The Admixture Model in Linkage Analysis

Jie Peng
D. Siegmund
Department of Statistics, Stanford University, Stanford, CA 94305

SUMMARY

We study an appropriate version of the score statistic to test the hypothesis of no linkage in a sample of phase known meioses and show that accurate approximations to the genome wide significance level can be obtained by using a large deviation approximation to evaluate Rice's formula for the expected number of upcrossings of a smooth random process.

1 Introduction

The core of the admixture model, proposed by Smith (1953) to deal with heterogeneous traits in genetic linkage analysis, is the hypothesis that a fraction λ of pedigrees is linked to the marker locus being tested while a fraction $1 - \lambda$ is unlinked. This model leads naturally to the vexing problem of testing for a *mixture* and has given rise to a large literature. See, for example, Chernoff and Lander (1995), Bickel and Chernoff (1993), Faraway (1993), Lathrop and Shoukri (1993), Chen (1998). In one simple, widely discussed formulation, we have a sample of N pedigrees, each of which contains k phase known meioses. The testing problem involves the general model of a sample of size N from a mixture of Bin(k, θ) and Bin$(k$,1/2) with weights λ and $1-\lambda$, respectively, where $0 \leq \theta \leq 1/2$. The parameter θ is the recombination fraction between a particular marker and a gene giving rise to the trait under consideration. The null hypothesis can be specified by either $\lambda = 0$ or $\theta = 1/2$, and partly because of this lack of identifiability, standard asymptotic likelihood theory does not apply.

One standard approach to this problem is that suggested by Davies (1977,1987), who uses Rice's formula for the expected number of upcrossings of a smooth Gaussian process to approximate the significance level of the max over θ of the standardized score statistic for testing $\lambda = 0$, but close examination shows that this approximation is often poor.

In Section 2 of this paper we discuss the arguably more appropriate translation of the genetic problem into a statistical problem involving mixtures suggested by Lander and Botstein (1986), where we test linkage to intervals between markers, not to individual markers. In Section 3 we present a modification of the Rice-Davies approximation similar to that of Rabinowitz and Siegmund (1995), which provides accurate approximations for the significance level as judged by comparison with simulations.

Although our methods apply reasonably generally, we consider the classical example of a sample of size N of k phase known meioses.

2 Description of the Model

We assume a sample of N pedigrees, with the nth pedigree providing k_n phase known meioses. Typically N is on the order of tens or hundreds while the k_n are in the range of one to five. Within the nth family, we let R_n denote the number of recombinations between the trait of interest and a fully informative marker. The most obvious special case is a fully penetrant dominant trait affecting a grandparent and k_n grandchildren, all of whom are genotyped. Then $k_n - R_n$ is the number of grandchilren sharing an allele identical by descent with the affected grandparent. For a homogeneous trait, the hypothesis of no linkage is that the R_n are distributed $\text{Bin}(k_n,\ 1/2)$, which is to be tested against the alternative that R_n is $\text{Bin}(k_n, \theta)$ for some $\theta < 1/2$. The parameter θ is the recombination fraction between a gene giving rise to the trait and the marker. The admixture model is an attempt to deal with heterogeneous traits by assuming that the affected members of some *pedigrees* have the trait by virtue of a gene linked to the marker, while affected members of other pedigrees are unlinked to the marker. The formal model is that R_n has a distribution that with probability λ is $\text{Bin}(k_n, \theta)$ and with probability $1 - \lambda$ is $\text{Bin}(k_n,\ 1/2)$; the null hypothesis of no linkage can be expressed by either $\lambda = 0$ or $\theta = 1/2$.

As suggested by Lander and Botstein (1986), this approach should be modified when, as is currently the practice, a large number of mapped markers distributed throughout the genome are used to search for evidence of linkage. The trait gene will typically lie in an interval between two markers, both of which used together provide more information, at least in principle, than the two taken separately.

Suppose the two markers have a known recombination θ_0 between them, and we want to test whether the trait gene is unlinked to these markers or lies in the chromosomal interval between them, say at a recombination fraction θ_1 from the first marker and θ_2 from the second marker. As a model for recombination, we assume the Haldane model of no interference, so $1 - \theta_0 = (1 - \theta_1)(1 - \theta_2) + \theta_1\theta_2$.

Let R_{00n} denote the number of meioses in the nth pedigree having no recombination between trait and first marker and no recombination between trait and second marker, R_{10n} the number of meioses having a recombination between trait and first marker but no recombination between trait and second marker, etc. For a homogeneous trait R_{ijn} for $i, j = 0, 1$ is a four-celled multinomial with parameters k_n and p_{ij} given in the case of linkage by

$$p_{ij} = \theta_1^i(1 - \theta_1)^{1-i}\theta_2^j(1 - \theta_2)^{1-j} \tag{1}$$

and in the case of no linkage by

$$p_{ij} = \theta_0^{i+j-2ij}(1 - \theta_0)^{(1-i)(1-j)+ij}/2. \tag{2}$$

In the admixture model the distribution is a mixture of (1) and (2) with weights λ and $1 - \lambda$, respectively, and we want to test the hypothesis that $\lambda = 0$. (Note that in this case the null hypothesis cannot be described as value for θ_1.) Let $\ell(\lambda, \theta_1)$ denote the log likelihood ratio of the mixture model relative to the null model, and let

$$J(\theta_1) = \mathrm{Var}[\partial\ell(\lambda, \theta_1)/\partial\lambda]|_{\lambda=0}$$

be the variance of the efficient score for testing $\lambda = 0$ at the value θ_1. Also let

$$Z(\theta_1) = \frac{[\partial\ell(\lambda, \theta_1)/\partial\lambda]|_{\lambda=0}}{[J(\theta_1)]^{1/2}} \tag{3}$$

be the normalized score statistic at θ_1. Since θ_1 is unknown, as a test statistic for a single interval, we propose following Davies (1987) and others to use

$$\max_{0\le\theta_1\le\theta_0} Z(\theta_1). \tag{4}$$

This should then be maximized over all intervals in order to cover the entire genome. We shall write $Z_i(\theta_1)$ when we want to emphasize that we are discussing the process Z on the ith marker interval. (A still more complete notation would also involve indexing by chromosome, but suppression of the chromosomal index should cause no confusion.)

3 Approximate p-values

It is easily verified that

$$\partial\ell(\lambda, \theta_1)\partial\lambda|_{\lambda=0}$$
$$= \sum_n\{(\frac{2\theta_1\theta_2}{1-\theta_0})^{R_{11n}}(\frac{2\theta_1(1-\theta_2)}{\theta_0})^{R_{10n}}(\frac{2(1-\theta_1)\theta_2}{\theta_0})^{R_{01n}}(\frac{2(1-\theta_1)(1-\theta_2)}{1-\theta_0})^{R_{00n}} - 1\}. \tag{5}$$

The variance of the nth term in this expression is

$$\sigma_n^2(\theta_1) = (\frac{2}{1-\theta_0}\theta_1^2\theta_2^2 + \frac{2}{\theta_0}\theta_1^2(1-\theta_2)^2 + \frac{2}{\theta_0}(1-\theta_1)^2\theta_2^2 + \frac{2}{1-\theta_0}(1-\theta_1)^2(1-\theta_2)^2)^{k_n} - 1, \quad (6)$$

and $J(\theta_1) = \sum \sigma_n^2(\theta_1)$. It is readily observed that when $k_n > 1$, the distribution of $Z(\theta_1)$ is positively skewed and can be extremely skewed. Although it is asymptotically normally distributed for large n, use of Gaussian process theory to approximate the distribution of (4) can lead to rather poor results.

Before turning to the general case, we consider briefly the case $k_n = 1$ for all n, which is closely related to a backcross (BC) in experimental genetics.

Suppose $k_n = 1$ for all n. Since $\sum R_{ij} = 1$, a generic term in the sum (5) can be re-written

$$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta_0}[2R_{00}-1] + \frac{(1-\theta_1)\theta_2}{\theta_0}[2R_{01}-1] + \frac{\theta_1(1-\theta_2)}{\theta_0}[2R_{10}-1] + \frac{\theta_1\theta_2}{1-\theta_0}[2R_{11}-1].$$

This is essentially the generic term in the numerator of the score statistic for interval mapping of quantitative traits in a BC design (Lander and Botstein (1989)), except that in that case the R_{ij} would be the average value of the phenotypic measurements in the four possible marker classes, two nonrecombinant and two recombinant at consecutive markers. Its distribution under the hypothesis of no linkage is symmetric, so is very well approximated by a normal distribution. When markers are not too closely spaced, Rice's formula as suggested by Davies (1977, 1987) has been shown to provide very good approximations to the distribution of (5) (Rebai, *et al.* (1994, 1995), Dupuis and Siegmund (1999)), while the approximation of Feingold *et al.*(1993) is useful for closely spaced markers. A brief discussion is contained in the Appendix. A compromise approximation that seems to cover both cases simultaneously has been suggested by Siegmund and Yakir (2003).

When $k_n > 1$, the Rice-Davies formula no longer provides adequate approximations, largely because the distribution of $Z(\theta_1)$ is skewed, even for large sample sizes. In the Appendix, we sketch a modification based on large deviation theory. We have also tried a simpler third moment correction in the spirit of that suggested by Tang and Siegmund (2001) for a different process. We found this to give a substantial improvement over the unmodified Rice-Davies formula, but still to fall short of the accuracy provided by the complete large deviation approach.

Table 1 gives numerical results comparing the Rice-Davies formula, our large deviations modification, and a Monte Carlo estimate based on 50,000 repetitions of the basic simulation. The sample size is $N = 100, 50, 20$ and values of $k_n = k = 1, 2, 4$ and 6 are considered. The results are given for a single chromosome representing roughly 1/22 part of the human autosomal genome. The thresholds were selected to give an exceedance probability of about 0.0022-0.0023, which is about 1/22 of the conventional 0.05 genomewide significance level. The results show that while the Rice-Davies approximation can easily be too small by an order of magnitude, our modification is very accurate and usually slightly conservative.

4 Discussion

In this paper (a) we have argued that a proper formulation of Smith's (1953) admixture model is that proposed by Lander and Botstein (1986) but then essentially ignored in the more recent literature, and (b) we have shown that the approximation suggested by Davies (1977, 1987) to the significance level of the maximum score statistic, which has proved useful in experimental genetics where statistics often have close to normal distributions, can be modified by the use of large deviation theory to give reasonable approximations for this problem as well.

A number of questions remain, regarding the reasonableness of the model and how it should be applied. The model is easy to criticize, since in its basic assumption that entire *pedigrees* are either linked or unlinked, it ignores the possibility that some individuals within a pedigree have their trait by virtue of mutations at one genetic location, while others have theirs by virtue of mutations somewhere else or for nongenetic reasons. It also assigns a single parameter λ to all pedigrees, no matter the size of the pedigree or the number of affecteds. While both these assumptions fly in the face of reasonable genetic models for complex traits, it is still possible that the simplifying structure they bring to an otherwise complex situation can be useful. To address this question requires some attention to power, particularly in comparison to more realistic, but less structured models. (For an interesting discussion of the artificiality of the admixture model and problems associated with the interpretation and estimation of λ, see Whittemore and Halpern (2001).)

An important question involves the value of the continuous process $Z_i(\theta)$, which interpolates between markers in the ith marker interval via the recombination fraction θ, in comparison with the simpler process consisting only of the discrete skeleton $Z_i(0)$. The analogous question has been studied in experimental genetics. It turns out that unless markers are fairly widely spaced or a breeding design is used that increases the expected number of recombinations between markers (analogous to using more distant relatives in human genetics), the interpolation adds little to the power to detect linkage although it does give a clearer picture of where in the interval between markers one can expect a linked gene to be found. See, for example, Darvasi and Soller (1995) and Dupuis and Siegmund (1999)

On a more technical level, while we have discussed only the case of phase known meioses, which means in effect that we assume data from at least three generations and complete penetrance of the trait, our methods are more generally valid. However, other models will often lead to less skewed data, so somewhat simpler approximations, e.g., along the lines of the third moment corrections suggested by Rabinowitz and Siegmund (1995) or Tang and Siegmund (2001) may suffice. (We computed such approximations in our current study but have omitted them because they seem inferior to the large deviation approximations we have discussed and are anti-conservative.) Phase unknown meioses, which are relevant when marker information from only two generations is available, are a straightforward example, but more interesting examples involve the reduced penetrance one expects to find in dealing with complex traits (cf. Liang (2002)).

5 Appendix

Let Y_t be a stochastic process with mean value identically 0, variance 1, and piecewise twice differentiable sample paths. Then

$$\mathrm{P}\{\max_{0 \le t \le t_0} Y_t \ge b\} \le \mathrm{P}\{Y(0) \ge b\} + \mathrm{E}(N),$$

where N is the number of upcrossings of the level b by the process Y_t, which under general conditions has expected value given by

$$\mathrm{E}(N) = \int_0^{t_0} \int_0^{\infty} y \mathrm{P}\{Y_t = b, \dot{Y}_t = y\} dy, \tag{7}$$

where $\dot{Y}_t$ denotes differentiation with respect to t and the probability under the integral sign denotes the joint probability density function of the indicated random variables. When the process is Gaussian, this joint density can be simplified at points where the sample paths are smooth, since Y_t and $\dot{Y}_t$ are jointly Gaussian and are uncorrelated, hence independent, and also $\mathrm{E}[\dot{Y}_t] = 0$. This leads to the formula

$$\mathrm{E}(N) = (2\pi)^{-1/2} \int_0^{t_0} \mathrm{P}\{Y_t = b\}[\mathrm{Var}\dot{Y}_t]^{1/2} dt. \tag{8}$$

Under the assumptions given, the probability density function of Y_t is simply the standard normal distribution, but we have left this less explicit form in display (8) to prepare for the following generalization, which is obtained by using a large deviation approximation to the joint probability distribution in (7).

To this end we define $\psi(\xi, \eta, t) = \log \mathrm{E}[\exp(\xi Y_t + \eta \dot{Y}_t)]$, and denote partial derivatives with respect to ξ, η, t by subscripts. Let $\xi_0 = \xi_0(t)$ be defined to satisfy $\psi_\xi(\xi_0, 0, t) = b$ and introduce the probability distribution

$$\mathrm{P}_{\xi_0}\{Y_t \in dx, \dot{Y}_t \in dy\} = \exp[\xi_0 b - \psi(\xi_0, 0, t)] \mathrm{P}\{Y_t \in dx, \dot{Y}_t \in dy\}. \tag{9}$$

Since $\mathrm{E}_{\xi_0}(Y_t) = b$, we should obtain a reasonable approximation to the joint distribution appearing in (7) by inverting (9) and applying a normal approximation to the P_{ξ_0} joint distribution of $Y_t, \dot{Y}_t$. In particular the marginal distribution of Y_t is given approximately by

$$\mathrm{P}\{Y_t = b\} \sim (2\pi \mathrm{Var}_{\xi_0} Y_t)^{-1/2} \exp[-\{\xi_0 b - \psi(\xi_0, 0, t)\}]. \tag{10}$$

The variance entering into (10) and more generally the moments of the P_{ξ_0}-joint distribution of $Y_t, \dot{Y}_t$ are easily obtained by differentiating $\psi(\xi, \eta, t)$.

It is easy to see that unlike the Gaussian case, $\mathrm{E}_{\xi_0}(\dot{Y}_t)$ and $\mathrm{Cov}_{\xi_0}(Y_t, \dot{Y}_t)$ do not vanish in general, although it seems reasonable to expect them to be small. If we assume they can be neglected, our approximation is of the form of (8), but with the marginal probability density of Y_t replaced by (10) and the variance of $\dot{Y}_t$ calculated under the probability P_{ξ_0}. Otherwise the approximation also involves the P_{ξ_0} mean of $\dot{Y}_t$ and correlation of $Y_t, \dot{Y}_t$, and is somewhat more complicated.

REFERENCES

Chen, J. (1998). Penalized likelihood-ratio test for finite mixture models with multinomial observations, *Canadian J. Statist.* **26**, 583-599.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial, *J. Statist. Plann. Inference* **43**, 19-40.

Darvasi, A., Weinreb, A., Minke, V., Weller, J.I. and Soller, M. (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134** 943-951.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* **64** 247-254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* **74** 33-43.

Dupuis, J. and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers, *Genetics* **151**, 373-386.

Faraway, J. (1993). Distribution of the admixture test for the detection of linkage under heterogeneity, *Genet. Epidemiol.* **10**, 75-83.

Feingold, E., Brown, P.O. and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent, *Am. J. Hum. Genet.* **53**, 234-251.

Lander, E.S. and Botstein, D. (1986). Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphis, *Proc. Nat. Acad. Sci. USA* **83**, 7353-7357.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* **121**, 185-199.

Lathrop, M. and Shoukri, M. (1993). Statistical Testing of Genetic Linkage Under Heterogeneity, *Biometrics* **49**, 151-161.

Liang, K. Y.

Rebai, A., Goffinet, B. and Mangin, B. (1994). Approximate thresholds of interval mapping test for QTL detection. *Genetics* **138** 235-240.

Rebai, A., Goffinet, B. and Mangin, B. (1995). comparing power of different methods for QTL detection. *Biometrics* **51** 87-99.

Siegmund, D. and Yakir, B. (2003).

Smith,

Tang, H. K. and Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models, *Biostatistics* **2**, 147-162.

Whittemore, A. and Halpern, J. (2001). Problems in the Definition, Interpretation, and Evaluation of Genetic Heterogeneity, *Am.J.Hum.Genet.* **68**, 457-465.

Table 1

Tail Probabilities. The number of phase known meioses is k, the sample size is N, the distance between markers is Δ, the number of (equally spaced) markers on a chromosome is M, and the threshold is b. The column headed P_0 gives the Rice-Davies approximation, P_1 is the approximation suggested in this paper, and MC denotes an estimate based on 50,000 repetitions of a Monte Carlo experiment.

k	N	Δ	M	b	P_0	P_1	MC
1	100	20	8	3.45	0.0027	0.0025	0.0022
				2.70	0.0278	0.0277	0.0250
2	100	20	8	3.70	0.0012	0.0024	0.0025
				3.00	0.0132	0.0191	0.0202
4	100	20	8	4.20	0.0002	0.0024	0.0022
				3.00	0.0163	0.0411	0.0367
6	100	20	8	4.90	1×10^{-5}	0.0025	0.0023
				4.00	0.0006	0.0148	0.0125
4	50	20	8	4.45	7×10^{-5}	0.0022	0.0022
				3.50	0.0032	0.0190	0.0155
6	50	20	8	5.30	1×10^{-6}	0.0028	0.0022
				4.50	7×10^{-5}	0.0112	0.0109
4	20	20	8	4.75	2×10^{-5}	0.0027	0.0023
				4.00	0.0005	0.0124	0.0098
6	20	20	8	6.25	5×10^{-9}	0.0024	0.0024
				5.50	5×10^{-7}	0.0069	0.0041
4	100	10	16	4.40	0.0001	0.0021	0.0022
				3.50	0.0049	0.0205	0.0192
4	100	5	31	4.45	0.0002	0.0026	0.0021
				3.50	0.0069	0.0289	0.0243