and "read loc vcf" now allow ranges including the chromosome (CHR:BP) - was just the basepair coordinates.

24-Apr-2018 (1.00b)

String searching admits meta-characters \< and \> for word beginnings and ends, and bracket expressions (wild-card lists of characters, [*ccc*]). The "read anno" command reads VCF INFO variables and appends the matching values to the Sib-pair locus annotations. Minor segfault from new code when no categorical variables were declared - fixed.

19-Apr-2018 (1.00b)

Bugs causing segfaults in main REPL loop and in wrcsv() found. Only uncovered when different evaluation order in compound test statement was chosen by compiler. For example, "`(ntwins > 1 .and. dataset%plocus(i,twinning) > KNOWN)`" failed when `twinning` was missing, even though `ntwins` = zero. Fortran compilers are allowed to evaluate logical expressions like this in whichever order they think more efficient.

Sib-pair now can read and write string values for categorical traits. The "set cat" command determines whether levels or labels of a categorical trait are printed.

The anova macro makes a nice ANOVA table for the fixed effects in a mixed model analysis (`var`, `mft`). No fancy corrections I'm afraid.

08-Apr-2018 (1.00b)

Bug in "read vcf" when IDs in form `ped_id` - miscounting pedigrees by one, so segfault. Fixed.

23-Jan-2018 (1.00b)

The Tarone chiq-square routine in "tab ped" checks for a prevalence of unity. There is a limit of 1000 records from "lis" and "ls" in interactive use, unless *plevel* > 0.

12-Jan-2018 (1.00b)

Problem with "merge geno" under unix when file has DOS line ends - if the last word was an allele, then it had a carriage return appended. Fixed by adding an additional test to the `nextword()` routine. The "merge geno" command now reads Illumina style reports (format: marker id GC_score A1 A2) and filters out genotypes with a quality score under a threshold, default 0.6.

21-Nov-2017 (1.00b)

The "read stats" command accepts an integer indicating the file column to read. The "keep where spectrum" command accepts multiple strings (but not wildcards - ranges can be mimicked using macro variables).

20-Oct-2017 (1.00b)

The "read stats" command will read in a value for each locus to locstats from a file. This can then be used by the "sum" and "keep" commands. The "test vcf" command's summary output has been enhanced. The "merge vcf" filtering can now use a whole locus quality score eg imputation R^2. The output of the "show kin" command summarizes the currently active "big" kinship matrix (usually an empirical kinship matrix for the entire sample). Enhanced zlib functionality.

05-Apr-2017 (1.00b)

Added "rare" command to count rare alleles in cases and controls. Tweaked `readmap()` to respect file suffixes - eg if the first few locus names in a .bim file looked numeric, it would fail to recognise the file type. The "test" command will list all discordant genotypes when given two IDs to compare. The "show ids" command can now subset on wild cards. Noted problem reading from a .bim file - if a locus name was purely numeric, which includes strings of the form ddddd-dd (allowable in Fortran), this locus was skipped. When the .bed file is then merged in, these will be skipped, fortunately without wider repercussions.

03-Feb-2017 (1.00b)

Another nasty bug in reading VCF files - reserved characters within ID strings were splitting single IDs into multiple IDs, so genotypes were off by the extra columns. Fixed.

16-Jan-2017 (1.00b)

The "tab polychoric" option gives polychoric correlation estimates for a pair of variables. The shell script wrapper needed quoting of command line arguments - fixed.

05-Dec-2016 (1.00b)

The "read grm" and "write grm" commands read and write GCTA's binary kinship matrix format files to and from Sib-pair's internal "big" kinship matrix. The "set kin" command allows the user to set the "big" `kinmat`, currently A, C or add a ridge constant ("head kin" allows one to check).

30-Oct-2016 (1.00b)

The "swap <trait>" command tests for and repairs likely allele swaps with particular levels of a categorical trait. The "test flip <trait>" command just tests for such allele swaps versus the trait. The "ass <trait> vcf" command compares allele frequencies at the different levels of a trait to that from a VCF file containing population allele frequency data as INFO variables (eg "AC" and "AN" -- "set vcf" changes these defaults).

24-Oct-2016 (1.00b)

Both "test vcf" and "merge vcf" can filter on a genotypic quality score.

17-Oct-2016 (1.00b)

The "test vcf" command compares genotypes in the current dataset to those in a VCF file. The "ref" modifier to the "recode" command replaces missing genotypes with wild type homozygotes, which can be filtered on an indicator trait.

13-Oct-2016 (1.00b)

The labels of genotypes in output from the "tab" command were incorrect if the alleles at that system included letters and numbers: fixed.

07-Oct-2016 (1.00b)

Polychoric correlations for flat contingency table entered via keyboard ("pol"). Familial polychoric correlations for ordinal traits ("des <trait>") with jackknife standard errors and derived WLS heritability estimate. The "file print" command includes CSV and tab-delimited output options, and can skip a requested number of lines at the beginning of the file to be read. Several commands now accept a "pos" keyword which then allows loci to specified as ranges of map positions. Commands reading VCF files can read those where ANNOVAR has prepended annotations ("ann" modifier keyword).

22-Jul-2016 (1.00b)

Weird bug when repeatedly "read loc vcf". Specifically test1.vcf (10 markers), test2.vcf (970 markers OK, 971 markers causes duplicate locus declarations). Was caused by storage expansion causing the locus name hash table to fall out of synch - fixed.

13-Jul-2016 (1.00b)

"show map", "keep/drop/undrop" and "list" now accept a map position or range of positions when the "pos" or "where pos" modifier is present.

15-Apr-2016 (1.00b)

Added "wri DISTmix". Case (ie nonpedigree) data without an ID field can now be read. The "hash" command keywords have been rationalized. The macro braces will expand a sequence of the form {N : M}.

28-Jan-2016 (1.00b)

Using index of variable eg "rename 1" out by number of automatic variables `ENVNUM`: fixed. Testing sex fits a mixture model to the number of available Y-locus markers.

21-Dec-2015 (1.00b)

The "read loc plink" command will flip the annotation reference alleles around if the first one is "0".

10-Dec-2015 (1.00b)

The "hash <file>" command now indicates matches in the untyped() array, accessible via "chosen".

08-Dec-2015 (1.00b)

The "keep whe chr|pos" commands allow the "trait" modifier to toggle subsetting markers or traits - useful for quantitative traits with a map position ie expression.

06-Dec-2015 (1.00b)

Sib-pair produced .bed files sometimes had extra bytes appended - I think when an old file was already present: fixed. The "file fasta <fil> index" command indexes a FASTA file. The "casecontrol" command no longer decorates the pedigree ID to produce a unique pedigree ID for each person (unless "new&quot is added). The "write snp" command will write ID-major datasets containing allelic dosages if the "dose" modifier is present.

27-Nov-2015 (1.00b)

Unfortunate bug in "merge vcf" - ignored possibility of unmatched ids, so quietly read in wrong data: fixed. Also fixed "read vcf" where was being passed uninitialized variables.

25-Nov-2015 (1.00b)

The "read loc plink" but not the "read bed" command was overwriting the automatic variables in the locus data tables, causing "lis" etc to be off count. Added "test flip <source>"

18-Nov-2015 (1.00b)

Cleaned up "keep whe r2"- although "keep" and "drop" are supposed to the same for this option, didn't seem to work correctly. Output from "blu" now mentions trait subsetting if it was requested. Did some work on "mulhom <trait&qt;" - though still does not include LD in simulation, so need to prune markers before use. The `(stats)` deals with NaN more nicely - treats as missing.

15-Nov-2015 (1.00b)

Reorganising the various strand commands. "test flips" merely tabulates inconsistencies between the annotation strand and the external map strand. The "flip fasta" and "test flips fasta" commands compare the annotation reference allele with the consensus sequence from an indexed (.fai) FASTA file - the former will flip or swap the **annotation**. The "file fasta" command summarizes the contents of the .fai file (which contains counts for each chromosome/sequence).

Found a long-term bug in "residuals" and "predict" command, where allelic coding was not used for first marker - would instead use mean of alleles, so usually no effect.

12-Nov-2015 (1.00b)

The "flip map" command will now flip the strand of the allele at a monomorphic marker if it matches the reference allele. Previously, these were skipped. If multiple matches for the marker are found in the map file, an allelic inconsistency in the later match no longer overwrites the locstat record of the preceding match.

01-Nov-2015 (1.00b)

If overwriting an existing map, if the new map was in different units, then the rescaling would affect unmatched (ie old) loci if the new map file format wasn't VCF etc: fixed. If loci are declared by sequential "set loc" statements, checking for duplicate names currently takes increasing amounts of time as the the dataset grows. This testing is already disabled for PLINK, VCF etc, but now can be optionally turned off for all cases: "set check names on|off"

27-Oct-2015 (1.00b)

The "sdt" and "clreg" commands can stratify on a variable, and strata may span pedigrees.

19-Oct-2015 (1.00b)

"llm" sample weight not initialized, so standard usage no longer worked: fixed. Increased capacity of hash table (table of prime number sizes) to 79999987 records. Speeded up locus hashing for "read bed" and "read vcf" - this was being done incrementally, rather than once at the end: fixed. The "show spectrum" command tabulates (nucleotide) allele spectra for markers.

08-Oct-2015 (1.00b)

"rename map" did not update locus name hash: fixed. If logfile already exists, new commands are now appended rather than overwriting.

02-Oct-2015 (1.00b)

The "mul" required the markers to be sorted in map order to estimate F_{roh} - now sorts internally. The inbreeding coefficient for each indidividual can be written to a trait. The "read map" command did not recognize some VCF files (with a nonstandard header): fixed. The "test map" command hung if the first marker was unmapped: fixed. The "tab" and "llm" commands allows a sample weight to be utilized ("sampleweight" and "weight" keywords).

21-Sep-2015 (1.00b)

Updated strings in GFF files the Sib-pair recognizes. Tightened up "flip map" to recognize the major allele of ambiguous (CG, AT) SNPs, match on position as well as name, and to check and update all annotation records of reference alleles. The latter is used by "write vcf", so one can match required strandedness. (Standard errors for allele frequencies can be saved in the `allele_class` container - currently used by the test for the major allele in "flip".) Fixed reading of annotation reference alleles so always Ref then Alt. Using "sum tab" after many commands affecting SNPs such as "flip" gives tabulation of changes or differences.

Part of my current pipeline for imputation is:

```
macro chr=1
read bin genos.bin.gz
keep where chromosome %chr
order $mm
# drop indels
drop where spectrum "ID"
# test strand problems using allelic spectrum and HWE
test strand
# merge data where multiple assays same SNP
# allows for strand swaps
test map merge
```

```
# drop the superfluous duplicates
drop whe test >= 2
# drop triallelic markers
drop whe all > 2
pack
# match strand to HRC reference data (record in annotation)
flip map
/home/davidD/Genetics/Maps/HRC.r1.GRCh37.autosomes.mac5.sites.tab.gz
# filter for VCF output
set loc use aff
if (protyp > 0.9) then use = y
write vcf chr%chr.vcf use
$ bgzip chr%chr.vcf
```

15-Sep-2015 (1.00b)

Added "read vcf" to read IDs from a VCF file. The "wri sna" writes a SNP-major dosage file for the WOMBAT "--snap" association option.

11-Sep-2015 (1.00b)

The output from "dis" for SNPs as calculated by the cubic equation approach sometimes chose an inferior solution when haplotype counts were low - fixed. The "write vcf" command now allows filtering on an indicator trait (only individuals not missing at the trait will be written. Also, default behaviour when writing VCF files is to make the major allele the reference allele, unless the annotation includes contrary information (EDITED: "[R/A]", where A is the alternate allele, and R the reference allele).

24-Aug-2015 (1.00b)

Improved behaviour of `(intersect)`, `(list-select)` etc when given empty lists: now eg, `(list-select '() '()) => '()).`

22-Aug-2015 (1.00b)

Added "quantile_normalization" command.

21-Aug-2015 (1.00b)

Added an additional automatic variable "chosen" which marks those records affected by or contributing to the last operation - use to identify merged records etc. Tidied up "file vcf" - prints all loci in an interval if plevel > 0 (used to be 1). Flat tables from "tab" of SNP genotypes now include row totals.

19-Aug-2015 (1.00b)

"ass <qua_trait> snp" does SNP regression with output including betas and SEs. Added "ped_id" option to "merge vcf" and "write vcf". "file vcf" prints more of the annotations. The "recode nuc" command converts "A/B" to nucleotides where the appropriate annotation is present for the locus eg "[C/T]". Matrix inversion automatically tries adding a small ridge constant when encountering problems. The data used by "var" is now rescaled internally so BOBYQA will be happier. The "nor" command tests for quantitative trait normality across multiple traits (eg gene expression datasets), and allows subsetting on the resulting statistic. The "sum" command provides an FDR estimate for each locus.

31-Jul-2015 (1.00b)

"test map <file>" compares the current map to that in the specified file.

27-Jul-2015 (1.00b)

"read locus vcf <file> human" recognizes chromosomes 23-26. The "head" and "tail" commands now allow the standard locus type subsetting eg "head loc $A". Fixed reading male X haplotypes from VCF files - were being set to missing.

10-Jul-2015 (1.00b)

Fixed problem with "merge vcf" where missing genotypes were sometimes given spurious values.

29-Jun-2015 (1.00b)

Added "sum get <varname> to extract numerical values from locus annotations. One can now "select ped in file".

11-Jun-2015 (1.00b)

The `load` command wiped the rest of the Scheme command buffer - fixed. Also added `lochash-update!`, and `paste`.

12-May-2015 (1.00b)

The command "get sibling" did not correctly sample all eligible siblings - fixed. Added extra output from "kin inb": number with F > 1/16.

17-Apr-2015 (1.00b)

The empirical P-values for quantitative trait association analysis using the "assoc" command were occasionally overly significant, which would disappear on rerunning. Was due to OMP parallelization of calls to `doanova()` - which uses dataset%untyped as a work array. This array needed to be private - fixed. Another OMP problem with "neff" causing segfaults - as workaround. stripped out OMP directive.

17-Mar-2015 (1.00b)

Implemented `read-char` and `peek-char`.

12-Mar-2015 (1.00b)

"mgt" sometimes gave spuriously significant results where pedigrees with no genotypes at that marker did have trait data. Fixed, by imputing completely missing pedigrees to mean dose in "gpe".

11-Mar-2015 (1.00b)

Added "get mzt", and dropped MZ contributions from "get sib".

09-Mar-2015 (1.00b)

Fixed segfault if added covariates to "mgt".

04-Mar-2015 (1.00b)

Noticed "test map" skips the last marker - fixed.

26-Feb-2015 (1.00b)

Cleaned up "keep where pos" bug when specifying a chromosome. Some new Scheme commands: `environment-bound?`, `unique` and `duplicated`, `intersect`, `setdiff` and `union`.

19-Feb-2015 (1.00b)

Fixed bug in reading .bim files with long allele strings (heterozygote genotypes could be stored in both orders eg A/G and G/A, causing segfaults).

17-Feb-2015 (1.00b)

Cleaned up bugs in "ken" bivariate survival analysis code, adding in bootstrap and permutation tests. Allowed "read bed" to accept long allele strings.

04-Feb-2015 (1.00b)

The "pack" command now works before pedigree data is read in: this allows manipulation of locus lists before merging in select data. Commands that calculate inbreeding coefficients automatically switch to MC methods if the kindred is too large. Renaming of loci can avoid checking for existing duplicates (with current hashing, slow for very big lists) If test for genotype discordance between MZ twins or duplicate samples, the number of discordances is stored in locstat, so it can be examined using the "sum" and "keep"/"drop". Summary F statistics are available to Scheme via `stat-result` (1=Fis, 2=Fis, 3=Fst). This allows me to implement a macro that gives all pairwise F_{st} between multiple populations.

06-Jan-2015 (1.00b)

Using `loc-set!` does not update the locus name hash table - added hash update check to "ls". Finally, "undrop $t" works as expected. The Moskvina et al effective sample size is accessible using `stat-result` eg

```
#
# ad hoc gene-based etc test window <trait> <m1>...<mN>
#
macro window
 keep %0
 ass %1
 neff
 eval (let ((neff (stat-result "stat")) \
            (loci (ls "m"))) \
          (pchisq (* -2 (apply + (map log (map locstat loci)))) (* 2 neff)))
```

01-Jan-2015 (1.00b)

The "set err" command can set the acceptable mistyping rate and minimum number of markers used for the fast test of duplicate samples or MZ twins ("test dup"). Crosstabulations using the "tab" command can include missing as a category by adding "showmiss" modifier. The "mgt" command accepts a list of covariates for the quantitative trait mixed model. Fixed bug in "test map merge" where small map distances (1 bp) were rounded to zero.

26-Aug-2014 (1.00b)

Added "protyp" automatic variable to supplement "numtyp". The "test map" command checks for duplicated map positions and can merge these if requested. The default behaviour of Sib-pair is to always try to fit all data into memory (the default memory size threshold is now "0", interpreted as "do not test"). Duplicates are marked as a locus statistic, accessible using the "sum", "keep" and "drop", commands. The "show ids dup" command now can save a similar style indicator to a trait variable, and the "tes dup ids mer" can test for identity and merge genotype and phenotype data for identical individual IDs (in different pedigrees). The "read probs" command handles Beagle, Impute2 and Mach, and the merge key can be specified more precisely. The "merge loci" command can merge data for two loci.

23-Jul-2014 (1.00b)

Implemented long term plan to make "automatic" individual-level variables such as sex and number of typed markers directly accessible to statistical procedures (such as "tab", "reg", "llm"). Fixed a couple of garbage collection bugs in intrinsic Scheme procedures that produce lists. Added `loc-set!`, (`locpos` and `locpos-set!`, which normal users will not need), `set-sex!`, `allele-freqs`. The per locus Mendelian error rate is now saved for analysis (by "summary"). One can now select loci based on observed homozygosity (main use where homozygosity zero or one). Cleaned up reading of integers to support exponential notation as well as K,M,G as suffixes to decimal numbers eg 3.2K. Sex and Mendelian error checking specifically flags Y-locus heterozygotes. The *evdtail()* function returns the naive result when the tail is flat.

02-Jul-2014 (1.00b)

Added more XLispStat statistical procedures to Sib-pair Scheme (`sort`, `rank`, `order`, `quantile`), a number of standard R5RS procedures (`char=?` and friends, by aliasing to the equivalent string procedure); fixed some bugs along the way (eg scanner treatment of backslashed backslashes).

22-Jun-2014 (1.00b)

Fixed bug(s) in "strat" so correct association statistic calculated - command now documented. Added in option to specify a covariate other than a SNP. Fixed theshold for new memory allocation for Scheme to *memsiz*/2, as used by Nils Holm (http://www.t3x.org/) to improve performance of his version of miniscm - this makes a big difference. Similarly rewrote `(define-macro)` to follow his more standard implementation. Moved car/cdr compositions to Fortran, and added "(dir)" to list top-level variables and closures, radixes to `(number->string)`, a high precision timer `(system-clock)`, and the XLispStat commands `sample`, `select` (here called `list-select`) and `which`.

27-May-2014 (1.00b)

The "set loc" command can recognize map positions of the form "<chr>:<coord>".

07-May-2014 (1.00b)

Minor enhancement to "recast affection": recodes "0/1" or "1/2" to "n/y" for categorical variables. The `(pmvnorm)` Scheme procedure can estimate multivariate normal CDF probabilities using either the Mendell-Elston or Genz algorithms. Fixed up printing of some tables with large counts (eg contingency tables were usually limited to five digits).

07-Apr-2014 (1.00b)

Added TOMS717 (Bunch et al 1993) implementation of Mendell and Elston (1974) algorithm for approximate MVN probabilities. Seems to give reasonable results a lot quicker than mvnstdp.

19-Mar-2014 (1.00b)

Refactoring (which mainly involves moving stuff into modules). Added BOBYQA as alternative optimizer - seems twice as fast fitting variance components and MFT models, but this may reflect less strict tolerances.

07-Feb-2014 (1.00b)

Refactoring. Added "rename map" to automatically update names of loci with the same map position as in a reference map file (currently VCF only). Added a "write vcf" command.

10-Jan-2014 (1.00b)

The "test" output of genotypes was garbled - fixed. The inverse kinship matrices used by "mqls" etc are saved for reuse, with some gain in speed. When the kinship matrix is nonpositive definite, the Moore-Penrose inverse is now calculated. Some code refactoring.

12-Dec-2013 (1.00b)

Fixed problem in "mqls" command - contribution of untyped but phenotyped individuals was being calculated incorrectly for the original (pedigree based kinships) algorithm (these were not large effects, so I only noticed them when comparing to results using the new empirical kinship algorithm). Added Scheme commands `(set-pedigree-name!)`, `(set-individual-name!)`, `(allele-names)`, `(current-second)`, `(set-data!)`.

21-Aug-2013 (1.00b)

The "mqls" and "wqls" commands can use an empirical kinship matrix to do a whole sample corrected analysis a la ROADTRIPS. Also added modifiers to allow use of robust or naive variance estimators, and add a ridge constant. Further improvements to the "test strand" command.. Wider locus names allowed in most tabular outputs. The "test lange" command allows use of the Lange-Goradia elimination algorithm for Mendel checking. Tidied up stored P-values for "qtl full" (these are multimarker, and weren't lining up with the correct locus).

08-Jul-2013 (1.00b)

The "set chi-square" command selects whether a Pearson or Gibbs (LR) chi-square is used as the test statistic for categorical trait association analysis. The Scheme `(data-counts)` command produces a tabulation of trait value levels.

08-Jun-2013 (1.00b)

Fixed segfault in "merge bed" (only if target locus to merge declared as marker instead of SNP). The "merge bed" command "id" modifier causes the merge key to be individual ID rather than pedigree plus individual ID.

05-Jun-2013 (1.00b)

The "keep/drop where spectrum" command selects SNPs with the matching allelic spectrum.

03-Jun-2013 (1.00b)

The "sum dump" command dumps all the P-values or test statistics to a file. By default, missing sexes that cannot be inferred via the sex of a mate are no longer filled in with a random value. The "impute sex" command uses sex chromosomal data to impute sex. The `(format)` statement allows binary and octal output formats to be specified, and now allows tabs ("`~T`"). Categorical trait annotations giving level labels

are no longer truncated: they are stored as a Scheme/macro variable called labels_<trait>.

28-Feb-2013 (1.00b)

Chengrui Huang pointed out that .bim files produced by "write bed" didn't have a space before the minor allele when the marker is monomorphic: fixed. Added gene-dropped Jonckheere-Terpstra for allelic nonparametric SNP association test.

04-Jan-2013 (1.00b)

The "test strand" command tests if observed genotypes at a SNP are consistent with a mxture of strands ie four alleles but only two types of heterozygote. The "keep" and "drop" commands can subset on the number of observed alleles for a marker. MENDEL type binary files are recognized by the "read bed" command.

14-Nov-2012 (1.00b)

The "sho chr" command behaved badly when non-standard chromosome identifiers were used (infinite loop): fixed.

08-Nov-2012 (1.00b)

Fixed bug in SNP data after "subped" command. Added `(map-position-set!)`, `(chromosome-set!)` to Scheme. The "set append" controls how duplicate locus declarations are treated (skip or name mangle). Extra output from "test sex", and "set sex"can specify the homozygote to heterozygote error rate for X-loci used for the likelihood calculation. The "set mis" command can specify a missing data token for missing genotypes. The "kee whe all <op> <number_alleles>" selects markers based on the number of alleles. The "lod" command can estimate two-locus haplotype GPEs or dosage scores for the first locus (the latter can be used to estimate single-locus *ibd*).

25-Oct-2012 (1.00b)

Added "p" option to "file print". Fixed bug in permutation P-value for case kinship test. Further tweaks to the "merge bed" command: sometimes had to run twice to get to work. The "show chrom" command now prints range of map values for active loci on each chromosome as well as count.

09-Oct-2012 (1.00b)

Segfault due hash table being too small: fixed. The "sho chr" command shows map lengths. The "wri sol gen" command would skip some eligible loci when markers were subsetted: fixed. The "file wc" gives the length of the longest word in a file.

05-Oct-2012 (1.00b)

Changes in "varcom" had introduced errors, as starting value AE model fitting was not correctly initialized: fixed. Added "merge dose" to read in PLINK format SNP allele dose files.

04-Oct-2012 (1.00b)

Cleaned up multilocus IBD routine: it was confused when sets of loci were presented out of map order (recall that it clusters markers within a set distance). The "varcom" command now runs an AE model first to obtain reasonable starting values - some variance components linkage models were failing to converge.

02-Oct-2012 (1.00b)

"wri snp" writes a PLINK SNP-major `.tped` format file, while "wri snp" remains the row-major allelic dosage file that ROADTRIPS reads. The "set app" option controls treatment of duplicate locus names: the default is still to modify the name by adding a version number, but such declarations can now be skipped, for example if a file merge is to be performed. The "wri csv" now allows the genotype missing data token to be different from the the phenotype missing data token. The "var" command can use an empirical kinship matrix ("within pedigree", recalling the pedigree may contain unrelated components. Behaviour of "merge bed" fixed so that loci declared as markers rather than SNPs are still updated correctly. The "mqls" now uses the empirical null variance rather than the theoretical variance (under HWE), as described in Thornton and McPeek [2010].

14-Sep-2012 (1.00b)

Added permutation P value for case mean kinship compared to rest of pedigree.

12-Sep-2012 (1.00b)

Fixed bug in "merge vcf": was not reading complete line.

17-Aug-2012 (1.00b)

Added "merge mach"

13-Aug-2012 (1.00b)

The "hwe" command with no arguments would skip non-SNP X-linked markers: fixed. The "read map" command can now recognize and read VCF files with map information. Prettification of some output eg "surv" output now includes an extrapolated empirical P-value, and respects declared column separator. The `csimped()` subroutine made thread safe: this affected the "assoc" command for quantitative traits for OpenMP enabled code.

04-Aug-2012 (1.00b)

Fixed problem with "file vcf" when only one locus being searched for (would not find). Test for duplicate locus names was missing some cases after a code revision: fixed. Note that for speed reasons this test is not used by all locus reading commands.

03-Aug-2012 (1.00b)

Added "keep where near". Reorganised command line calling of Sib-pair to allow easy use with a file browser by associating a file type with Sib-pair: if the first argument ends in ".in", ".bin", or ".bin.gz", it will be included or opened. The `wrpap()` function now sorts *trip.dat* (previously these sometimes had to be sorted before PREPED would accept them). Reorganised the "merge plink" command so it is closer to the other merge commands, in that it will update genotypes for previously declared SNP markers. This allows merging of multiple `.bed` files. Fixed "test locus" handling of sex chromosomal SNPs: was skipping these. The "sum tab" command includes a cumulative proportion of markers with associated P-value less than or equal to bin threshold.

12-Jul-2012 (1.00b)

The "merge probs" command can use IDs as well as a merge key to match up individuals to genotypes. Tinkered with "read map" to infer map units based on size of position values. Improved "file transpose" speed on large files by buffering. Added search range to "file print"

05-Jul-2012 (1.00b)

The "read map" command will read position information from a GVF file (eg Homo_sapiens.gvf.gz from Ensembl).

04-Jul-2012 (1.00b)

Added commands for reading VCF files: "file vcf", "read loc vcf", "merge vcf".

22-Jun-2012 (1.00b)

Added a couple of utilities "file transpose" and "file inverse". Width of pedigree and individual ID columns increased for many output tables.

13-Jun-2012 (1.00b)

Fixed behaviour of hashing of long locus names: was hashing full length of entered name rather than maximum stored length (currently 20 characters).

24-May-2012 (1.00b)

Fixed segfault in "merge genotypes". Due to `matrix_copy()` shrinking target matrix to size of first matrix, ignoring any requested size increase.

18-May-2012 (1.00b)

Added "set map units"

01-May-2012 (1.00b)

Assorted minor cleanups.

27-Apr-2012 (1.00b)

The "homoz" command gets asymptotic P-values. The "write dot" command includes doubled edges from parent to marriage node for inbred matings. The "test dup" command can test for duplicates across the entire dataset. Code is multithreaded for most single locus tests, with four-fold speed improvement on eight-core machine.

12-Mar-2012 (1.00b)

Tidied code. Automated reading of old Sib-pair binary pedigree files (reads version of file from header).

02-Mar-2012 (1.00b)

Added virtual memory for large genotype datasets: stored in memory. A performance hit due to refactoring around this, and I will continue to look at this. If zygosity indicator declared categorical, was not correctly dealt with: fixed.

01-Feb-2012 (1.00b)

Fixed segfault in "surv" when scanning markers. The "sho sex" command tabulates counts of each sex. If compiling using the Open64 compiler, stream i/o now works.

05-Jan-2012 (1.00b)

Subtle bug in `fgz_rewind()`: this needs to zero `zbufpos` (the module-wide global used by `fgz_read()` to record state of the read buffer). Added command to merge in most likely genotypes based on genotype probability files written by BEAGLE and Minimach.

19-Dec-2011 (1.00b)

Added "show chromosomes". PLINK map file physical positions now kept in locus annotations, and written out by "write map plink".

13-Dec-2011 (1.00b)

Added `(isatty?)` to Scheme

11-Dec-2011 (1.00b)

Still garbage collection problems with Scheme: hopefully all fixed with saving of "value" register (not saved in miniscm, but is in tinyscheme version of code).

01-Dec-2011 (1.00b)

The "rel" caused a segfault if a SNP was chosen to be added to the output: fixed. The "seg" command segfaulted for X-loci: fixed. The "read loc plink" command now accepts an "append" modifier (it is much quicker for large numbers of named loci than declaring loci in the script). The "tail map" and "tail loci" commands give the correct number of records. The scheme `(help)` command now accepts a search string.

18-Nov-2011 (1.00b)

Various Scheme data accessors. The "sdt" command accepts the "ped" modifier, so test is within pedigree rather than within sibship (immediate idea is to test for mating asymmetry in allele frequencies). Further fiddling with Windows "sib-pair.ini" file.

15-Nov-2011 (1.00b)

"read loc merlin" did not allocate enough space for "S2" loci (ie skipped), even though they are needed - Sib-pair reads in all the data, then marks "S2" markers as dropped.

08-Nov-2011 (1.00b)

Fixed some infelicities in "mft", mainly degrees of freedom of tests. Added Scheme accessors to phenotype data: `(nobs)`, `(npeds)`, `(pedigree-name)`, `(pedigree-size)`, `(pedigree-members)`, `(individual-name)`, `(father)`, `(mother)`, `(imztwin)`, `(sex)`, `(data)`. The "kin case" command prints out the correction factor of DeGeorge and Rosenberg [2009] for F statistics calculated from families. Added "sml age" (age of neutral allele based on frequency, population size). Added "pop" modifier to "grr case" for when controls are from the general population (ie unknown trait status). AND, finally squashed bug in Scheme interpreter, where garbage collection clobbered work variables used by some subroutines (weren't declared to the GC).

13-Aug-2011 (1.00b)

Standard errors for fixed effects part of "varcom" model now provided.

21-Jul-2011 (1.00b)

The "mft" command carries out VC analysis under the threshold model. The "set mft" command control the integration (Alan Genz's MVNDST.

13-Jul-2011 (1.00b)

Upgraded "fil pri" utility slightly.

08-Jul-2011 (1.00b)

Added "rea ped nop nos".

30-Jun-2011 (1.00b)

Added "wri pli".

16-Jun-2011 (1.00b)

Added a few more scheme functions, notably `char-upcase`, `char-downcase`, `list?`, `make-closure`, `string-ref`. Revised some functions so match standard eg `gcd` can have zero or many arguments.

08-Jun-2011 (1.00b)

Fixed a few more deviations of Sib-pair Scheme from standard, including making quasiquoting and thus Scheme macros work properly.

24-May-2011 (1.00b)

To produce Beagle marker list, added "wri loc bea". Handling of allele names for compressed SNPs repaired

13-May-2011 (1.00b)

Can write parent-offspring trios for Beagle: "wri bea ... tri". The "permute" command permutes trait values within pedigrees.

18-Mar-2011 (1.00b)

Reorganised "write morgan" so consistent with MORGAN 3 command language. Added dispersion estimate to "reg" output. The "order" command no longer segfaults when there are no eligible markers for "$mm".

11-Mar-2011 (1.00b)

The "edit" command did not recognize categorical traits: fixed. Added "mztwin find" to set a zygosity indicator using genotype data (hard threshold on genotype discordance). Fixed segfault in "fpm" when categorical covariates were included in the model, and an observation was missing, and another due to an uninitialized variable. Some cleaning up of code, mainly fixing `intent` of subroutine arguments. Where an unobserved genotype had been imputed to include unobserved alleles (by the "sim" command), this lead to error messages from the `getnam()` function: if this occurs now, `getnam` silently replaces the allele with the commonest observed allele.

24-Feb-2011 (1.00b)

The printed fixed effects parameter estimates from "varcom" were from the "E" model, rather than the best fitting VC model: fixed. This didn't affect testing of fixed effects by LRT.

09-Feb-2011 (1.00b)

Adds `(locstat-init!)`, Davis and Resnick empirical P-values for "hwe".

03-Feb-2011 (1.00b)

The "dis" command did not work on compressed SNPs: fixed.

13-Jan-2011 (1.00b)

Added extreme value distribution parametric survival analysis. Refactored glm code in passing. Tweaked "llm" so that `(stat-result)` can obtain model G.O.F. P-value.

17-Dec-2010 (1.00b)

The recent refactoring broke binomial regression with "reg" -- fixed. The "set mod" command allows automatic genotypic coding of marker loci in the regression based procedures (previously one created a suitable categorical variable).

07-Dec-2010 (1.00b)

Refactored "varcom" "clreg" and "fpm" so categorical variables handled correctly.

05-Dec-2010 (1.00b)

Now reads labels for levels of categorical variables from the locus annotation, where the annotation contains "value=label" pairs. The labels are used for all analysis output. As part of refactoring of scanner code to support this, old behaviour of compressing whitespace in annotations was removed.

29-Nov-2010 (1.00b)

"merge genotypes" will merge in data read as one record per genotype. Sib-pair Scheme bindings for EGGX/proCALL graphical library.

18-Nov-2010 (1.00b)

Fixed "keep where pos" so includes snp data.

03-Sep-2010 (1.00b)

Adds "write map plink". Hopefully fixes regression in "kin roa".

29-Aug-2010 (1.00b)

Added "set pwd", so can use file chooser to select working directory.

27-Aug-2010 (1.00b)

Empirical P-value for QTDT was incorrect if optimized ("-O2 -no-fast-math") compilation of Fortran code: the summary statistic generated by a simulation that should have been equal to the observed statistic was systematically smaller. Fixed by making comparison use epsilon rather than zero.

16-Aug-2010 (1.00b)

Added "`(time)`", and changed "date" to give today's date if no arguments.

16-Aug-2010 (1.00b)

Fixed evaluation of "`(let* () ...)`".

15-Aug-2010 (1.00b)

The "allelic" modifier to "llm" fits an allelic model for the first marker in the model: ie HWE for genotype frequencies, and multiplicative effects of allelic dose. The "`(locstat-set!)`" allows modifying the recorded statistic for a locus. The "`(append)`" procedure now is correctly variadic.

13-Aug-2010 (1.00b)

Now Sib-pair Scheme's "`(format)`" recognizes "~w,dF" for a flonum, and a new "`(stat-result)`" procedure allows access to results from the last model fit: "pval", "lik", "npars", "lrt", "df", "stat", "var".

10-Aug-2010 (1.00b)

Memory usage output now uses long integer size()'s. The "strat" command now allows comparisons. The "`(loctyp)`" correctly deals with categorical and SNP types, and "`(format)`" can read simple Fortran format statements. Direct writing to gzipped files now possible (zlib).

30-Jul-2010 (1.00b)

Added `(map-position)`, `(chromosome)`, and `(version)` to Scheme interpreter. The "kin roa" command was printing the coefficient of relationship rather than kinship coefficient: fixed.

28-Jul-2010 (1.00b)

Fixed "write bin" and "read bin" so large SNP datasets don't fail during the unformatted write (presumably due to buffer size in 32 bit versions of compiled code). The (internal) twin pointers were not written correctly by "subpedigree": fixed. Made changes allowing compilation with openf95.

26-Jul-2010 (1.00b)

Added "keep|drop in <file>: selects loci listed in file (one locus per line). Cleaned up syntax for "keep where coverage". Fixed "merge plink" so not fazed by extra phenotypes.

14-Jul-2010 (1.00b)

The "keep <locus name>" command crashed if there were no loci already declared: fixed. Added a "keep|drop where in <file>" option.

09-Jul-2010 (1.00b)

Added "keep where snp", "kin roadtrips". The "wri snp" command writes a SNP-major (ie one row per SNP) genotype file where SNP genotypes are encoded 0, 1, 2. The "list" now respects the `plevel`. Fixed bug in "kin" (was miscounting when trying to identify MZ twin pairs).

28-Jun-2010 (1.00b)

Minor fixes to on-line help. The "loc rel" command uses chromosome location from the map as well as from annotation. The "plo" will use categorical as well as quantitative variables to choose symbol type.

24-Jun-2010 (1.00b)

"wri csv" did not print out compressed SNP names: fixed.

15-Jun-2010 (1.00b)

The "sib" command could produce P-values of NaN (if only half-sibs), which failed to be sorted correctly by `srank()`: fixed.

25-May-2010 (1.00b)

The "lrt" command was not working when comparing "var" variance components models with covariates present: fixed.

21-May-2010 (1.00b)

The "wqls" test would crash if only one trait level was observed for a marker: fixed.

20-May-2010 (1.00b)

Some operations such as "subped" did not work on a compressed SNP dataset: fixed. Also fixed intermittent bug after "clear" due to locus name hash not being cleared.

19-May-2010 (1.00b)

The macro loop "{ $m }" was ignoring compressed SNPs: fixed. The "$<class>" list function now can order by test statistic, and "$mmr" does give a reverse map order.

14-May-2010 (1.00b)

The "typ <trait>" command was ignoring compressed SNPs: fixed. Added "sum tab" to bin P values or summary statistics.

11-May-2010 (1.00b)

Most of JAPI GUI library can now be called from Sib-pair Scheme (I have skipped the graphical routines: canvas etc).

07-May-2010 (1.00b)

Fixed quotation bug for empty macro variables (where passing to Scheme for evaluation). The output of more (now most) "set" commands respects the `plevel`. The "repeat" macro allows iteration of a command ("repeat <n> <commands>"). This is quite short:

```
eval (define (s-repeat n cmdstring) \
(let loop ((i 1)) \
(if (<= i n) \
(begin \
(run cmdstring) (loop (+ i 1))) \
(format "# end of ~d iterations~\%" n))))
macro repeat
eval (s-repeat %1 "%+2")
;;;;
```

05-May-2010 (1.00b)

Quantitative and categorical trait WQLS test ("wqls") working. The former may need some later refinement.

29-Mar-2010 (1.00b)

Fixed reading of long integers into Scheme (increased read buffer length from 20 characters/digits to 40). The variance correction for the corrected chi-square ("mqls") is now printed in output when *plevel* greater than zero).

22-Mar-2010 (1.00b)

Added double precision reals (back) to Scheme interpreter, along with `(pnorm)` etc. This will make Scheme portion of old binary images unreadable.

17-Mar-2010 (1.00b)

Rejection sampling approach to simulating QTL genotypes under the mixed multifactorial threshold model implemented: the "sim qtl" command.

12-Mar-2010 (1.00b)

Extrapolation of empirical P-values when observed statistic exceeds all simulated values following Davis and Resnick [1984].

09-Mar-2010 (1.00b)

"fpm" offset command was broken: fixed.

08-Mar-2010 (1.00b)

Added "ass fre", "ass maf", "ass ris", which give different sets of summary results for SNPs.

26-Feb-2010 (1.00b)

Added sex testing via Y markers. The "mit" and "yha" association commands would analyse each other's loci: fixed.

25-Feb-2010 (1.00b)

Fixed problem with Windows version: was a constant (PORT_STANDARD in module ioports) whose definition was incorrectly ifdef'ed out. Thanks to Bahram Namjou, who reported the problem.

24-Feb-2010 (1.00b)

"kin ibs" actually added. Added nancycats.in example. Found regression problem with Windows version of Sib-pair: is not reading in files correctly. Temporarily, I have rolled "sib-pair.exe" back to 2009-09 version.

23-Feb-2010 (1.00b)

Refactored IBS-based kinship estimation (adding "ibs ibd" and "kin ibs"). Added "qch" and "pow" commands, and "ncp" modifier to "pch". The "reg" command now calculates a model likelihood, which it passes for use by "lrt"

12-Feb-2010 (1.00b)

"read locus plink" was reading blank lines as a locus: fixed. Estimation of IBD from IBS.

06-Jan-2010 (1.00b)

`(string->number)` now does BOZ transformation if requested (as per R5RS).

5-Jan-2010 (1.00b)

Added "tet". The "sml" and related commands print expected tetrachoric correlations for relative pair types. The "pairs" print style prints numbers as integers when appropriate. Result of last algebraic evaluation or simple commands (eg "tet") can be saved to a macro variable

24-Dec-2009 (1.00b)

In some cases, `open_infile()` failed to check if file already open.

23-Dec-2009 (1.00b)

Expanded syntax of "keep where" to allow explicit comparison operators where applicable.

21-Dec-2009 (1.00b)

Gene-dropping P-values for "surv". Documented and improved BLUE allele frequency estimates for subgroups. The PLINK ".bim" and ".fam" files may be gzipped. The "sho mis" command summarizes patterns of missingness by locus.

11-Dec-2009 (1.00b)

Added "-" as missing parental ID marker. Added "macro <- pval" to save P-values eg from log-rank and Kruskal-Wallis tests. Made sure variables declared as categorical get printed by `pedout()`. Fixed bug in "kru" if variables not a consecutive sequence in dataset (due to undeclared interface).

03-Dec-2009 (1.00b)

Fixed bug in measuring record length for gzipped files: was sometimes a few characters short. Started adding support for Unix pipes (currently affects "file cat" and "print".

30-Nov-2009 (1.00b)

"mqls" fails gracefully if unable to allocate a large enough pedigree covariance matrix.

25-Nov-2009 (1.00b)

Fixed bug in `unidens()` when observations too close together.

19-Nov-2009 (1.00b)

Reading gzip compressed input files via *zlib* is now faster: uses `gzread()` and a buffer, rather than `gzgets()`.

10-Nov-2009 (1.00b)

The "surv" command carries out the log-rank test.

05-Nov-2009 (1.00b)

Fixed bug in tabulation level printing. The "lifetable" command variable types were broken, fixed, and made declaration of strata slightly more logical.

03-Nov-2009 (1.00b)

Sib-pair can now call zlib directly to uncompress data files. Three-way and higher tabulations can now be represented as a crosstabulation with the last variable as multiple columns. This includes the "kru", so that the strata can be combinations of multiple variables. A variant of the DerSimonian-Laird type random effects model is available for binary trait analysis in the presence of a (categorical) covariate: the "strat" command. Fixed bug in the Scheme `read-line` command reading from the standard input. The "read map" command now checks for files where record take the form "chr mappos marker_name" (Allan MacRae pointed out this problem reading map files generated by Mega2).

23-Oct-2009 (1.00b)

Fixed bug when adding a new marker locus to an existing pedigree file. If extra storage space needed, would clobber last previously declared marker as expanded array. Also fixed behaviour of "mqls" when subset of pedigrees selected and when MZ twins are present.

21-Oct-2009 (1.00b)

Added "read locus merlin snp". Finished prettifying P-value printing (could fail if P-value=0).

19-Oct-2009 (1.00b)

Implemented MQLS association test of Thornton et al 2007.

29-Sep-2009 (1.00b)

Implemented BLU allele frequency of McPeek et al 2004.

23-Sep-2009 (1.00b)

Bug in genotype imputation where iterative peeling gave zero likeihood: fixed so now give missing value rather than NaN.

22-Sep-2009 (1.00b)

Small P-values from association analysis now given in exponential notation, rather than as 0.0000. The "mgt" command give VC association analysis for quantitative traits with pedigree-based imputation of missing genotypes.

18-Sep-2009 (1.00b)

Code cleanup (especially of intent statements) so compiles without complaint using Sun Fortran compiler. Still some problems compiling with the Intel compiler, which may be compiler bugs. Fixed stutter in reading long lines of inline data: would sometimes join words together if fell at read buffer length (currently 20000 characters).

08-Sep-2009 (1.00b)

Added "set mem" to preallocate sufficient memory for a dataset, cleaning up `expand_geno()` etc along the way. Added general ability to read gzipped map, locus and pedigree files (decompresses to a work file using system call to gzip). If declared many extra loci, sometimes hash table completely filled and was slow to reexpand: fixed.

03-Sep-2009 (1.00b)

The "mer/upd" commands were recently broken by the changes for compressed SNPs: fixed. Stopped repeating "run" from (often) leading to a segfault.

02-Sep-2009 (1.00b)

Added "$A" and "$D": all active and dropped loci respectively.

01-Sep-2009 (1.00b)

The "kee/dro whe" commands were recently broken for compressed SNPs: fixed. Fixed a couple of array bound-width overruns of no apparent effect.

28-Aug-2009 (1.00b)

Recent changes to `setup_plink()` broke "read hapmap": fixed.

27-Aug-2009 (1.00b)

Assorted tidying up. Changed *zp()* to Alan Miller's code, so more accurate for extreme deviates. Hashing in parser, so evaluation of algebra significantly faster for large datasets. Reading PLINK binary files now allows for parents who do not have their own record in the *.fam* file. More work on "mgt" command.

20-Aug-2009 (1.00b)

With *plevel*=-2, the pvalue from "dis <marker>" was not being calculated to be saved for the "sum" command: fixed.

18-Aug-2009 (1.00b)

Refactored code to allow 8 bit storage per allele if requested, and added additional classes of variable type (separate classes for mitochondrial and Y chromosome genotype data), as well as removing some hidden limits on the maximum size dataset that algebra could be performed on (parser was limited to a maximum of

100000 markers, but would give incorrect results rather than failing outright). Most of these changes are invisible to the user.

23-Jul-2009 (1.00b)

If `iter`=0, the summary P-value from TDT for the "sum" command was that from the genotypic test, which really requires simulation to avoid being sometimes too liberal. Now defaults to the global allelic test P-value. The maximum allowable length of the format string for the "file print" command has been extended from 40 to 256 chars: this allows nicer scripting. Also fixed case for "dis loc1 loc2", when *loc1* is not active but *loc2* is. Improved map reading (eg EnsEMBL BioMart ordering is: name, chromosome, position).

20-Jul-2009 (1.00b)

Fixed regression reading in large numbers of variables (due to error in new locus hashing).

10-Jul-2009 (1.00b)

Implemented "update inline" so can update from inline data. Put "set work" back (was in documentation but not in program). Change "mzt" algorithm so that MZ twins with nonadjacent IDs are no longer flagged as suspicious. Imputation of missing sexes now correctly deals with MZ twins. Speeded up "read map" (now hashed search). Fixed bug in "davie": if last family contains no offspring, then segfaulted.

03-Jul-2009 (1.00b)

Broadened "und" range of selection criteria (eg chromosomee or map interval or maximum allele frequency). If twin indicator trait specified but dropped from analysis, then the "mzt" would not respect "merlin" type zygosities: fixed.

02-Jul-2009 (1.00b)

For a binary trait, "des" now gives the recurrence risk for MZ twins. Segfault if using "var" with more than one marker locus as covariate: fixed. Fixed "dis" when one SNP monomorphic in crosstabulation: no longer gives NaN.

26-Jun-2009 (1.00b)

Now reads "NA" as missing for quantitative values (`fval()` defaulted to 0) - note that NA is still not usable as a missing parent indicator.

12-Jun-2009 (1.00b)

Removed line length limit to `(read-line)`. The "read bin" read of a Scheme image did not allow for the possiblity that extra procedures might be added to the interpreter: if incompatible version number, skips loading the image.

10-Jun-2009 (1.00b)

Added internal version of `(format)` command to Scheme: knows about "~[:num:][@][ASD~%]", but can't deal with nonatomic arguments. Remember to escape "%" in scripts.

09-Jun-2009 (1.00b)

Fixed value of maximum active pedigree size after "sub" (was setting to size of dataset). Prettification of output from "kin inb".

05-Jun-2009 (1.00b)

Finally got around to implementing delete-*d* jackknife SE for familial correlations from "des" command -- added "set jack" to control *d*. List of locus names is now hashed.

31-May-2009 (1.00b)

The "read bin" and "write bin" commands save the twinning and sex indicators and an image of Scheme, so macros and macro variables are retained as well. These commands are backwardly compatible.

25-May-2009 (1.00b)

The "update" and "merge" commands allow matching just on individual ID (for completeness, added "hash id" command.

22-May-2009 (1.00b)

Added in command to read HapMap type genotype files: "rea hap".

19-May-2009 (1.00b)

Added avuncular and cousin correlations to output of "des" for quantitative traits. Added "wri loc sage <fil> par" to write a modern style SAGE parameter file.

28-Apr-2009 (1.00b)

Updated Windows initialization script so moves into Desktop -- was starting up in the installation directory, which does not work well for users without administrator privilege. If the dataset is already read in, the "set twi" command now automatically checks for genotypic discordances and sets up the internal twin indicator array (it previously assumed that zygosity was declared before the dataset was read in, so this had to be done manually using the "mzt" command).

23-Apr-2009 (1.00b)

Segfaults for PLINK `.bed` files where families are present. This is partly because parental identifiers were uninitialized plus parental pointer not correctly set. However, as currently implemented, Sib-pair does not allow children to precede parents or have parents without records in the `.fam` file. Temporarily, this is "fixed" by setting the parents to missing in these cases. Gu Zhu asked for Merlin map positions to be written to 6 decimal places (this seems to have been increased): done.

06-Apr-2009 (1.00b)

Cleanups around GLMMs.

03-Apr-2009 (1.00b)

Added Tarone score test for extra-binomial variation. Added heritability (etc) estimates for logistic-normal and Weibull GLMM as:
$h^2 = V_{RE}/(V_{RE}+C)$, with $C_{Bin}=pi^2/3$, and $C_{Weibull}=pi^2/6$.

31-Mar-2009 (1.00b)

Fixed linkage disequilibrium chi-square printed by `cubicld()`, the routine called for diallelic markers with *plevel*=0: this was twice as large as it should have been. The "sho map" command allows loci to be specified,

and gives recombination distances when traversing the specified map in either direction (originally only allowed forward ordering).

24-Mar-2009 (1.00b)

Fixed pathname length restriction for "out", and restriction on tail calls from macros (due to a superfluous semicolon appended to the macro text).

23-Mar-2009 (1.00b)

Added "wri sas" to make a SAS command file with inline data "cards". The "set prev" command specifies a trait prevalence for the "ass gen" command, which now produces attributable risks and sibling recurrence risks for each locus.

18-Mar-2009 (1.00b)

Fixed bugs in "wri arl" if trait (population indicator) not specified, and where missing alleles were "0" when should have been "x". Added "tab ped <trait>" to tabulate the trait values for each pedigree in a compact fashion (alternative was to create a pedigree indicator variable and crosstabulate versus that). The "wri dot" command allows specification of the colours used to represent the values of the trait. Got rid of NaNs for Fis for monomorphic loci. Added Scheme `(char->integer)`, `(integer->char)` and

27-Feb-2009 (1.00b)

The command "keep whe covered" keeps/drops markers based on whether genotypes have been observed in all categories of a trait. The "sav" modifier keyword for "fpm" can save BLUPs or genotypes to a variable.

23-Feb-2009 (1.00b)

Added "wri sib" to create Sib-pair script with pedigree data inline. Improved Scheme `(write)`, `(display)` and `(newline)` so can write to a port, as specified by R5RS. Implemented QQ plot for "sum plo".

19-Feb-2009 (1.00b)

Fixed segmentation fault in "mzt drop" when some loci currently dropped from analysis (these are not thinned by this command).

16-Feb-2009 (1.00b)

The "snp" command recodes a dummy quantitative trait representing SNP genotype. The Scheme `(display)` and `(write)` functions finally allow writing to an (open) output port. Fixed up "wri csv" coding of MZ twins where more than one set in a sibship. The "exit" command now exits completely from Sib-pair when called within a script -- "quit" still returns to the topmost REPL.

09-Feb-2009 (1.00b)

The "hap mit|yha" command combines haploid markers into haplotypes (with conservative imputation of unobserved parental haplotypes), while "mit|yha hap" prints frequencies of such haplotypes. The "mit|yha" command will analyse categorical and binary traits. Fixed up bug in P-value simulation for "mit|hap", where first haplotype frequency got used twice.

06-Feb-2009 (1.00b)

Refixed printing of haploid genotypes in tables (would give funny values or a blank). Can calculate IBS based genetic distances for haploid markers. The "kee whe num" command allows a proportion or a number.

05-Feb-2009 (1.00b)

Added file chooser for "set dat".

3-Feb-2009 (1.00b)

Fixed bug in "residuals": intermittently segfaulted if missing covariate data.

18-Jan-2009 (1.00b)

Added ability to keep or drop a marker based on the proportion of markers genotyped (compared to the maximally typed marker).

12-Jan-2009 (1.00b)

Spurious errors messages from "sim" for markers about allele names: fixed. Were due to `rdfreq()` upgrade to read allele names: this is now only allowed for "set fre". Fixed error message from Scheme `(reverse)`

07-Jan-2009 (1.00b)

The plot command can use different plot symbols to represent a third categorical variable. Tidied up bits of Scheme (eg `string-split` skips leading tabs as well as spaces). If you have wget, the "getNCBIsnp" command (macro) can download information about a SNP.

05-Jan-2009 (1.00b)

Using "update/merge", duplicate IDs can be flagged. The current print mask can be saved as a macro variable. The `hashids()` function was silently testing the 0th element of an array -- fixed.

1-Jan-2009 (1.00b)

Minor improvements to Scheme interpreter: "`(run)`" now evaluates the command immediately; "`(load)`" has been implemented.

23-Dec-2008 (1.00b)

Wrong CIs for MZ female twin pairs from "twi": fixed.

19-Dec-2008 (1.00b)

Reimplemented "wri arl", so that multiple populations can be extracted.

17-Dec-2008 (1.00b)

Odds ratios 95%CIs for Nx2 contingency tables used variances rather than SEs -- fixed.

15-Dec-2008 (1.00b)

Principal components analysis for traits and for IBS marker-based estimation of genetic distance between individuals. Multidimensional scaling for IBS. "Nonparametric" twin survival analysis via Kendall's tau for censored data.

11-Dec-2008 (1.00b)

Rationalized Mendelian error checking for data that is already read in. Recent changes here had broken the "gpe" HOPS example. Added option to read in allele names using the "set fre" command.

09-Dec-2008 (1.00b)

Fixed "pack loci", where uninitialized variable led to segfaults. Added an "upd" (or "ins") modifier to the "copy" command: this only updates the target when the existing value of target variable is missing.

05-Dec-2008 (1.00b)

The "ibs" command retooled: now writes mean IBS sharing for all pairs or a subset based on a trait.

02-Dec-2008 (1.00b)

Added odds ratios (plus 95%CIs) for Nx2 contingency tables. Fixed count of variables written by "write bin" (this makes old binary files unreadable).

23-Nov-2008 (1.00b)

Fixed annoying bug in nuclear family Mendelian error checking: sometimes set a genotype in the next pedigree incorrectly (carried across from error in last sibship). Introduced in August reorganization. Also fixed up drawing of Mendelian errors for "test loc", which included imputed genotypes.

21-Nov-2008 (1.00b)

Improved counts of updated records for "merge": these included cases where the original record and update record were both a missing value.

20-Nov-2008 (1.00b)

Fixed error in "flip" command. The "upd" and "mer" commands behave gracefully when a variable in the update list is currently dropped from analysis. Speeded up calculations by "dis" for SNPs (solving cubic equations).

06-Nov-2008 (1.00b)

Added "wri ped ... hea" to write a line containing the names of the variables at the head of the GAS type pedigree. The "upd" and "mer" warn if they have encountered duplicated records in the file they are reading updates from.

04-Nov-2008 (1.00b)

"rea bin" and "rea pli" are allowed long file names, like all the other such commands.

03-Nov-2008 (1.00b)

The "fst" command prints the locus and total H_0, H_S and H_T, when *plevel* > 0.

31-Oct-2008 (1.00b)

Fixed up dumb indexing error in "dis" -- affected the contribution of inferred phased genotypes where number of alleles at two markers unequal, so effect on results usually invisible, but caused segfaults on some platforms, and wrong answers if informative pedigree data. The "wri map mer" command wrote a blank if no chromosome had been specified for the map -- fixed by setting a default, and carrying along any chromosomal localizations of preceding marker.

30-Oct-2008 (1.00b)

Fixed P-values for "ass snp" (were one tailed). Prettified some output.

24-Oct-2008 (1.00b)

Fixed problem with "rea map pli"

23-Oct-2008 (1.00b)

Added "set gen" and "set mis" to set the output allele separator and missing value token. Prettified header for printing pedigree data to screen.

22-Oct-2008 (1.00b)

Had inadvertently blocked default action of the "rec" command on markers -- renumbering alleles to consecutive integers: fixed (in passing, allows recasting of all quantitative or all binary traits). With `plevel` set to -1, the "pri" command no longer outputs the selection criteria or totals.

20-Oct-2008 (1.00b)

Fixed problem when including markers as a fixed effect in variance components analysis (introduced with new storage structure in August).

17-Oct-2008 (1.00b)

The "com" command had been deactivated in April 2008 by careless editing to include haploid loci.

14-Oct-2008 (1.00b)

Added "mod" modulo operator.

26-Sep-2008 (1.00b)

The "rea loc lin" was broken by the change in storage: fixed. Also fixed problem writing Post-Makeped `.ppd` files when strong inbreeding is present. Added "sum ucsc" to write `WIG` format files -- this can be used to add a custom track for the Golden Path browser. Documented "hash <fil>": this matches IDs in the current dataset with those in a file. And "test sex" was also broken by the storage change: fixed.

24-Sep-2008 (1.00b)

Added "neff" command (implements Moskvina V, Schmidt KM Genet Epidemiol 2008 32: 567-573). I will speed this up eventually.

23-Sep-2008 (1.00b)

The `complete()` function was broken for sets of variables mixing markers and traits with the change in storage model: fixed. Expanding the number of loci had stopped working intermittently for the same reason: fixed.

22-Sep-2008 (1.00b)

Vertical printing respects the table separator ("set tab").

19-Sep-2008 (1.00b)

Default action of "update" is now to update all loci. Added "merge", which is "update" but doesn't overwrite existing non-missing values. Trying "copy" (copy active loci from individual A to individual B). Added "set pri ver" to print records vertically. Fixed "typ" when *plevel* > 1. Now can use "which where".

12-Sep-2008 (1.00b)

Reintroduced "wri cri"

11-Sep-2008 (1.00b)

Two-point lod score linkage analysis, currently only for codominant markers. The "typ" command crashed: fixed and speed improved. File chooser for "read bin". The "set nha" command allows one to change the model size for the "dis" command. The "rec" command was broken for markers with the change in storage model: fixed.

04-Sep-2008 (1.00b)

SNP genotype data can be stored as 4-bits per genotype. This is less flexible than the ordinary storage format for genotypes: not able to pack, for example. Currently only from a PLINK .bed file ("read plink <fil> com"). Added Scheme `locnotes-set!`.

01-Sep-2008 (1.00b)

Found bug in regression prediction if more than one marker. The subsequent markers would use *b+mean_allele_size*: replaced with *b*mean_allele_size*.

29-Aug-2008 (1.00b)

Increased Scheme string buffer size to 5000 characters (to match `macbody`).

28-Aug-2008 (1.00b)

Marker alleles are now stored internally as integer*2, rather than real*8 as previously. The "sho mem" command shows the internal pedigree data arrays (similar to R `str()`). The "lis" and "ls" commands can be interrupted. The "set tab" command sets the column separator for summary tables such as those from "fre snp", "hwe", "ass" and "sum".

21-Aug-2008 (1.00b)

Reintroduced reading of "unformatted" (in the Fortran sense) Sib-pair working pedigree file: "rea|wri bin <fil>". These are currently large files, but on systems where gzip is available, can be automatically compressed and decompressed ("wri bin <fil> com"). The file is a Fortran unformatted write of the locus and pedigree arrays, and so will be compiler and platform specific. Fixed allocation of space for loci from "rea mer" etc (was `newsiz-nloci` rather than `newsiz-MAXLOCI`). "lif" accepts "0" in the start variable position (all periods start at time=0). The "rea plink " now reads the chromosome assignment into the new chromosome variable.

20-Aug-2008 (1.00b)

Now "typ" defaults to just the active loci (unless *plevel* > 1). Fixed "pack" handling of chromosome location.

15-Aug-2008 (1.00b)

Fixed minor bug in reading chromosome numbers (was allowing "ch" as a prefix, and overwriting a correct number from earlier in the same line of text).

12-Aug-2008 (1.00b)

Added a chromosome element to the map. Reads the chromosome number from the annotation (searches out "chr[omosome] (NN|X|Y)", or from map files that include that information. Is respected by "$mm" option. A complementary "set chr" command added.

11-Aug-2008 (1.00b)

The "tes loc" allows selective testing for Mendelian errors. Fixed "clr" and "sdt" to work with only one stratum/family. The "fil pri" utility accepts a search string in a "sub-awkian" way.

08-Aug-2008 (1.00b)

Found bug in string searching, so that "1" matched "*1": fixed. This was only noticeable using "pri ped". Now "sho ped lis" gives a list of active and inactive pedigrees, and "pri ped" will reply with a list of active pedigrees if none were specified. Fixed unitialized variable in `zp()`.

07-Aug-2008 (1.00b)

Found bug in test for inequality of genotypes ie "a/a" ^= "a/b" gave `FALSE`, because they share an allele in common: fixed. Some output prettification.

06-Aug-2008 (1.00b)

Added "rea pli" to read PLINK `.bed` and ancillary files, as well as "rea loc pli" to read the PLINK `.map` locus file.

05-Aug-2008 (1.00b)

The "kin <case>" command now prints out coefficients of fraternity for the pairs of affecteds. Output also prettified.

29-Jul-2008 (1.00b)

Gu Zhu pointed out that the "wri mendel" and "wri loc mendel" are giving problems for MENDEL 8.0 format. Binary traits were being written as "2" and "1", but locus file had "AFFECTED" and "NORMAL": harmonized to latter pair. All commands now recognize "new" (the locus and map files were "free").

24-Jul-2008 (1.00b)

Added "set pri" to change format of "pri" command. The default format is now that obtained by the "wri" command ie a rectangular matrix of results.

23-Jul-2008 (1.00b)

Added "loo" command to show inbreeding or marital loops. The "kin" command now takes a "dom" modifier, which prints the non-full-sib relative pairs where the coefficient of fraternity is nonzero. This has been added to help assess the effects of marital loops. The output from "pai" also now includes a flag ("MZ","FS" or "Bi") at the end of each record.

18-Jul-2008 (1.00b)

If *plevel>0*, the "rel" command now prints out a list of shortest paths between *ego* and all other pedigree members.

15-Jul-2008 (1.00b)

The "wri loc sage" command skipped X-chromosome markers: now gives entry for these as well as autosomal markers (thanks to Joan Bailey-Wilson and Peter Lipman for pointing this out).

09-Jul-2008 (1.00b)

To make working with SIMWALK2 easier, the "trait" modifier to "wri men" and "wri loc men" uses "`1`" and "`2`" instead of "`NORMAL`" and "`AFFECTED`".

07-Jul-2008 (1.00b)

The "typed" command gives counts of phenotyped individuals for all loci versus a stratifying variable.

02-Jul-2008 (1.00b)

Minor code cleanups.

30-Jun-2008 (1.00b)

Found problem in FBAT/RC-TDT implementation for untyped matings giving rise to {*A*/*B*, *A*/*A*} offspring: this needed further conditioning on the number of *A*/*B* and *A*/*A*} offspring. Fixed.

24-Jun-2008 (1.00b)

Reintroduced sibship disequilibrium test, as conditional logistic regression: "sdt" and "clr".

20-Jun-2008 (1.00b)

Noticed combination of APM empirical P-values via inverse-Z approach seems to be biased (away from null) if many small pedigrees are analysed. Reinstated Fisher chi-square approach as primary method, with

inverse-Z combined P printed as an additional column when *plevel*=1.

19-Jun-2008 (1.00b)

Reshuffled loops in "assoc" so that categorical trait analysis does not give useless error messages about "level not found". Fixed minor bug in *marginal_table()* which might have affected some printed proportions (but not counts). Found partly uninitialized array in *rctdt()*: fixed. Found incorrect pointer in *doapm()* pedigree trimming: fixed. Prettified tables from "tab".

16-Jun-2008 (1.00b)

Reading exponential notation (eg 1.0e-04+1) was still broken -- this time should be last!

13-Jun-2008 (1.00b)

Adds "grr cas" command giving penetrances etc from case and control allele frequencies.

11-Jun-2008 (1.00b)

Adds "ass snp" command giving brief output for SNP association analysis.

06-Jun-2008 (1.00b)

The "rec num" was setting non-letter code values to missing: fixed by passing through other values unchanged. The "clear" had stopped working: fixed *cleanup_hash()* which needed to check ia hash table had been allocated.

05-Jun-2008 (1.00b)

Noticed similar problem with semicolons in comments within macros: fixed by stripping comments as read macro in (and saves space!).

04-Jun-2008 (1.00b)

Reverted treatment of ";" by "$", as this broke macros using "$". This means semicolons in a system command need to be escaped. Repeating the "run" command (without first issuing a "clear" command) caused Sib-pair to give a run-time error (as dataset memory is already allocated): fixed.

02-Jun-2008 (1.00b)

Fixed up treatment of escaped characters by "$" (evaluated then passed to system). Haplotype (nloci>2 or with trait) labels were out of order in output of "hap" command: fixed.

31-May-2008 (1.00b)

Fixed up treatment of ";" by "$" (passes to system) and "echo" (respects escaped ";"). Added "fact" function.

30-May-2008 (1.00b)

Efficiency of ID hashing improved. Added Scheme string search (`substring?`). Prettified output from "reg" association testing. Hopefully fixed intermittent bug in Windows file picker (tmpfile name could have junk at end).

28-May-2008 (1.00b)

"wri csv" segfaulting when previously declared twinning variable dropped and "pack" issued: fixed. Fixed undefined constants in *isatwin()*. Calculated inbreeding from "hbd" was wrong part of matrix: fixed (the HBD calculation was never unaffected).

27-May-2008 (1.00b)

The "update" command allows phenotypes in the dataset to be updated from a second file.

22-May-2008 (1.00b)

Postscript histogram also produced by "his". If "rep" modifier is used for regression, then genotypes are automatically imputed (previously had to issue "set ana imp").

16-May-2008 (1.00b)

Added an option to "hist" for the number of bins. Removed limit on length of file name in "read ped".

15-May-2008 (1.00b)

Removed errant decimal point that meant single digit numerical alleles were suddenly not recognized in genotype expressions. Added "file cat".

14-May-2008 (1.00b)

Added in "pnorm" and "qnorm" functions. Multiple imputation now adds in estimate of allelic effect and standard error following Rubin [1987].

08-May-2008 (1.00b)

Bug in imputed genotypes logistic regression, choosing wrong column of data: fixed. Also changed degrees of freedom in tabular output to npar-1 (note this only applies to logistic regression but not to Poisson/Exponential/Weibull). History skipping bottom three commands, since its counter does not include the new header lines: fixed. The *getvar()* function used to access macro variables used an uninitialized variable: fixed. Free format MENDEL 8.0 map and definition files now can be written (add "free" at end of command). Now the default action of "rec <tra>" is to recast from binary to quantitative or vice versa.

01-May-2008 (1.00b)

The value of *maxact* being incorrectly set by "select": fixed.

29-Apr-2008 (1.00b)

"fre snp" output was aligning character alleles too far right (disappearing ;)): fixed. Weibull GLMM giving NaN likelihoods: hopefully trapped. Can escape percent sign within a macro function body.

24-Apr-2008 (1.00b)

Exponential notation no longer has to be quoted to correctly handle a signed exponent on the command line (never affected the reading of data, even inline). Log has datastamp at head.

22-Apr-2008 (1.00b)

Further shuffling of order of evaluation of macros so that one can iterate within loops.

21-Apr-2008 (1.00b)

Checks for legal survival times and Yazdi et al [2002] heritability estimator for Weibull model in "fpm". Sometimes "wri csv" did not print any trait information (if first time "write" called): fixed (again).

15-Apr-2008 (1.00b)

Macro variables can be embedded in strings by delimiting using brackets in the usual fashion eg "%(a)". Evaluation of macro variables in macro functions re-delayed until the function is called. When generating a perfectly informative marker in linkage with a target using "sim", there was linkage disequilibrium induced by the ordering of alleles in genotypes (in nuclear families, for example, even numbered alleles cosegregated with the larger allele of the target). These indicator alleles are now shuffled.

09-Apr-2008 (1.00b)

The "rec" command now allows mapping of multiple values to multiple values. Prettified "tab" output so that numbers that are too large for the current print format do not print as "***", but are truncated or printed as integers as appropriate.

08-Apr-2008 (1.00b)

The "twin" analysis giving strange results when the twin indicator is missing for a pair of siblings -- being scored as DZ even though do not meet criterion (eg "twin trait zyg == 1"). Fixed.

07-Apr-2008 (1.00b)

Minor fixes. The "sim ped" command was not giving requested minimum number of offspring (in the second generation only) -- changed, but note that the second generation of families always have a minimum of one child (rather than zero). The reported number of active pedigrees after an algebraic "select" was not being updated -- fixed. Some prettification of output and updating of documentation and examples.

04-Apr-2008 (1.00b)

Changed matrix inversion routine used to evaluate the likelihood for "varcom" from AS7 to LINPACK's dgedi. This fixes problems analysing MZ twin data when the heritability is very high.

02-Apr-2008 (1.00b)

Compiles again using Intel ifort (needed appropriate declarations of unlink and rename). An automatic print format for "wri", to avoid the Fortran "***" for numbers too large for the specified format (only for the native format at the moment).

01-Apr-2008 (1.00b)

Compiles again using SunStudio sunf95 (needed appropriate declarations of systemic specific routines eg hostnm, time, signal). However, persistent FPE within *zp()* (AS66), which seems to be compiler specific.

31-Mar-2008 (1.00b)

Sib-pair now looks for and "includes" a file "sib-pair.ini" along the full search path, starting in the present working directory and then the work directory (if specified). This does not occur if the "-f" command line option has been used. The "echo" command now evaluates macro variables and respects C style escaped characters: this means percent can be escaped, and "\n" used for a newline and so forth.

28-Mar-2008 (1.00b)

"ass" was broken for X markers by MZ twin code -- fixed.

25-Mar-2008 (1.00b)

Fixed haploid analysis: uninitialized variable meant skipping analysis. Now "recode letter" "flip" etc work for haploid markers.

18-Mar-2008 No. 2 (1.00b)

Fixed up "freq" so that gives allele frequencies for haploid markers. Fixed genotypes in "tab" with *plevel*=0. Better trapping of problems in "mit" eg no trait variation.

18-Mar-2008 (1.00b)

Added haploid marker haplotype association code for quantitative traits.

17-Mar-2008 (1.00b)

famcor() now handles MZ twins cleanly.

16-Mar-2008 (1.00b)

Added handling of MZ twins to "var", "fpm" and all routines relying on MC genotype and IBD simulation except the FBAT routines.

12-Mar-2008 (1.00b)

Fixed annoying bug in "fpm" Weibull analysis due to starting values not being correctly set. Fixed Z-value for "VE(F+R)/VE(F)" (was assuming null of zero). Added "set log" so can set name of logfile .

05-Mar-2008 (1.00b)

Further tinkering with "keeldro whe r2" , so that keep and drop give expected results. Minor output cleanups. Added code so that *wrcsv()* respects locus order from *lorder()*: fixes locus order from "wri men <fil> new".

03-Mar-2008 (1.00b)

"keeldro whe r2" now checks r^2 of the current candidate versus all markers subsetted to date (was checking only the nearest included marker. Segfaulting in some haplotype association jobs where more than two alleles per locus due to error in calculation of scatter matrix indices: fixed. Found longstanding typo in *loglin()* (*par* instead of *pars* in call to *bsub())*: fixed.

28-Feb-2008 (1.00b)

Table of new and old IDs from "uni". Repaired count printed by "anc" (was uninitialized). "keeldro whe r2" allows selection of markers based on disequilibrium r^2 measure.

21-Feb-2008 (1.00b)

Prettification of peeling output.

20-Feb-2008 (1.00b)

Fixed the other iterative peeling bug.

19-Feb-2008 (1.00b)

Fixed some bugs in iterative peeling algorithm (not all done yet ;)). The system "$" command is no longer macro evaluated.

15-Feb-2008 (1.00b)

Quantitative TDT detailed output had format error when no informative individuals: fixed. Added Mendel 8 pedigree format support: "wri men <fil> new".

14-Feb-2008 (1.00b)

MC IBD calculations ignoring inbreeding contributions: fixed.

12-Feb-2008 (1.00b)

Added iterative peeling algorithm for single locus likelihood. The "gpe" command now defaults to this deterministic algorithm (next, recursive IBD).

06-Feb-2008 (1.00b)

Added "more" command. Fixed calculation of maxact after select statement.

05-Feb-2008 (1.00b)

Macro function names now have precedence over hardcoded commands. For example, if macro "plot" exists, the command "plot" will evaluate to the macro, while "plo", "plotter" etc will evaluate as usual to the standard *plot* command. This obviously allows aliasing. Echo hack refined.

31-Jan-2008 (1.00b)

Had broken starting values for "fpm" poisson GLMM with an offset: fixed.

31-Jan-2008 (1.00b)

Single-site Gibbs part of MCMC genotype sampler not updating terminal individuals: fixed. Reduced required workspace for some routines.

30-Jan-2008 (1.00b)

Marker allele frequencies being written to *header.dat* rather than *popln.dat* for "write loc pap": fixed.

29-Jan-2008 (1.00b)

The "edit" command broken for markers: fixed.

25-Jan-2008 (1.00b)

Stuff to improve handling of large pedigrees: improved time taken to read in such pedigrees greatly (had left in an unoptimized ID search *tabid()*); made some print widths for numbers slightly larger; extended length of dummy IDs, so as to accomodate large animal breeding type pedigrees with many unspecified parents; reimplemented MC inbreeding estimation routine "kin inb mc" for very large pedigrees. Fixed up big gap in lines written by "wri loc loki".

17-Jan-2008 (1.00b)

Long IDs in diagram showing Mendelian errors now left truncated; all IDs better centred.

14-Jan-2008 (1.00b)

If a set of contiguous markers are completely linked (defined as map distance less than 0.1 cM), a variance-weighted average IBD is now calculated for the set for "qtl full". A particular set can be specified using "var <tra> aqe <mar1> + <mar2>...". As a result, the list of covariates also need to separated by "+". Fixed interaction between "pack" and "join". The latter now has an "interactive" interface when no names are specified.

07-Jan-2008 (1.00b)

The "dis" command can analyse more than two loci, and can carry out haplotypic association analysis (autosomal and assuming all individuals unphased). Currently, it is too slow for more than five-marker haplotypes.

02-Jan-2008 (1.00b)

The "fil del" and "fil ren" delete and rename files. The Scheme `read-line` function can now read from stdin.

10-Dec-2007 (1.00b)

Combining "order" with "pack" not keeping the correct columns of data: fixed.

24-Nov-2007 (1.00b)

Fixed ID printed when person is own parent (intermittently was a nearby person from the same pedigree instead). Assorted prettification of output. Finished macro (see example script "where.in") that implements subsetting for commands eg "where (male) assoc trait".

20-Nov-2007 (1.00b)

Bug in generation number saved to a variable by "gen": fixed. Added a "rev" modifier keyword, so that generation number can be from the bottom of the pedigree (ie "present" generation) rather than top. More summary output from "reg rep". Streamlined effects of the "set imp" command. "rea loc mer" deals with "S2"

correctly. Prettified output.

19-Nov-2007 (1.00b)

The assignment operator now regenerates starting values for unobserved genotypes, so MCMC procedures run without complaint on copies of marker loci.

16-Nov-2007 (1.00b)

Executable compiled with the Intel Fortran 95 compiler ifort now working correctly: gcc based compilers use Unix escaped characters, so "\" needs to be escaped for g95, but not ifort (affected line continuations and quotation marks in strings).

09-Nov-2007 (1.00b)

F-distribution P-values were from the lower tail: changed to upper tail. Prettified writing of dates.

08-Nov-2007 (1.00b)

Postscript plot of "fpm" loglikelihood trace now automatically generated. Added Scheme string comparison operators, `open-input-file`, `close-input-port`, `read-line`. There seems to be an intermittent garbage collector problem. Documented the "which" command. Added "kbp" modifier for "read map".

05-Nov-2007 (1.00b)

Added documentation for "keep whe tes": this allows retention of marker loci based on the last appropriate association or linkage analysis.

02-Nov-2007 (1.00b)

The "replicates" modifier to the "reg" command gives multiple imputation automatically. The Scheme commands `locord`, `locnotes`, and `loctyp` access locus information (position in list, notes and the locus type). Fixed up handling of zero survival times in Weibull regressions. Using "fixedweibull" as a model in "reg" fixes the shape parameter. Prettyfication of output.

25-Oct-2007 (1.00b)

The "set fre" command allows one to specify the population allele frequencies for a marker for use in (some) MCMC algorithms. The "set ana" command allows inclusion of imputed genotypes in regression analyses -- these will be automatically reimputed each time "reg" is called. The idea of this is to allow multiple imputation association analysis. An alternative is to use the output from the "gpe" command. This gives genotype probability estimates and an expected allele dose, if requested, for a marker. The "set freq" command is respected by "gpe". Locus notes now appear in "fre snp" output. The "reg" Weibull regression occasionally gets stuck estimating the shape parameter. An additional option "sha <val>" allows one to specify a better starting value, if necessary. The "sum" command summarizes P-values from the "dis" command, which may occasionally be useful, mainly in the search for tagging SNPs for a given target SNP. The "set ple" message can be blocked by adding the "quiet" modifier (for use in scripts).

19-Sep-2007 (1.00b)

For Merlin pedigrees, zygosity codes sometimes were "*" where they should be zero: fixed.

17-Sep-2007 (1.00b)

Write Beagle format unphased genotype data files.

24-Aug-2007 (1.00b)

"write mendel" did not know about MERLIN coding of multiple sets of twins: fixed.

20-Aug-2007 (1.00b)

Added another multipoint homozygosity statistics to "mul" output. Added "<-all <mar>" to save a list of alleles for a marker.

17-Aug-2007 (1.00b)

Added "<-ls" to save state of active locus list. Macros now allow statements to extend over multiple lines.

16-Aug-2007 (1.00b)

Macro variables can save the state of the "bur", "che", "epo", "imp", "ite", "min", "ple", "pwd", "sex" and "twi" settings, using "mac <macro_var> <- <*setting*>. The "unselect" statement can now roll back in a stepwise fashion through multiple "select" statements (previously it would undo all the "select"s at once). The "string-split" Scheme command splits a string into a list of words.

10-Aug-2007 (1.00b)

Fixed dummy allelic coding of missing data for (first) marker included as covariate for "var" and "fpm". These were being filled in with the allele frequency rather than twice the allele frequency.

06-Aug-2007 (1.00b)

"mzt" did not honour MERLIN coding of multiple sets of twins: fixed.

03-Aug-2007 (1.00b)

Now can "undrop where" based on annotation search.

02-Aug-2007 (1.00b)

MERLIN coding of multiple sets of twins now correctly input and output. Introduced "even" and "odd" trait comparison operators for this purpose.

31-Jul-2007 (1.00b)

Minor Scheme memory allocation fixed. Added Scheme multilist map.

30-Jul-2007 (1.00b)

Added "sum" to summarize results of statistical tests of linkage or association.

25-Jul-2007 (1.00b)

Added minimal Scheme interpreter ("eval"). Main visible effect of this is that unlimited macro variables and functions can be declared. These variables are stored as strings in the Scheme environment. From Scheme,

Sib-pair commands can be passed back for execution using "`(run <string>)`".

12-Jul-2007 (1.00b)

The ":" operator can be used to glue sequences of expressions (cf ";" which does the same for commands). It is automatically inserted between sets of bracket-enclosed expressions. This allows block if statements for variable manipulation. Nested if statements cleaned up.

25-Jun-2007 (1.00b)

Deleting records by "famnum" did not work -- fixed.

20-Jun-2007 (1.00b)

Postscript plotting added.

19-Jun-2007 (1.00b)

Fixed recently introduced bug in handling of covariates by "fpm". Added "log10" function.

15-Jun-2007 (1.00b)

Fixed "show ids" output (did not show first pedigree ID). Smaller "show" output.

14-Jun-2007 (1.00b)

"write pap" reintroduced. "set liab" added (automates inclusion of a liability class for Linkage format pedigree and locus files).

13-Jun-2007 (1.00b)

Fixed problem in manipulating data ("pack") due to addition of a MZ twin pointer to the data structure (packer() writes this to a scratch file). Fixed problems with "join" which would sometimes baulk if the same individual in two pedigrees to join had parents listed.

01-Jun-2007 (1.00b)

Fixed problem in measuring longest pedigree file line length (missed counting some whitespace).

21-May-2007 (1.00b)

More output formatting fixups.

18-May-2007 (1.00b)

More column widths for locus names in brief output now 14 characters wide. "hea loc" and "hea map".

17-May-2007 (1.00b)

Added "%" to evaluate a macro (variable) within a command. Gu Zhu pointed out SNPs with a minor allele frequency of 0.5 get printed twice in the brief output of "fre". Column width for locus names in all brief output now 14 characters wide.

11-May-2007 (1.00b)

Overall multilocus heterozygosities and F statistics now calculated. Added "sex" modifier to "rea cas".

08-May-2007 (1.00b)

Each individual's multilocus heterozygosity estimated by "mul" if a trait is not specified. HWE for X-linked markers works again. The writing of Eclipse format locus files reinstalled (was already documented in help). Some cleanup of output for categorical data association.

02-May-2007 (1.00b)

Added routine to point out IDs involved in impossible loops eg own grandfather. Braces "{" now "echo"ed to output without being evaluated as marking a macro loop. Some prettification (including adding the map position to the output from "sib"). Fixed bug in *tabmat()* -- array size would not expand as required.

27-Apr-2007 (1.00b)

Minor fixes to "wri mor" and "let".

21-Apr-2007 (1.00b)

Fixed newly introduced bug in regression involving markers. Fixed old bug in marker covariate handling in "var".

20-Apr-2007 (1.00b)

Fixed newly introduced bug in printing genotype data with "wri". And another one printing the header. Added code to test the maximum record length in the pedigree file and create an appropriate size buffer for reading it.

19-Apr-2007 (1.00b)

Fixed handling of dropped loci in expressions (these were supposed to be temporarily undropped if named in an expression, but this was broken). Fixed minor problem in "test sex" (too much output).

18-Apr-2007 (1.00b)

Fixed "famnum" and "index" for operations other than "select" and "count" -- could not be used to create new variables etc.

17-Apr-2007 (1.00b)

Added "julian" and "greg" functions. Some tinkering with "lif".

14-Apr-2007 (1.00b)

Fixed bug in multinomial "assoc" gene-dropping P-values (table not being reset for each simulation).

13-Apr-2007 (1.00b)

The "fstat" command writes F-statistics to a table. The "assoc" command can now perform multinomial analysis ("cat" modifier). The "out" command now appends rather than overwrites if the target file already exists. Some output prettification.

05-Apr-2007 (1.00b)

The "keep where position" command allows subsetting of loci by map position. The "whi" command tells you the index number for a locus -- most Sib-pair commands allow this to be used instead of the name of the locus, but there was no utility to match up names to numbers. Chan Hee Park found a problem reading in tab delimited files under Windows, which is now hopefully fixed.

03-Apr-2007 (1.00b)

Added in "llm" command. Converted all references to "und" to use the full name "undrop" rather than "undelete", as is this is more descriptive. Removed "%" as an alias for "$". Some column widths in output tables increased to cater for the age of GWAS.

28-Mar-2007 (1.00b)

Added in "read cases". The "read ppd" fixed -- was reading wrong columns for sex etc (code not brought over correctly from F77 version).

26-Mar-2007 (1.00b)

Mendelian error output now clearer for X-linked markers. Extended "lif" command.

22-Mar-2007 (1.00b)

Documentation updated. Added "lif" command. Fixed output from "get all".

15-Mar-2007 (1.00b)

Gu Zhu pointed out that "wri hap" did not work if the alleles were already numeric (set them to missing). Now alleles other than A,C,G,T pass through unchanged.

09-Mar-2007 (1.00b)

Various bug fixes: "he1" had stopped working, and the addition of a haploid marker class had broken some parts of "drop", "keep" and "show".

06-Mar-2007 (1.00b)

Added some routines for handling of haploid data (aimed at mitochondrial or Y haplotype type data), the most useful being "test hap", but "tab $h" will list and count the haplotypes, "reg", "var" etc will allow a test of association. Expanded some help text to aid searches. Added option to read command line arguments at startup ("--help", "--include", or a quoted Sib-pair command eg "nsp95 '1+1'). I doubt the latter will see much use, but may be of assistance for shell scripting. Added some missing documentation for the "set twin" command -- thanks to Chris Oldmeadow for pointing this out.

01-Mar-2007 (1.00a)

Added toggle ("set fba") to turn off multigenerational imputation for FBAT procedure. This then gives the same behaviour as the FBAT and PBAT programs. Added the count of active pedigrees to the "wri" output preamble.

23-Feb-2007 (1.00a)

Fixed another bug in interaction between loops and macros.

20-Feb-2007 (1.00a)

The "get" command summarizes trait values in relatives of a given degree of relationship with ego. The "get all sample <trait> <newtrait>" option will generate a permutation sample of a trait.

16-Feb-2007 (1.00a)

The "test dob" command checks parent-offspring ages or DOBs for consistency. The history is no longer diverted by the "out" command. Documented the "test sex" command.

13-Feb-2007 (1.00a)

Bug in loops fixed. Windows text based file picker cleaned up.

12-Feb-2007 (1.00a)

Penrose sib-pair linkage analysis implemented.

02-Feb-2007 (1.00a)

The "out" command allows redirection of output from the screen to a file. The GUI is activated by "set gui on" -- if off, a text-based file and directory browser is called instead.

31-Jan-2007 (1.00a)

Fixed up error in "fpm" MC estimate of parameter standard errors (again!) -- the wrong denominator was giving too large values. Macros being run by an *include*d script no longer finish prematurely. The body of a macro no longer appears in the history or logfile.

The Windows (and even linux, if desired) version now can call a graphical file browser, if the file name is not specified on the command line (eg "inc"). This is the Java AWT file dialogue called via the crossplatform/language JAPI library. This is slower on older computers due to a delay in starting up the Java runtime. On linux, the java processes ("japi kernel") seem to persist sleeping. Obviously, the Java runtime must be present on your computer, along with the older AWT library.

12-Jan-2007 (1.00a)

Prettified "show macro" output. The "twin" command does binary as well as quantitative traits. One level of nesting of "includes" now allowed.

08-Jan-2007 (1.00a)

Added "twin" command to perform simple analyses of a classical twin design. Added "wri fba" as a synonym for "wri asp".

02-Jan-2007 (1.00a)

Added "ibd" to the "assoc" command: this activates gene-dropping conditional on *ibd* at another marker (presumably more informative for *ibd* in that region). Incidentally, improved results from multiple calls to *sim* when conditional on *ibd* at a marker, and implemented updating of the starting values for unobserved

genotypes whenever *drop()* has been used.

21-Dec-2006 (1.00a)

Added the "seg" command. Added implicit loops (the "{" command).

11-Dec-2006 (1.00a)

"wri loc super" added.

08-Dec-2006 (1.00a)

"wri loc morgan" was writing imputed genotypes to the ".par" file -- fixed.

04-Dec-2006 (1.00a)

The "sim ped" command added.

29-Nov-2006 (1.00a)

The "ito" command added.

28-Nov-2006 (1.00a)

The "%+N" macro variable added. The "write loc rel" command will change the chromosome based on locus name if it is a "D" name (otherwise it uses the map positions *mod* 1000 eg positions written as 22xxx are interpreted as xxx cM on chromosome 22).

24-Nov-2006 (nsp95)

"rank" and "blom" now reimplemented.

23-Nov-2006 (nsp95)

Re-repaired weighting of contributions of multiple chains/replicates in "fpm". The "test" command now documented.

21-Nov-2006 (nsp95)

Repaired weighting of contributions of multiple chains/replicates in "fpm". The "pwd <newdir>" command now successfully changes directory. "rea mer" added as an alias for "rea ped".

16-Nov-2006 (nsp95)

Added "%%" and "%0" macro variables. Fixed Solar pedigree file where missing sex is not allowed (defaults to "f"), and in phenotype files where dichotomous trait need to be numeric coded.

15-Nov-2006 (nsp95)

Added "dec loc" to automate declaration of loci. Expanded online help including an example.

06-Nov-2006 (nsp95)

Added "wri mer" command to automatically write an MZ indicator variable. The "wri phe" command writes a white-space delimited phenotype file for FBAT. The "help" command output for the "wri" commands has been expanded, so there is one line describing each type of output.

19-Oct-2006 (nsp95)

Fixed bug in MZ indicator variable written for MENDEL and SOLAR -- wrong column read if marker locus precedes or reordering performed. Imputation of missing ages or birth years performed within families by "imp <age>".

18-Oct-2006 (nsp95)

Dale Nyholt pointed out that the MZ twin indicator was not being included in SOLAR pedigree files -- this and a household indicator (pedigree number) are now written. Another of his suggestions: MERLIN mapfile chromosome number is now appended to the notes as "(chrN)". Profile likelihood confidence bounds are now given for V_A in the output from "var". Added in first attempt at WLS analysis of familial correlations in output from "des". The BLUPs written by "fpm" will be for the pedigree environmental or maternal effects if an additive genetic component was not included in the model. Some speed increases obtained by optimizing reading in pedigrees (but still too slow).

12-Oct-2006 (nsp95)

Fixed bug in "qtl" -- lod score was always zero for "full" model. Started a WLS heritability analysis of familial correlations in "des".

10-Oct-2006 (nsp95)

Added a "sim" modifier to "reg", which gives the gene-dropping P-value for the first marker locus in the formula. This allows genetic association survival analysis etc. Bug squashing, notably in "var" where "CE" model bombing intermittently, and in macro facility, which now nests calls sensibly.

03-Oct-2006 (nsp95)

Added simple macro facilities ("mac", "mac <nam> del" and ", "sho mac"). To make this easier, ";" can now be used to separate multiple commands on the same line.

29-Sep-2006 (nsp95)

Added the "join" and "show ids" commands. The "wri hap" and "wri loc hap" commands ease writing for haploview (recoding nucleotide letter codes to "1..4" automatically). Imputation, checking and genotype starting value generation streamlined: always runs Lange-Goradia exclusion algorithm to test errors, and always uses gene-dropping with rejection to generate the starting genotypes: *imp* can be -1 (nil), 0 (starting values generated), 1 (unequivocal imputation), 2 (see starting values for genotypes). Almost all commands back in and working ("hap" one exception).

31-Aug-2006 (nsp95)

Port to Fortran 95 nearing completion. New features include:

- Testing of sex assignment using sex-linked markers including Amelogenin.
- Summary of Mendelian errors by pedigree and by locus (similar to that generated by the *taberrors* shell script).

- Nicer summary of monozygotic twin assignment test results.
- Penalised nonparametric ML estimation of MCMC posterior modes.
- Another quantitative trait TDT, following the simplified regression model formulation of Gauderman et al 2003.
- Recoding nucleotide letter codes to/from numerical code.
- Information about dataset memory utilisation.

07-Jul-2006 (1.00a17)

Fix documentation of "wri csv".

06-Jul-2006 (1.00a17)

Actually included the reading of zygosity indicator trait values of "MZ" in the publically available code. Fixed processing of twinning indicator in "wri mendel" (reading wrong column).

27-Jun-2006 (1.00a17)

Added documentation for "rea ppd" and "rea loc mer xli".

22-Jun-2006 (1.00a17)

Smart truncation of locus names fixed for more than 26 collisions.

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" and allowed addition of a genotype to the diagram (within the ideogram). Added in "marcom" function to count up maximum number of typed markers shared between ego and 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 undrop 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 "undrop" 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 "undrop" 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/undrop 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).

Appendix: Embedded Scheme Commands

The Scheme implemented within Sib-pair supports only a subset of the language. The only atomic data types are booleans, (long) integer and double precision real numbers, and strings. Characters (and character procedures) are handled as strings of length one.

<	Numerically less than
<=	Numerically less than or equal
=	Numerically equal
>	Numerically greater than
>=	Numerically greater than or equal to
-	Integer subtraction
/	Integer division
*	Integer multiplication
+	Integer addition
abs	Absolute value
and	Logical and
append	Append to end of list
apply	Apply function to list
apropos	List matching functions
assoc	Find matching pair in an "alist" (list of paired values) using **equal?**
assq	Find matching pair in an "alist" using **eq?**
assv	Find matching pair in an "alist" using **eqv?**
atom?	Is a simple number or string?
begin	Start a block
boolean?	Is a boolean?
car	First element of pair or list
cdr	Remainder of pair or list
caar	First element of first element of pair or list
cadr	Und so weiter
cdar	
cddr	
caaar	
cdddr	
caadr	
cadar	
caddr	
cdaar	
cdadr	
cddar	
call/cc	
call-with-current-continuation	
case	Conditional branching
char<?	Is char lexicographically less?
char>?	Is char lexicographically greater?

char<=?	Is char lexicographically less than or equal?
char>=?	Is char lexicographically greater than or equal?
char-alphabetic?	Is char alphabetic?
char-ci<?	Is case-insensitive char lexicographically less?
char-ci>?	Is case-insensitive char lexicographically greater?
char-ci<=?	Is case-insensitive char lexicographically less than or equal?
char-ci>=?	Is case-insensitive char lexicographically greater than or equal?
char-numeric?	Is char a digit?
char-whitespace?	Is char a tab, space, CR, NL, HR?
char->integer	Return ASCII code
char-downcase	Return lower case
char-upcase	Return upper case
closure?	Is a closure
cond	Conditional branching
cons	Create a Lisp pair of elements
cons-stream	
current-second	Return seconds since epoch
define	Define a variable or function
delay	
display	Print a variable or function
do	Looping construct
else	
eq?	
equal?	
eqv?	
error	Print an error message
eval	Evaluate a *SEXPR* (Scheme expression)
even?	Test if even
exact->inexact	Cast to real
exit	Exit Sib-pair completely
expt	Integer exponentiation
force	
for-each	Like map, but used for side-effects only
format	Print formatted data
gc	Force a garbage collection
gcd	Greatest common divisor
gensym	Create a new unique symbol
get-closure-code	
help	Minimal information about Scheme
if	Conditional branching
inexact->exact	Cast to integer
integer?	Is an integer
integer->char	Return character for given ASCII code
lambda	Function
lcm	Least common multiple

length	Gives length of a list
let	Declare a local variable
let*	
letrec	
list	Create a list
list?	Test if a list?
list-ref	Gives Nth element of a list
list-tail	Drops the first k elements of a list
ls	List Sib-pair variables
macro	
make-list	Create a list of given length
make-string	Create a string of given length
map	Apply a function to each element of a list in turn
max	Maximum of arguments
member	Test if present in a list
memq	Test if present in a list using equal?
memv	Test if present in a list using eqv?
min	Minimum of arguments
modulo	Give modulo
negative?	Test if negative integer
newline	Print a newline
new-segment	
not	Logical negation
null?	Test if a null
number?	Test if an integer number
number->string	Convert from number to string
odd?	Test if odd
or	Logical or
pair?	Test if a Lisp pair
peek-char	Read one character without advancing
positive?	Test if a positive integer
print-width	
procedure?	Is a function?
quasiquote	Allow some of quoted expression to evaluate
quit	Leave the Scheme interpreter
quote	Quoted expression left unevaluated
quotient	Integer division
random	A random integer from U(1, N)
read	
read-char	Read one character from standard input or file
read-line	Read one line of input from standard input or file
remainder	Remainder
reverse	Reverse ordering of a list
run	Run a Sib-pair command
set!	Set the contents of an existing variable

set-car!	Set contents of first element of pair or list
set-cdr!	Set contents of rest of pair or list
sqrt	Integer square root
string?	Is argument a string?
string=?	Are strings equal?
substring?	Is string 1 a substring of string 2?
string<?	Is string lexicographically less?
string>?	Is string lexicographically greater?
string<=?	Is string lexicographically less than or equal?
string>=?	Is string lexicographically greater than or equal?
string-append	Concatenate strings
string-length	Length of argument string
string-ref	Get kth character from string
string-set!	Set the value of a substring
string-split	Split a string into a list of words
string->number	Convert from a string to a number
string->symbol	Convert from a string to a symbol
substring	Get the value of a substring
symbol?	Is a symbol?
symbol->string	Convert from a symbol to a string
system	Passes commands to the operating system
unquote	
unquote-splicing	
write	
zero?	Test if equal to zero

There are a number of Sib-pair specific builtin commands for example allowing access to locus data:

apropos <str>	lists commands containing that string.
help	Information about this Scheme implementation.
isatty? <fil>	test if interactive session.
file-exists? <fil>	test file.
file-list <nam>	list contents of a directory.
file-delete <fil>	delete a file.
open-input-file <fil>	open a port.
close-input-port <port>	close a port.
read-line [<port>]	reads in next line from stdin or open file.
format <port> "{~[:num:][@][ASD~%]}" <args>...	formatted output.
string-split <str> [<sep>]	splits string on white space or optional char.
substring? <sub> <str>	returns start of substring in string.
regexp <sstring> <str>	test if search regular expession matches string.
system <cmd>	passes command to shell.
getenv <nam>	returns value of environment variable.
date	returns current date and time.
time	returns current time as number of seconds since epoch.

seq <sta> <fin> [<step>]	generate sequence.
filter <test> <list>	filter contents of list
environment-bound? <name>	test if symbol bound to a variable

version	prints Sib-pair version.
run <cmd> ...	runs a Sib-pair command.
pass-command <cmd> ...	stores Sib-pair commands to the buffer for evaluation once you return to the usual Sib-pair prompt.

Sib-pair locus dataset accessors:

ls [<typ>]	creates a list of locus names (of given type "adhmqx".
nloci [<typ>]<	returns total number of loci.
loc <index>	returns locus at that position in the locus list.
loc-set! <idx> <name>	set name of locus at that position in the locus list.
lochash-update!	update hash of locus names
locord <loc>	returns position of a locus in the locus list.
locnotes [<loc>..<loc>]	returns notes for a locus.
locnotes-set! <loc> <str>	rewrites notes for a locus.
loctyp [<loc>..<loc>]	evaluates type of a locus ("adhmqx").
loctyp-set! <loc> <typ><	set type of a locus ("adhmqx").
map-position [<loc>..<loc>]	returns map position for locus.
map-position-set! <loc> <pos>	sets map position for locus.
chromosome [<loc>..<loc>]	returns locus chromosome.
chromosome-set! <loc> <chr>	sets locus chromosome.
locstat [<loc>..<loc>]	returns last P-value for a locus.
locstat-init! <title>	initializes all locus P-values.
locstat-set! <loc> <value>	sets P-value for a locus.
locrank <loc>	returns rank of locus test statistic.
stat-result ["pval|lik|npars|lrt|df|stat|var"]	returns result of last model.

Sib-pair phenotype dataset accessors:

nobs	number of individual records.
npeds	number of pedigrees.
nactpeds	number of active pedigrees
active-status	activity status of pedigrees
set-active-status! <idx> <lev>	set activity status of pedigree
active-pedigrees [<idx>...]	list of active pedigree names
pedigrees [<idx>...]	list of pedigree names.
pedigree-size [<idx>...]	size of ith pedigree.
pedigree-members [<idx>...]	list of indices of pedigree members
individual-pedigree [<idx>...])	give pedigree ID for index
individual-name [<idx>...]	give ID for index.
set-pedigree-name! "<ped>" <newname>	Set pedigree ID string .
set-individual-name! <idx>|"<id>" |("<ped>" "<id>") <newname>	Set ID string for individual.
individual-index [<id>|<idx>...]	give index for ID.
insert-record! <idx>|(<ped> <id>) ['after]	insert data row before index.

father [<idx>...]	father indices.
mother [<idx>...]	mother indices.
imztwin [<idx>...]	MZ twin pointer.
sex [<idx>...]	sex value for individual(s).
set-sex! [<idx>...] <sex>	set sex value for individual.
data <loc> [<idx>]	phenotype for individual(s).
set-data! <idx>|"<id>" |("<ped>" "<id>") <loc> <val>	Set phenotype for individual.
data-counts <loc> [<loc>]	tabulation of phenotype levels.
allele-freqs <loc>	tabulation of alleles.

There are a number of builtin statistical functions (20100323 onwards), calling the appropriate Fortran routines.

pnorm <Z>	Gaussian upper tail probability
qnorm <p>	Gaussian quantile for given upper tail probability
pchisq <X^2> <df> [<ncp>]	Central and non central chi-square upper tail probability
qchisq <p> <df>	Chi-square quantile for given upper tail probability
fp <F> <df_1> <df_2>	F upper tail probability
pgamma <X> <df>	Incomplete gamma integral
lgamma <X>	Log gamma
dbeta <X> <shape1> <shape2>	Density of Beta distribution.
bivnor <Z_1> <Z_2> <r>	Bivariate Gaussian upper tail probability
pmvnorm <p> <dir> <cor> ['genz]	Multivariate Gaussian tail probability
pchisqsum <X> <alpha>	Upper tail probability for distributions of quadratic forms in normal variables

And a few statistical data manipulation routines, that follow the XLispStat equivalents:

sort <list>	Sort a list of numbers
order <list>	Indices of order of numbers in list
rank <list>	Rank of numbers in list
quantile <list> <p>	quantile(s) of list
sample-seq <N> <size> ['replace]	sample a sequence 1..N
sample <list> <size> ['replace]	sample a list
differences <list>	Lag-1 differences in list
which <list>	Indices of non-null elements of list
unique <list>	Indices of unique elements of list
duplicated <list>	Indices of repeat elements of list
intersect <list> <list>	Common elements of lists
setdiff <list> <list>	Elements unique to first list
union <list> <list>	Elements of both lists
list-select <list> <idx-list>	Sublist using indices
stats <list>	N, missing, mean, variance, min, max of list.

Sib-pair Scheme also now (2010-05-9) contains builtin functions to call the JAPI (java application programming interface) library (http://www.japi.de), which allows building GUIs etc. JAPI allows Fortran code to interface the Java AWT (Abstract Windowing Toolkit). The JAPI website details the commands.

j_start	Start JAPI listener

j_quit	Stop JAPI listener
j_frame	Create a frame
j_panel	Create a panel
j_borderpanel	Create a borderpanel
j_dialog	Create a dialogue
j_button	Create a button
j_radiobutton	Create a radiobutton
j_radiogroup	Create a radiogroup
j_checkbox	Create a checkbox
j_list	Create a list
j_add	Add an object to a container
j_setcolor	Set the foreground colour
j_setcolorbg	Set the background colour
j_setnamedcolorbg	Set the background colour
j_getselect	Return index of selected item
j_select	Select an item
j_deselect	Deselect an item
j_fileselect	A file dialog
j_filedialog	A file dialog
j_enable	Activate an object
j_disable	Deactivate an object
j_additem	Add an item to a list
j_seperator	Draw a separator
j_textfield	Create a textfield object
j_textarea	Create a textarea
j_setborderpos	Place an object using borderlayout
j_setrows	Set number of rows eg textarea
j_setcolumns	Set number of columns eg textarea
j_getrows	Get number of rows eg textarea
j_getcolumns	Get number of columns eg textarea
j_getlength	Get length in pixels of object
j_getselstart	Start of selected text
j_getselend	End of selected text
j_selecttext	Select text
j_gettext	Get text from label or list item
j_getseltext	Get selected text
j_getitem	Get a list item
j_label	Create a label
j_getcurpos	Get current cursor position
j_setcurpos	Set current cursor position
j_setfont	Set font
j_settext	Set text in object
j_inserttext	Insert text at position in eg textarea
j_replacetext	Replace text at position in eg textarea
j_delete	Delete text

j_dispose	Free resource
j_menubar	Create a menubar
j_menu	Create a menu
j_menuitem	Create an item for a menu
j_pack	Pack layout of frame or panel using layout manager
j_show	Make object visible
j_hide	Hide object
j_keylistener	Listen for key stroke if object active
j_getkeycode	Returns pressed key code
j_getkeychar	Returns pressed key ASCII code
j_mouselistener	Listen for mouse activation
j_getmousebutton	Return pressed mouse button
j_nextaction	Return next action of any object
j_getwidth	Get width of object (pixels)
j_getheight	Get height of object (pixels)
j_getpos	Get position of object (X.Y)
j_setpos	Set position of object (X,Y)
j_setsize	Get size of object (pixels)
j_setalign	Get layout alignment for object
j_setborderlayout	Layout manager
j_setgridlayout	Layout manager
j_setflowlayout	Layout manager

Sib-pair Scheme (2010-11-29) also contains builtin functions to call the EGGX/ProCALL graphical library (http://www.ir.isas.jaxa.jp/~cyamauch/eggx_procall/), which allows plotting and building simple GUIs etc under X Windows.

ggetdisplayinfo	Get X display info. Returns a list containing *ndepth* (8,16,24 bit), *nrwidth* (screen width), *nrheight* (screen height).
gopen	Open a window. Takes two optional arguments *nxwidth* (width of window in pixels) and *nywidth* (height of window in pixels). Returns the handle (integer window index) for the opened window.
gclose	Close the specified window. Takes one argument, the window handle.
gcloseall	Close all active windows. Takes no arguments.
newcoordinate	Set up new new coordinates system
newwindow	Change the coordinates system for a window. Takes five arguments: handle, x0, y0 (bottom left), x1, y1 (top right).
layer	Set drawing and display layers (0-7 per window). Takes three arguments: handle, index of display layer, index of drawing layer
copylayer	Copy one layer to another. Takes three arguments: handle, index of source layer, index of destination layer.
gsetbgcolor	Set background colour. Takes two arguments: handle, name of colour to set to (string).
gclr	Clear the specified window. Takes one argument: handle.
tclr	Clear the current terminal window
newpencolor	Change pen colour (0-14). Takes two arguments: handle, number (0-14) or colour name (black, white, red, green, blue cyan, magenta, yellow, dimgray, gray, darkred, darkgreen, darkblue, darkcyan, darkmagenta, darkyellow).
newcolor	

	Specify a new colour (X name). Takes two arguments: handle, name of colour to set to (string).
newrgbcolor	Specify a new colour (RGB values). Takes four arguments: handle, R, G, B.
newlinewidth	Specify line width. Takes two arguments: handle, line width.
newlinestyle	Specify line style. Takes two arguments: handle, line style (0=solid, 1=dotted).
pset	Draw a point. Takes three arguments: handle, xcoordinate, ycoordinate.
drawline	Draw a line. Takes five arguments: handle, x0, y0, x1, y1.
moveto	Move pen to point. Takes three arguments: handle, xcoordinate, ycoordinate.
lineto	Draw line from current pen location to point Takes three arguments: handle, xcoordinate, ycoordinate.
drawrect	Draw a rectangle. Takes five arguments: handle, x, y, width, height.
fillrect	Fill a rectangle. Takes five arguments: handle, x, y, width, height.
drawcirc	Draw a circle (ellipse). Takes five arguments: handle, x, y, xradius, yradius.
fillcirc	Fill a circle Takes five arguments: handle, x, y, xradius, yradius.
drawsym	Draw a symbol Takes three to five arguments: handle, x, y, optional type (1-10), optional size.
drawstr	Print a string on the window at specified position Takes four to five arguments: handle, x, y, string, optional size.
drawnum	Print a number on the window at specified position Takes four to six arguments: handle, x, y, string, optional size, optional number of digits.
gsetnonblock	Set event handling to be nonblocking
ggetch	Get a character from the keyboard if focus on window
ggetevent	Get a keyboard or mouse event for window
ggetxpress	Get a keyboard or mouse event for window