Some new analytic procedures in the

Sib-pair statistical genetics package, 2011

David L Duffy
Genetic Epidemiology Laboratory
Queensland Institute of Medical Research
http://www.qimr.edu.au/davidD#sib-pair

Sib-pair is an extensible software package for genetic data manipulation and analysis. It provides interactive access to a large number of standard analyses, as well as some methodological novelties. In the following, I will give an overview of the program, and then describe some areas of recent development: extrapolation of extreme Monte Carlo P-values, the use of the grouped jackknife for intraclass correlations in pedigree data, a multinomial version of the *WQLS* association test, and Markov Chain Monte Carlo for the fitting of Generalized Linear Mixed Models to pedigree data.

The first code for Sib-pair was written in 1995. It is all standard Fortran 95, and compiles using multiple compilers on multiple platforms, including mobile phones. Creeping featurism has continued to date (71000 lines of code) .

- Simple interpreted language, over 200 commands
- Embedded Lisp (Scheme) interpreter
- Commands for linkage, association, variance components, segregation ...
- Offers the usual record-wise operations on data -- algebra, logical conditions
- Family-centric data operations -- subsetting, pruning etc
- Some elementary databasing type operations – querying, merging, editing
- Flexible data export and scripting to use other programs

Currently, there are "canned" procedures to write out data files (pedigree, locus, map data) for:

Arlequin, Aspex, Beagle, Cri-map, Dot, Eclipse, FBAT, FISHER, GAS, GDA, GDT, Genehunter, Haploview, Linkage, Loki, MENDEL, Merlin, MIM, Morgan, MQLS, Pap, PLINK, RAM, RELPAIR, SAGE, SAS, Simwalk, SOLAR, Structure, Superlink, Wombat.

Sib-pair can read data files in the formats used by:

GAS, HapMap, Linkage (pre and post), MERLIN, PLINK.

Sample data manipulations that can be done easily in Sib-pair:

- Collect summary statistics for a specific class of relatives of *ego* into a new variable
- Extract an optimal set of unrelated cases and controls from a set of related individuals
- Select pedigrees containing a specified number of probands meeting multiple criteria
- Automated testing (and imputation) of pedigree member birthdates or ages
- Generate new IDs
- Simulate genetic (pedigree) data.

```
> get sibling mean height newheight
> let use=(numtyp > 0.95 and height > 1.85); casecon use
> select containing 3 where isfou and height > 1.85
> test dob
> impute age
> unique_ids
```

```
> simulate pedigrees 100 3 4 5; run
> set locus trait affection
> set prevalence 0.05
> simulate trait 0.5
> describe trait
> mft trait ae
```

Monte Carlo P-values in Sib-pair

Geneticists are great consumers of Monte Carlo P-values (17% of the first 500 Google Scholar results for "Monte Carlo P-values", accessed 2011-03-29). The key advantage of Monte Carlo based significance testing [Barnard 1963] is that it is an easily implemented approach to very many statistical problems where analysis is intractable, or at least, too difficult for the present writer. The classical genetic situations it commonly finds use in involve testing hypotheses on correlated data, and multiple testing of correlated hypotheses.

$$p_{MC} = \frac{1}{B}\sum_{i=1}^{B}[T_i \geq t] \quad ,$$

where T_i is the value of the test statistic calculated for the ith of B datasets simulated under the null hypothesis, eg $E(T)=0$, and t is the test statistic value calculated for observed dataset.

The use of automatic digital computers makes this type of procedure reasonably quick, but quantities of data are expanding more quickly in this era of whole genome data. Sequential Monte-Carlo significance testing is one method to minimize the amount of calculation [Besag and Clifford 1991], which can be shown to always be more efficient than the usual fixed-size approach [Silva et al 2008]. This approach extends the standard sequential testing procedure (eg Wald 1945) to simulation under the null hypothesis, and has been used in Sib-pair since 1996.

However, in the multiple testing situation we are often interested in estimating very small P-values accurately. The minimum possible magnitude of a Monte Carlo P-value estimate is:

$$1/(1+B) \ ,$$

where B is the number of simulated samples, and the Monte Carlo error around this value is binomial. When the critical test threshold α is set to, say, 5×10^{-8}, the power to reject the null hypothesis is far less than the appropriate analytic test, unless B is large. The *resampling risk* is the probability that repeating the Monte Carlo test would reach a different conclusion as to whether a result was significant at the given α, and the above definition of a Monte Carlo P controls this resampling risk.

In the case of genome-wide data with many "significant" tests, this can become tediously slow.

Approximating the cumulative distribution function of the test statistic

In the above setup, we have carried out a lot of simulations, and one would think would give a lot of information about the shape of the null distribution of the test statistic. It would seem a more effective use of these data to carry out some type of curve fitting, and estimate the quantiles from this model: a higher-order approximation to the tail area. We can then assign P-values much smaller than 1/(*B*+1) to extreme observed values. Another advantage of this approach is in the multiple testing situation, where we can combine information about the *cdf* from all the tests.

Approximating the tail of the cdf *using extreme value theory*

Several authors have pointed out that we are not particularly interested in the shape of most of the distribution. The extreme tail of most distributions tend to resemble one another, and exceedances over a threshold fall into two families of the *extreme value distribution*, Exponential or Pareto, depending on the finiteness of the *index of regular variation*, *a*. Hill [1975] suggested a simple estimator of *a* based only on a set of the highest order statistics for the sample from that distribution,

$$a(n/m) = m^{-1} \sum_{i=1}^{m} [\log(X_{(i)}) - \log(X_{(m+1)})] \ ,$$

where $X_{(i)}$ is the *i*'th order statistic, *n* is the total sample size, and *m* is the number of order statistics.

With a finite estimate of *a* in hand, then Hill [1975], and Davis and Resnick [1984] suggested estimating the tail probability as (per the Pareto),

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

Davis and Resnick [1984] show this estimate to be strongly consistent, and put bounds on this estimate. The variance of the estimate of $a(n/m) \sim 1/m$, but for optimal behaviour *m/n* should approach zero.

I am unaware of any previous application to a Monte Carlo type significance testing setup. It is used in Sib-pair to augment the sequential testing algorithm used for Monte Carlo P-values in a number of analyses.

Fig 1. Performance of Davis-Resnick estimator versus *B* (constant *m*) for normally distributed test statistic. Black box-plots represents D-R estimator (*m*=10 highest statistics), upper blue boxes the naïve MC procedure, and solid line the asymptotic P-value for a true Z={4,5,6,7,8,9,10}. Panels represent results for 100, 1000, 10000, 100000 pseudosamples (1000 samples per condition).

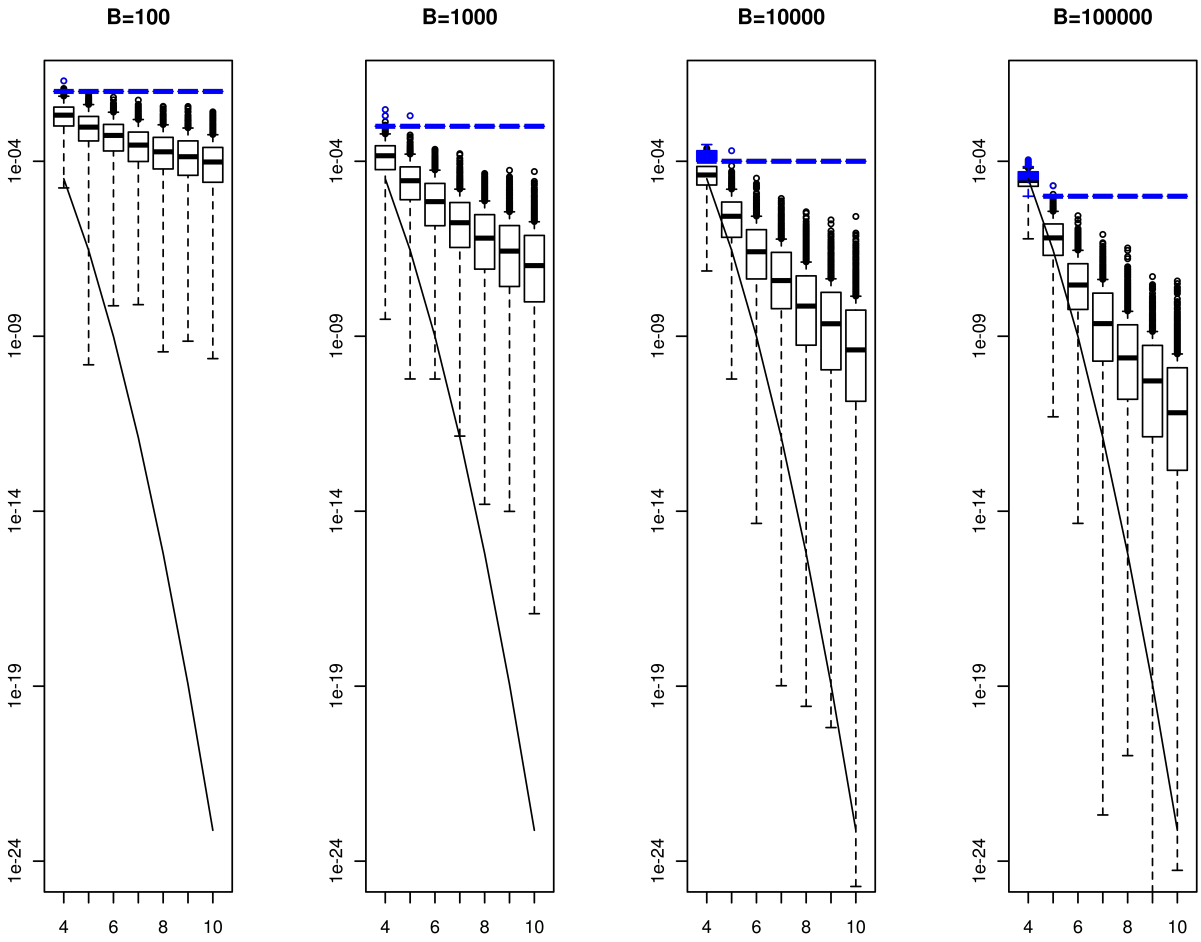


Fig 2. Performance of Davis-Resnick estimator versus *m* (constant *B*) for normally distributed test statistic versus *B*. Black box-plots represents D-R estimator, upper blue boxes the naïve MC procedure, and solid line the asymptotic P-value for a true Z={4,5,6,7,8,9,10}. Panels represent results for *m*= 5, 10, 20, 50 highest statistics in 10000 pseudosamples (1000 samples per condition).

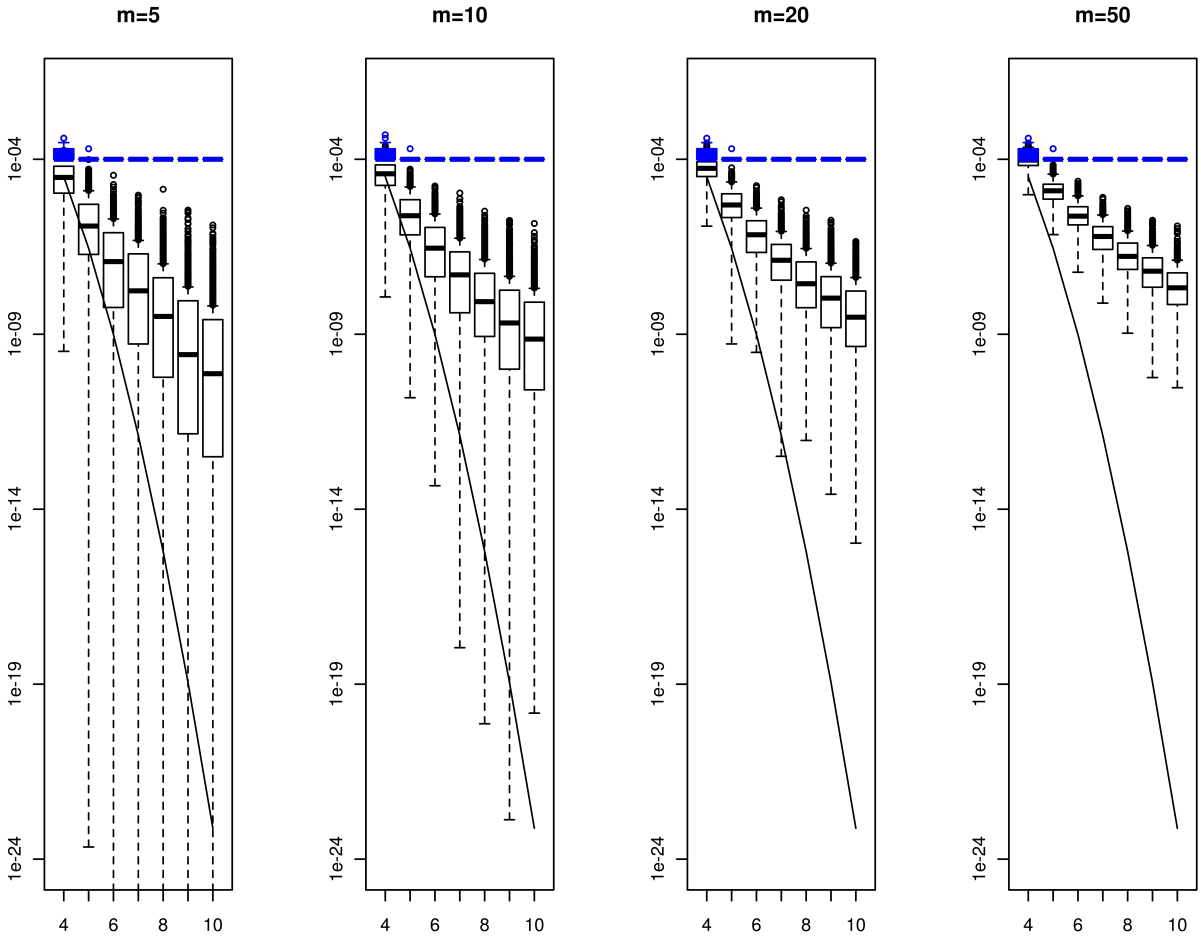
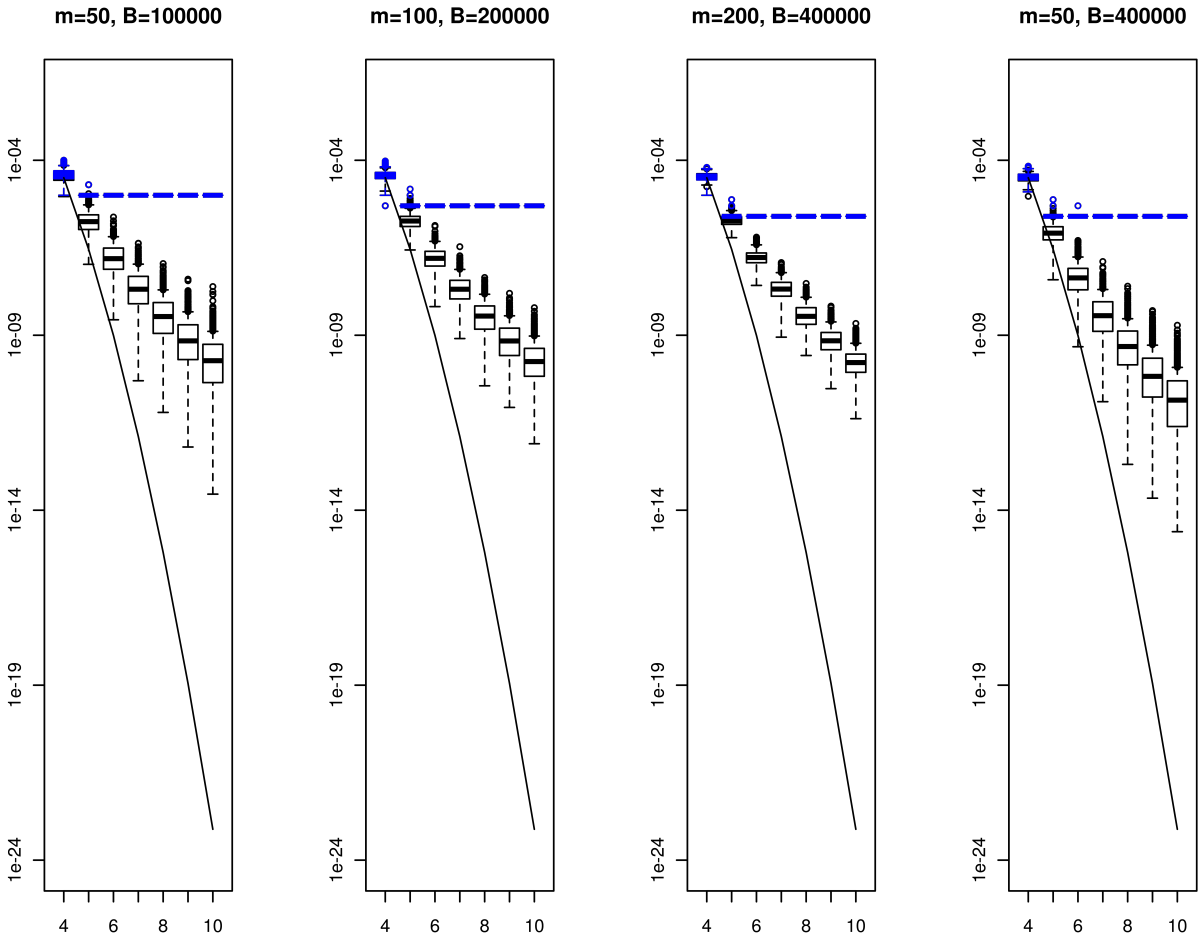


Fig 3. Performance of Davis-Resnick estimator with *larger* values of *B* for normally distributed test statistic. Black box-plots represents D-R estimator, upper blue boxes the naïve MC procedure, and solid line the asymptotic P-value for a true Z={4,5,6,7,8,9,10}. First three panels represent results for increasing *B*, holding constant *m/n*=0.0005; final panel shows reduction in bias by reducing *m/n* to 0.000125.



Resampling risk for this approach

This is easily defined, but hard to generalize upon since it depends on the size of the true effect, *n* and *m,* and the chosen significance threshold α.

A small real life example

Where the recombination distance is small, we can use a simple association X^2 with a gene-dropping test of significance to detect linkage within a single pedigree. Hall et al [1990] report on a pedigree in which a BRCA1 pathogenic variant is segregating. The LOD under a fully penetrant dominant model is 3.01, equivalent to a two-tailed P-value of $2x10^{-4}$ (two-tailed). I use the two-tailed P so we can appropriately compare it to the gene-dropping P-value.

With *B*=100000 replicates, in 1000 runs the mean Monte Carlo P was 0.00021 (SD=$4.6x10^{-5}$). With a threshold equivalent to a LOD=3 (P=0.0002), the resampling risk is 50% and the alternative hypothesis is accepted 50% of the time, as one would expect straddling a hard threshold. If one used a threshold LOD of 2 to flag results for further iteration, then with only 200 iterations, the Davis-Resnick P-value would exceed this on 87% of occasions (obviously, no naïve MC P-values could exceed 0.005).

Estimator	*B*	*m*	Mean estimated P (SD)	Reject H0 (Resampling risk)
Asymptotic	-	-	0.0002	-
DR	200	10	0.0014 (0.0019)	14% (24%)
DR	500	10	0.0009 (0.0011)	23% (35%)
DR	1000	10	0.0006 (0.0007)	33% (44%)
DR	2000	10	0.00052 (0.00046)	35% (46%)
DR	5000	10	0.00038 (0.00023)	29% (41%)
DR	10000	10	0.00030 (0.00014)	39% (48%)
DR	10000	20	0.00030 (0.00014)	41% (48%)
Naive	100000	-	0.00021 (0.000046)	50.0% (50%)

In summary, this approach has the advantage of requiring relatively small amounts of computation. It tends to be biased towards the null, but is not exact, in the sense that it can sometimes exceed the correct P-value. There is some room to develop adaptive methods to improve accuracy efficiently, but there is an interesting trade-off between the sampling error due to *m*, the number of extreme values used to estimate the P-value, and the bias, which will increase as *m*/*n* increases.

*Delete-*d *jackknife for intraclass correlations*

Another computer-intensive approach used by Sib-pair to analyse pedigree data is the **delete-*d* jackknife**. It is known that the standard **delete-1 jackknife** is inconsistent when data is correlated in nature. Deleting larger groups at a time gets around these problems (Shao and Tu, 1995).

The advantages of the jackknife are that it provides:

- An estimator with reduced bias
- An "automatic" "nonparametric" estimate of the sampling variance
- Cross-validation type model diagnostics (pseudo-values)
- (Sampling density estimation)

For experimental crosses and nuclear families, the *group* is easily and naturally the cross or family, but this breaks down for a single large kindred. The **random delete-d jackknife**, as the name implies, randomly selects subsamples (of a specified size) of observations for deletion.

$$V_{JD} = \frac{n-d}{dm} \sum_{i=1}^{m} \left(T_{n-d,i} - \frac{1}{m} \sum_{j=1}^{m} T_{n-d,j} \right)^2$$

where T is the test statistic, and m draws of size d from the n data have been made with replacement. In passing, this may be contrasted to the estimate where the data is partitioned into non-overlapping groups merely to ease computational load:

$$V_{JG} = \frac{m-1}{m} \sum_{i=1}^{m} \left(T_{n-g,i} - \frac{1}{m} \sum_{j=1}^{m} T_{n-g,j} \right)^2$$

I have previously used this approach to estimate jackknife standard errors for variance components, but this is slow, given the work required to estimate these quantities via maximum likelihood. They are much more competitive when applied to simpler statistics, such as the pairwise estimators of intraclass and interclass correlations.

Here are a few simulation results (500 replicates per condition) comparing the jackknife estimators for the intraclass and interclass correlation to ML solutions. The jackknife used the Sib-pair defaults: 200 pseudosamples with *d*=10.

True r	Sample Type	REML Mean estimated		ML Mean Estimated		Jackknife Mean Estimated	
		r	SE	r	SE	r	SE
0.33	100 sibships (size 2)	0.329	0.106	0.321	0.105	0.323	0.128
0.33	100 sibships (size 2-5)	0.328	0.077	0.323	0.076	0.322	0.071
0.33	100 sibships (size 5)	0.330	0.067	0.325	0.066	0.325	0.0
0.33	100 sibships (size 5-8)	0.332	0.062	0.328	0.061	0.330	0.044
0.33	100 sibships (size 8)	0.328	0.059	0.324	0.058	0.324	0.039

Generalized Linear Mixed Models (GLMMs)

Sib-pair offers fitting of GLMMs for the binomial-normal, probit-normal and multifactorial threshold model, poisson-normal, and weibull-normal. This is can be polygenic and/or single major locus or oligogenic/finite polygenic. Fitting is by a hybrid Markov Chain Monte Carlo algorithm, with Metropolis-Hastings slice sampler and Gibbs sampler steps. Unfortunately, this is still rather slow (and slow to mix). It does (!) give correct answers for standard nongenetic and genetic example datasets (eg segregation analysis, litter frailty toxicity trials). Eventually, I will extend the survival analysis to a mixed effects Cox Proportional Hazards model. Because of the speed issues, utility is currently limited to examining small numbers of candidate genes pointed to by other screening methods (in the GWAS context).

Bourgain et al *(2003) extended to categorical traits*

Bourgain *et al* (2003) presented a quasi-likelihood score test based on the weighted least squares analysis of linear models for a binary trait and allelic dose. This was subsequently extended by Thornton *et al* (2007). Sib-pair implements both the WQLS and MQLS tests for binary traits (as well as the BLUE for allele frequencies in pedigrees).

It is fairly simple to extend this to the multinomial case, at least for the **corrected χ^2** version of the test. The score equation takes the form,

$$S = U_2' I_{21}^{-1} U_2,$$

with.

$$U_{11} = 1' A^{-1} 1 ; U_2 = P' A^{-1} G ; U_{12} = 1' A^{-1} P ; U_{22} = P' A^{-1} P,$$

and,

$$I_{21} = U_{22} - (U_{21}' U_{11}^{-1} U_{21}).$$

P is a matrix of indicators for the levels of the trait; G, the matrix of allele counts; and A^{-1}, the inverse numerator relationship matrix. The test statistic for a k allele marker is,

$$T = \sum \sum (F^{-1})_{ij} U_{2(i)} I_{21}^{-1} U_{2(j)},$$

where F is the multinomial covariance matrix for the allele dose indicators.

An alternative approach which I have applied to SNPs is the so-called “left hand side” regression, fitted as a binomial-normal GLMM. This has the advantage of dealing more nicely with covariates because of the better behaved link function, and categorical and ordinal traits are conveniently dealt with. The covariance matrix for the allele doses between family members is no longer quite correct, however. This is not feasible for GWAS data (the Sib-pair GLLM implementation is very slow, but even fast codes such as those in the R *lme4* package still take 2-3 seconds per SNP in the family datasets we analyse. I suspect the multinomial WLS approach of Grizzle, Starmer and Koch, beloved of SAS PROC CATMOD users, might offer a faster alternative in a similar vein to Bourgain *et al*.

References

Barnard GA (1963) Discussion of "The spectral analysis of point processes" by MS Bartlett. *Journal of the Royal Statistical Society B*, 25, 294.

Besag J, Clifford P (1991). Sequential Monte Carlo p-values. *Biometrika*, 78(2):301-304.

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**: 612626.

Daniels HE (1954). Saddlepoint approximations in statistics. *Annals of Math. Stat*. 25:631-650.

Davis R, Resnick S (1984): Tail Estimates Motivated by Extreme Value Theory. *Annals of Statistics* 12: 1467-87.

Davison AC, DV Hinkley (1988). Saddlepoint approximations in resampling methods, *Biometrika* 75: 417-431.

Davison AC, Hinkley DV (1997). Bootstrap methods and their application. Cambridge University Press.

Duffy DL (2011). Sib-pair Version 1.00Beta. [Computer Program]. QIMR: Brisbane.

Hill BM (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* 3:1163-1174.

Hope ACA (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(3):582-598.

Shao J, Tu D (1995). The jackknife and bootstrap. *New York: Springer.*

Silva R, Assuncao. R, Costa M (2009). Power of the Sequential Monte Carlo Test. *Sequential Analysis: Design Methods and Applications* 28: 163-174.

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.

Wald A (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117-186.