Ann. Hum. Genet. (2002), **66**, 211–221 © *University College London*
DOI: 10.1017/S0003480002001112 *Printed in the United Kingdom*

Mapping quantitative effects of oligogenes by allelic association

W. ZHANG[1], A. COLLINS[1], G. R. ABECASIS[2], L. R. CARDON[2] AND N. E. MORTON[1]

[1] *Human Genetics Division, Duthie Building (MP 808), Southampton General Hospital,*
Tremona Road, Southampton SO16 6YD, UK

[2] *Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK*

(Received 10.10.01. Accepted 11.3.02)

SUMMARY

Regression analysis of a quantitative trait as a function of a single diallelic polymorphism has been extended to allelic association by composite likelihood under the Malecot model for multiple markers. We applied the method to 10 single nucleotide polymorphisms (SNPs) spanning 27 kb of the angiotensin-I converting enzyme (*ACE*) gene in British families, localising a causal SNP between G2530A and 4656(CT)3/2 in the 3′ region, at a distance of 21.6 ± 0.9 kb from the most proximal SNP T-5491C. Neither they nor the *I/D* polymorphism is causal. To clarify genetic parameters we applied combined segregation, linkage and association analysis. Stronger evidence for the 3′ region was obtained, with significant evidence of a lesser 5′ effect as reported in French and Nigerian families. However, rigorous confirmation requires that the causal SNPs be identified. Both Malecot and parametric analysis appear to have high power by comparison with alternative methods for localizing oligogenes and their causal polymorphisms.

INTRODUCTION

Positional cloning is the process by which loci of unknown structure and function, but with some effect on a phenotype of interest, can be identified by mapping to a small region that may be refined and ultimately cloned prior to sequencing. Mapping may be through *linkage* measured by sex-specific recombination without regard to genotype, *allelic association* measured by dependence of allele frequencies at two loci without regard to sex-specific recombination, *chromosome rearrangement* that alters gene number or function, or by combination of the three approaches. The locus that is positionally cloned is usually related to disease through affection or a quantitative trait. *Major genes* have effects so large that they are capable of causing disease in the absence of other predisposing factors, environmental or genetic. *Oligogenes* have smaller effects that act cumulatively and are harder to detect. *Polygenes* have effects so small, and the number affecting a particular trait is so large, that current mapping methods offer little hope of detection except for alleles at recognised major loci. Positional cloning of major loci is now routine, and genetics is committed to development of powerful methods for oligogenes.

Here we present an extended regression approach to map causal SNP(s) by allelic association under the Malecot model for multiple markers. This method (which is weakly parametric or 'non-parametric' because gene frequencies and effects are not specified) was then applied to 10 SNPs of the locus for angiotensin-I converting enzyme (*ACE*) in random British families (Keavney *et al.* 1998). Results were compared to parametric combined segregation/linkage/association analysis. Both methods localized a major causal SNP in the 3′ region and

Correspondence: Professor N. E. Morton, Human Genetics Division, Duthie Building (Mailpoint 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK. Tel: +44 (0) 23 8079 6536; Fax: +44 (0) 23 8079 4264.
E-mail: nem@soton.ac.uk

Table 1. *Haplotype frequencies and phenotype means in a random sample*

	Predictive SNP			
Causal SNP	P	P′	Total	Mean
G	$QR+\rho Q(1-R)$	$Q(1-R)-\rho Q(1-R)$	Q	μ_G
G′	$R(1-Q)-\rho Q(1-R)$	$(1-R)(1-Q)+\rho Q(1-R)$	$1-Q$	$\mu_{G'}$
Total	R	$1-R$	1	—
Mean	μ_P	$\mu_{P'}$	—	—

P, P′: Alleles at predictive SNP; G, G′: Alleles at causal SNP; μ_P, $\mu_{P'}$: Phenotype means, predictive SNP; μ_G, $\mu_{G'}$: Phenotype means, causal SNP; Q, R: Allele frequencies; ρ: Association.

a minor 5′ causal SNP. The latter was not evident in an alternative approach, showing that the new method will be helpful for localising causal SNPs of quantitative traits with oligogenic inheritance using multiple markers.

A MODEL FOR QUANTITATIVE EFFECTS OF OLIGOGENES

Major genes are usually rare and therefore studied in families that provide haplotypes in which presence or absence of a disease-related allele may be reliably inferred. The candidate region may be narrowed by recombination and refined by merging associated microsatellite marker alleles, calculating for each marker locus an appropriate measure of association ρ, and estimating the location of a disease locus in a high-resolution marker map (Collins & Morton, 1998; Lonjou *et al.* 1998). In contrast, oligogenes are usually common enough so that their allelic associations are more efficiently estimated in cohort or case-control studies (Morton & Collins, 1998), and their effects small enough that no reliance can be placed on merging alleles by a significance test. Diallelic markers such as single nucleotide polymorphisms (SNPs) are convenient, as well as more efficiently typed than the microsatellites that have proven useful for allelic association with major loci and for linkage. Our objective is to use a set of SNPs typed in a candidate region to determine the most likely location for a causal site, together with standard errors and significance levels. This is achieved by modelling the pattern of association with disease in the SNP map. Here we assume a random sample of families, like the QTDT program that estimates single-locus regression coefficients. We assume a single causal SNP in a candidate region having an effect on a quantitative trait y. The allele frequencies at the causal SNP are Q, $1-Q$ for alleles G, G′, respectively. An allele count x takes the values 0, 1 for the presence of the allele in a haplotype or 0, 1, 2 for the count in genotypes, and gives the regression $y = a_G + b_G x$, where $b_G = \mu_G - \mu_{G'}$, and μ_k is the mean of y for the kth allele. Often, as in the example we present here, the causal SNP is not identified, and so μ_G, $\mu_{G'}$ and therefore b_G are unknown, and we rely on association with markers in a region.

Now consider a predictive (marker) SNP with frequencies R, $1-R$ for alleles P, P′ respectively, ordered so that in expectation allele P is associated with allele G at the causal SNP (Table 1). Expressing μ_P and $\mu_{P'}$ in terms of μ_G and $\mu_{G'}$ we obtain, from Table 1, $\beta = \mu_P - \mu_{P'} = \rho b_G Q/R$, from which

$$\rho = \beta R / b_G Q. \tag{1}$$

The association parameter ρ is optimal in theory, most robust to variation in marker allele frequencies, and in trials on human data is at least twice as efficient for positional cloning as the regression coefficient (Morton *et al.* 2001). When b_G and Q are unknown we cannot fit a model directly to ρ, but βR and β may be modelled. We will show how together they estimate the unknown parameter Q. In this indirect way a weakly parametric method is able to estimate parameters not explicit in the model, but less reliably than a correct parametric model. By incorporating multiple markers in composite likelihood, the analysis is more powerful than if each marker were considered separately.

STATISTICAL ANALYSIS

The number of diallelic polymorphisms in the human genome is greater than one million, and the density in coding regions is greater than 1 per kb. With present methods only a fraction of these will be tested even in candidate regions, and so it is unlikely that a causal polymorphism will be included in a random set of associated SNPs. Therefore the largest values of βR in a sample may well be less than $b_G Q$. Even if a causal SNP is sampled, the estimate of $b_G Q$ is subject to error and will not be the maximum likelihood estimate unless expressed as a parameter, say $T = b_G Q$. Since $0 \leqslant \rho \leqslant 1$ (Collins & Morton, 1998), we adopt the convention that $0 \leqslant \beta \leqslant b_G$ and so $0 \leqslant T < b_G$. The Malecot equation for isolation by distance has been used to describe the decline of association with distance (Collins & Morton, 1998; Lonjou *et al.* 1998) and therefore gives a location for a causal site within a candidate region. If we express ρ in terms of β, R and T, the equation for the ith associated polymorphism becomes

$$\rho_i = \beta_i R_i / T \; = (1-L) M \exp(-\epsilon d_i) + L, \quad (2)$$

where L is an error term, representing the probability of spurious association through population stratification (allelic association between unlinked markers which could occur in inter-racial crosses or nonrandom mating), or the constraint that $\rho_i \geqslant 0$. L can exceed zero but has rarely been tested in other methods, although less efficient ways to control for population stratification have been advocated and indeed are in use (Ewens & Spielman, 1995). M is the proportion of disease alleles transmitted from a unique founder haplotype (and so is 1 if disease alleles are monophyletic and less than 1 if they are polyphyletic). The distance d_i between marker i and the causal locus should be measured in centimorgans (cM) if the genetic map is accurate, or on the physical map (kb) otherwise (Lonjou *et al.* 1998). When available, a linkage disequilibrium map will give more reliable localization than either cM or kb (Maniatis *et al.* 2002). As with major genes (Collins & Morton, 1998), distance d_i may be expressed as $d_i = \delta_i (S_i - S)$, where S is the location of the causal SNP, S_i is the map location of the ith associated SNP, and $\delta_i = 1$ if $S_i \geqslant S$ or -1 otherwise. This notation is required to make the likelihood differentiable with respect to S. The parameter ϵ reflects recombination and time. It depends on the number of generations t during which the haplotypes have been approaching equilibrium and the pressure to disrupt them by recombination θ, mutation, and perhaps selection; in the simplest case $\epsilon d = \theta t$. Pairs of SNPs in the region give an estimate of ϵ, assumed constant. The swept radius (Morton *et al.* 2001) $1/\epsilon$ is the distance at which linkage disequilibrium (LD) declines to $e^{-1} \approx 0.37$ of its original value.

If we assume a single monophyletic causal SNP in a region, then the parameter M is 1 and so equation (2) becomes

$$\beta_i R_i = T\{(1-L) \exp(-\epsilon d_i) + L\} \equiv z_i. \quad (3)$$

The composite likelihood makes the customary assumption that deviations of ρ from the Malecot expectation are independent normal variables with information proportional to $K\rho$ (Lindsay, 1988; Devlin *et al.* 1996). The logarithm of the composite likelihood is $\ln L = -\Sigma_{i=1}^{m} K_{zi} (\hat{z}_i - z_i)^2/2$, where $i = 1, 2, \ldots, m$ SNPs are considered, $\hat{z}_i$ is an estimate of $\beta_i R_i$ with expectation z_i and information $K_{zi} = 1/R_i^2$ var (β_i). Since the variance of β_i is $V_i / 2n_i R_i (1-R_i)$, where V_i is the variance of deviations from regression and n_i are the observations (individuals) contributing to the β_i estimate for the ith SNP, the information about z_i is

$$K_{zi} = 2n_i (1-R_i)/R_i V_i. \quad (4)$$

This gives $\chi_1^2 = (\beta_i R_i)^2 \, K_{zi}$ to test the hypothesis that $z_i = 0$ and an alternative expression for V_i as $2n_i R_i (1-R_i) \beta_i^2 / \chi_i^2$. Iterative maximum likelihood estimation follows Collins & Morton (1998) with no realistic alternative to using the residual variance as error to allow for sources of variation not in the model. The confidence interval around an estimate of S identifies where a causal SNP should be sought.

In the above analysis the effect $\hat{z}$ of the causal SNP is not resolved into its allele frequency Q

and regression coefficient b_G. Cardon & Abecasis (2000) have developed a theory to estimate β_i in families in their QTDT program and set upper and lower limits to Q. We can obtain a direct estimate, however, as $Q = T/F$ with approximate variance Var $(T)/F^2 + T^2$ Var $(F)/F^4$, where $T = b_G Q$ as defined above, and F is the expected value of b_G and is a parameter in the Malecot equation:

$$\beta_i = F\{(1-L)\ \exp(-\epsilon d_i)+L\}, \qquad (5)$$

with weight $K_{\beta_i} = 1/\mathrm{Var}(\beta_i) = 2n_i R_i(1-R_i)/V_i$. Since β_i shows more confounding with allele frequencies than ρ_i we expect equation (3) to be more reliable than equation (5).

The frequency of the G′P haplotype is $R(1-Q)-\rho Q(1-R) = (R-Q)\rho + R(1-Q)(1-\rho)$, where the frequency in founders is $R-Q$ as in Table 1 of Collins & Morton (1998). Although true in expectation, the constraint $R_i \geqslant Q$ can be violated by genetic drift or origin of the allele $\mathrm{P_i}$ subsequent to the causal SNP. In the latter event, association with the G allele could be either positive or negative. We do not constrain allelic frequency Q to be less than the smallest R_i, since that would be an increasingly serious mistake as the number of associated markers increases. Three factors protect against error. Firstly, the alleles in a small candidate region are chosen so that their correlations with the trait y are positive and their correlations *inter se* are almost all positive. Then their complementary alleles have the same correlation matrix, but correlations with y are of opposite sign. We assign the R_i to the vector with the greater χ^2 in the test of $\epsilon = 0$. Secondly, the combination of equations (3) and (5) to estimate Q does not impose $R_i \geqslant Q$ even when one of the markers is assumed to be a causal SNP. Thirdly, if the initial location S does not coincide with a marker, Q cannot be constrained by R_i, even if iteration converges to a marker location. Experience is required to evaluate these options.

The parameters to be estimated are L, ϵ, S, and T or F. On the null hypothesis all parameters except S are zero. χ^2 is calculated as the difference between the minus 2 natural log likelihood ($-2\ \ln L$) of the two models (hypotheses) compared, divided by the error variance for the more general model if the residual is significant. If there is evidence against ϵ and $T = 0$ then a causal SNP exists and is estimated to be at location S. The above algorithms have been implemented in the ALLASS program, which is available from http://cedar.genetics.soton.ac.uk/public_html/.

MALECOT ANALYSIS OF THE *ACE* LOCUS

The angiotensin-I converting enzyme is coded by the *ACE* locus on human chromosome 17q23 (Mattei *et al.* 1989). *ACE* is polymorphic for one or more determinants of enzyme level (Rigat *et al.* 1990). Ten markers were detected by sequencing four subjects with discordant enzyme levels (Villard *et al.* 1996). Keavney *et al.* (1998) performed a cladistic analysis of the common haplotypes in 27 kb of the *ACE* locus. They inferred two ancestral clades A and B, differing at all markers, and a recombinant clade C with frequencies 0.43, 0.31, and 0.16 respectively in founders. Clades B and C are associated with a codominant increase in enzyme level. These data were submitted to single-locus analysis by the QTDT program, using the β_w regression coefficient which extracts information only from families with a parent who is heterozygous for the marker allele (Abecasis *et al.* 2000*b*). This analysis concluded that the critical region is indistinguishable from polymorphisms G2215A, I/D, and G2350A in the 3′ half of *ACE*, spanning more than 10 kb. To increase resolution we made three extensions: from β_w to the more powerful β regression coefficient provided by the QTDT program; from β to the still more powerful association parameter ρ (equation 1); and from single marker analysis to multi-marker analysis by the Malecot model (equations 2–5).

These data consist of 666 individuals in 83 random British extended families. Pedigrees range from two to three generations, consisting of 4 to 18 individuals each. Plasma ACE levels were measured for 405 individuals and standardised within sex to give a phenotype y.

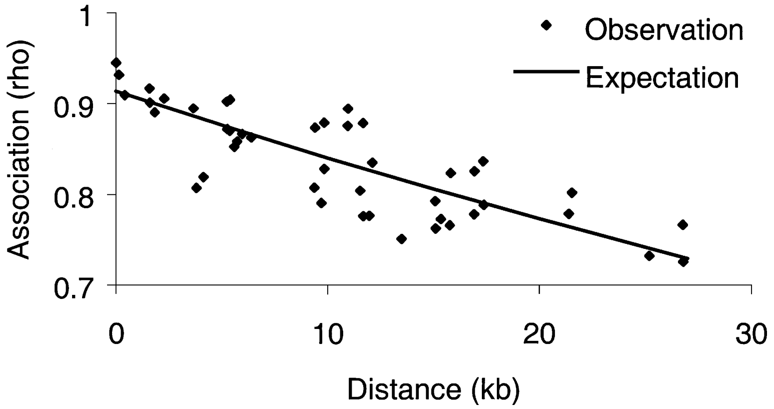

Fig. 1. Allelic association for pairs of SNPs in the *ACE* data.

There was no significant association between the phenotype and age and body mass index. Ten SNPs in the *ACE* gene were genotyped. Fourteen families were genotyped but have no phenotype information, reducing the analysis to 553 individuals in 69 families.

Before analysing the ACE phenotype we examined association ρ between the 45 pairs of SNPs (Table 3) using the ALLASS program. Haplotypes were inferred by SIMWALK (Keavney *et al.* 1998; Sobel & Lange, 1996). In contrast to less densely mapped regions (Morton & Wu, 1988; Jorde *et al.* 1994) the decline of association with distance is highly significant even over this short distance (Fig. 1). The Malecot model for $L = 0$ gives estimates $\epsilon = 0.0084 \pm 0.0009$, $M = 0.914 \pm 0.009$, with $\chi^2_{43} = 41.48$. Fitting L simultaneously gives $\chi^2_{42} = 40.50$, a nonsignificant improvement ($\chi^2_1 = 0.98$), indicating that L is not different from zero. Accepting that L is zero, the average SNP in this region has about 9 percent of polyphyletic origin by mutation or gene conversion (100% minus 91.4%). The swept radius of association is $1/\epsilon = 119$ kb, agreeing with other evidence that association typically extends far beyond the 3 kb predicted by simulation (Kruglyak, 1999). As shown in Table 3, and proven algebraically, the estimates of correlation for allele counts cannot exceed the estimates of association (Morton *et al.* 2001).

To derive a point estimate and confidence interval from the phenotypic composite likelihood requires two decisions. First, the vector of positive associations with clade B alleles corresponds to smaller gene frequencies (mean 0.43) than the negative associations with clade A (mean gene frequency 0.57). In Table 4, the test of $\epsilon = 0$ gives $\chi^2_2 = 29.45$ in the first case (model 3 vs. 2, divided by the error variance 15.64/7) and only $\chi^2_3 = 0.86$ in the second (model 11 vs. 10). All SNP correlations under the hypothesis that the low ACE determinant is ancestral are positive. This is consistent with the clade frequencies, and the cladogram, in suggesting that clade A is ancestral, clade B was derived by multiple mutations, and clade C arose by recombination between A and B. It is tempting to suggest that clades A and B characterised different populations, perhaps Paleolithic hunters and Neolithic farmers with subsequent origin of clade C, but evolutionary evidence is weak. The hypothesis that clade B introduced a factor for elevated enzyme level gives values of z that are higher in the segment that clade C putatively received from clade B than in the segment that corresponds with clade A (Table 2). On the contrary, the hypothesis that clade B is ancestral and clade A a derivative with reduced enzyme level gives allele frequencies greater than the putatively older clade, and uninformative values of z.

A second decision is to substitute β for β_w in order to use all the information in the data. β is estimated from all phenotyped individuals with genotype information. Both β and β_w are provided by the QTDT program (Abecasis *et al.* 2000*a*). The less efficient β_w, corresponding to

Table 2. *Data on the ACE locus* (*Keavney* et al. *1998*)

Marker from		Clades			Associated allele				
5′ to 3′	Location kb	A	B	C	P	β	R	χ^2	$z = \beta R$
T-5491C	0	T	C	T	C	0.695	0.332	64.18	0.229
A-5466C	0.025	A	C	A	C	0.707	0.335	65.03	0.236
T-3892C	1.599	T	C	T/C	C	0.746	0.426	84.47	0.317
A-240T	5.251	A	T	A	T	0.677	0.330	61.73	0.222
T-93C	5.398	T	C	T	C	0.678	0.333	59.99	0.225
T1237C	10.979	T/C	C	C	C	0.769	0.576	95.38	0.442
A2215G	15.108	G/A	G/A	A	A	0.879	0.495	125.60	0.434
I/D	16.945	I	D	D	D	0.864	0.502	127.42	0.433
G2350A	17.372	A	G	G	G	0.884	0.501	130.23	0.442
4656(CT)3/2	26.796	3	2	2	2	0.894	0.501	129.84	0.447

Table 3. *Associations at the ACE locus* (*ρ above diagonal, correlation below*)

Marker	T-5491C	A-5466C	T-3892C	A-240T	T-93C	T1237C	A2215G	*I/D*	G2350A	4656(CT)3/2
T-5491C	●	0.945	0.902	0.873	0.905	0.895	0.763	0.826	0.789	0.726
A-5466C	0.928	●	0.917	0.903	0.870	0.876	0.793	0.778	0.837	0.767
T-3892C	0.723	0.749	●	0.895	0.807	0.808	0.751	0.773	0.766	0.732
A-240T	0.865	0.895	0.725	●	0.932	0.859	0.829	0.776	0.835	0.802
T-93C	0.893	0.866	0.656	0.928	●	0.853	0.791	0.805	0.777	0.779
T1237C	0.544	0.539	0.610	0.525	0.522	●	0.820	0.867	0.863	0.824
A2215G	0.530	0.560	0.647	0.581	0.555	0.715	●	0.891	0.906	0.879
I/D	0.569	0.545	0.662	0.539	0.560	0.762	0.882	●	0.910	0.879
G2350A	0.543	0.586	0.655	0.578	0.539	0.766	0.897	0.910	●	0.875
4656(CT)3/2	0.505	0.543	0.636	0.562	0.549	0.719	0.873	0.874	0.868	●

Table 4. *Malecot model: Composite likelihood, all data* (*values in parentheses fixed by hypothesis*)

Model	Hypothesis	$-2\ln L$	D.F.	ϵ	T	S
	Tests of general model fitting T:					
1	General ($\hat{\text{L}} = 0.283$)	15.03	6	0.084	0.597	21.752
2	$L = 0$	15.64	7	0.040	0.545	21.859
3	$L = \epsilon = 0$	81.45	9	(0)	0.318	—
4	$L = \epsilon = T = 0$	943.87	10	(0)	(0)	—
	standard error, $L = 0$	—	—	0.005	0.035	1.279
	Assuming $L = 0$ and fitting T:					
5	T1237C causal	78.73	8	0.017	0.363	(10.979)
6	A2215G causal	36.75	8	0.042	0.460	(15.108)
7	*I/D* causal	27.84	8	0.038	0.462	(16.945)
8	G2530A causal	26.30	8	0.038	0.466	(17.372)
9	4656 (CT) 3/2 causal	24.93	8	0.027	0.520	(26.796)
	Switching $R \rightarrow 1-R$ and fitting T:					
10	General ($\hat{\text{L}} = 0.434$)	9.43	6	0.618	0.946	22.377
11	$L = \epsilon = 0$	10.29	9	(0)	0.421	—
		$-2\ln L$	D.F.	ϵ	F	S
	Fitting model for F:					
12	General ($\hat{\text{L}} = 0.622$)	1.13	6	0.135	1.087	22.103
13	$L = 0$	1.69	7	0.014	0.945	22.910
	standard error, $L = 0$	—	—	0.005	0.061	3.683

Models 12 and 13 used the alternative method that estimates F instead of T in order to calculate the gene frequency Q of the causal SNP. $\hat{\text{L}}$: L is estimated at this value.

the TDT of Ewens & Spielman (1995), discounts information in parents and children of homozygotes in order to protect against allelic association between unlinked genes, which has not been detected in populations without racial mixture, is efficiently controlled by the L parameter of the Malecot model, and is not relevant to localisation within a region known on other evidence to have a quantitative trait locus (QTL). As an example, the maximum lod was 28.28 when estimating total association based on β, whereas the lod is only 14.59 for the TDT test based on β_w at the same SNP G2350A, losing about half of the information.

The ACE phenotype agrees with $L = 0$ ($\chi_1^2 = 0.24$ for model 2 vs. 1, Table 4) but residual χ^2 is significant ($p = 0.02$, residual variance $= 15.03/6$), and the estimate of ϵ is much greater than for pairs of SNPs in the same interval (0.04 vs. 0.0084). Contributing factors may include more than one causal SNP and likely origin of clade C through recombination. A causal SNP is localized at 21.75 ± 1.28 kb between G2530A and 4656(CT)3/2 (model 1, Table 4). Neither they nor the *I/D* polymorphism is likely to be causal, since the smallest of the three values of $-2 \ln L$ exceeds the maximum likelihood value (model 9 vs. 2) by $\chi_1^2 = 4.16$.

Evidence for a causal SNP in the 3′ region is highly significant (model 3 vs. 2), since $\chi_2^2 = 29.45$ corresponds to $\text{lod}_1 = 5.58$ (Collins & Morton, 1998). However, the effect is so great that the five proximal SNPs converge to $\epsilon = 0$, obscuring possible evidence for a 5′ causal SNP of lesser effect. We therefore examined the two regions separately in the same sample, but ignoring SNPs in the other region. The seven markers in the 3′ region identify the same causal SNP at 21.63 ± 0.94 kb, with $\chi_2^2 = 52.70$, $\text{lod}_1 = 10.49$. The standard error is reduced because residual variation is no longer significant, and the highly significant lod is further increased by exclusion of markers in the 5′ region. The five most proximal markers suggest a causal SNP in the promoter region at 2.59 ± 0.34 kb. Despite this small standard error and highly significant evidence against $T = 0$ ($\chi_1^2 = 329.82$), there is little evidence against $\epsilon = 0$ ($\chi_2^2 = 5.53$). Equation (5) gives $Q = 0.596 \pm 0.072$ for the 3′ causal SNP and 0.512 ± 0.217 for the 5′ causal SNP. Although significant by a large-sample test, the standard error of the latter is too great for reliable estimation of allele frequency or effect.

PARAMETRIC ANALYSIS

Analysis of families controls population stratification (if any), and increases power by direct estimation of gene frequency and effect. To reduce 10 SNPs to a smaller number of haplotypes we performed principal component analysis assuming the SNP genotypes are quantitative (coded as 0, 1, and 2 for allele counts). The second eigenvector was of opposite sign for SNPs 1–5 in the 5′ region, and 6–10 in the 3′ region. Treating the latter separately and regressing the ACE level y on those principal components, we found only the first eigenvector (designated $y_{3'}$) to be significant in stepwise regression. Repeating this process for the 5′ region with $y_{3'}$ included, $y_{3'}$ and the first eigenvector (designated $y_{5'}$) were significant. The four quadrants defined by $y_{3'}$ and $y_{5'}$ correspond to clades A, B, C, and a small residual clade D with 3′ resemblance to clade A and 5′ resemblance to clade B. This procedure assigns a small number of ambiguous haplotypes to the clade with the most similar pattern of SNPs. The four clades defined in this way were treated as alleles A, B, C, and D, with founder frequencies 0.459, 0.316, 0.197, and 0.028, respectively.

With this information we applied the parametric COMDS program under a model of two loci with linkage and association at one of them (Morton *et al.* 1991; Shields *et al.* 1994). This program was written for a single marker locus and is not optimal for multiple SNPs. COMDS partitions a pedigree into nuclear families, with pointers and probands to control ascertainment bias. Since this is a random sample, pointers and probands are not required here. When a pedigree is partitioned an individual may appear as both child and parent. To avoid inflation of the type I error without sacrificing efficiency when parents and children are considered jointly, the phenotype was omitted for the child in such cases, retaining markers if typed. To prevent confounding of skewness or kurtosis with either locus, COMDS uses a polychotomy instead of a quantitative trait, assigned to affected as severity and to normals as diathesis, with frequencies estimated from the sample. Information was maximised by defining affection on the median and taking equal frequencies within each of the 9 classes provided by the program, giving 18 equally frequent classes. Scaling parameters for the effects of the two loci on diathesis and severity (called B, S, B_m, and S_m in COMDS) are expected to be unity under a true hypothesis. This is confirmed for tenable hypotheses. Further simplification is provided by assuming that the allelic effects are additive within and between loci ($D = D_m = 0.5$), that linkage to the marker is complete ($\theta = 0$), and that the unlinked modifier has gene frequency $Q_m = 0.5$. These assumptions are supported by statistical tests.

The other parameters to be estimated are displacement between homozygotes at the linked locus (T) and the modifier (T_m), gene frequency Q for a single causal site at the linked locus, and coupling frequencies c_i representing allelic association between that site and the ith haplotype with frequency q_i for $i = 1, \ldots, 4$ (MacLean *et al.* 1984). Then $c_i q_i$ is the frequency of haplotype i in coupling with the susceptible allele at the causal site. Since $Q = \Sigma c_i q_i$, the vector of coupling frequencies is constrained by $c_j = (Q - \Sigma_{i \neq j} \mathrm{c}_i \mathrm{q}_i)/\mathrm{q}_j$ and therefore is of rank 3. We also tried more complicated models, but none was significantly better than this parsimonious set.

There is no evidence for dominance or against complete linkage by either the single locus or 2-locus test (Tables 5 and 6). The latter gives slightly higher lods in the presence of allelic association. The lod for linkage in the presence of association is 18.49, compared with a maximum of 6.88 for linkage by the QTDT program. The lod testing association in the presence of linkage is 31.73, compared with a maximum of 28.28 for total association (β) by the QTDT program, which does not give the other comparisons in Table 6. In this example a parametric model is more powerful than non-parametric methods. Combining the linkage and association tests, $\mathrm{lod}_1 = 33.58$ for the parametric model. Note that the QTDT was applied to the two alleles of a SNP in pedigrees, whereas COMDS used four haplotypes as alleles in nuclear families. Composite likelihood in ALLASS does not extract linkage

Table 5. *COMDS: Major locus with and without additive modifier (values in parentheses fixed by hypothesis)*

Model/Hypothesis	$-2 \ln L$	Major locus				Recombination	Modifier		Coupling frequencies			
	-2000	D	T	Q	V	θ	T_m	V_m	A	B	C	D
With modifier												
1. No association or linkage	361.19	(0.5)	1.96	0.184	0.290	(0.5)	1.85	0.426	0.184	0.184	0.184	0.184
2. Linkage, no association	347.95	(0.5)	1.90	0.584	0.439	(0)	1.48	0.274	0.584	0.584	0.584	0.584
3. Association, no linkage	281.56	(0.5)	1.86	0.504	0.432	(0.5)	1.30	0.210	0.034	0.982	0.855	0.340
4. Association+linkage	196.39	(0.5)	1.78	0.470	0.396	(0)	1.41	0.248	0	0.984	0.809	0
5. 3′ region effect only	199.93	(0.5)	1.74	0.487	0.378	(0)	1.41	0.249	(0)	0.950	0.950	(0)
6. D estimated	193.25	0.577	1.77	0.471	0.400	(0)	1.39	0.242	0.003	0.976	0.820	0
7. Association+$\hat{\theta}$	196.39	(0.5)	1.78	0.470	0.396	0	1.41	0.248	0	0.984	0.809	0
8. 5′ effect only	279.37	(0.5)	1.35	0.344	0.206	(0)	1.73	0.373	(0)	0.999	(0)	0.999
No modifier												
1. No association or linkage	365.47	(0.5)	2.23	0.684	0.535	(0.5)	(0)	—	0.684	0.684	0.684	0.684
2. Linkage, no association	353.57	(0.5)	2.10	0.617	0.521	(0)	(0)	—	0.617	0.617	0.617	0.617
3. Association, no linkage	291.78	(0.5)	2.02	0.443	0.503	(0.5)	(0)	—	0.047	0.837	0.767	0.211
4. Association+linkage	222.45	(0.5)	2.01	0.416	0.492	(0)	(0)	—	0.036	0.843	0.679	0
5. 3′ effect only	230.27	(0.5)	1.90	0.399	0.434	(0)	(0)	—	(0)	0.778	0.778	(0)
6. D estimated	219.93	0.581	1.89	0.426	0.464	(0)	(0)	—	0.030	0.855	0.721	0
7. Association+$\hat{\theta}$	221.75	(0.5)	2.04	0.421	0.509	0.036	(0)	—	0.034	0.847	0.697	0.017

Table 6. *COMDS: χ^2 tests of hypotheses*

			With modifier		No modifier	
Null hypotheses	Models	D.F.	χ^2	Lod_1	χ^2	Lod_1
No association or linkage	1 vs. 4	3	164.80	33.58	143.02	28.91
No linkage	1 vs. 2	1	13.24	2.88	11.90	2.58
No association|linkage	2 vs. 4	2	151.56	31.73	131.12	27.32
No dominance ($D = 0.5$)	4 vs. 6	1	3.14	—	2.52	—
Complete linkage ($\theta = 0$)	4 vs. 7	1	0	—	0.70	—
No linkage|association	3 vs. 4	1	85.17	18.49	69.33	15.05
No 5′ effect	5 vs. 4	1	3.54*	—	7.82**	—

1-tailed p-value: $* < 0.03$; $** < 0.003$.

information, and the test for association is not directly comparable. It is approximately $(943.87–81.45)/(2 \ln 10)(81.45/9) = 20.69$ (model 4 vs. 3, Table 4). As the most powerful of the three tests, the value of $\chi^2_1 = 3.54$ for no 5′ effect on COMDS (1-tailed $p < 0.03$, and $p < 0.003$ if no modifier) is suggestive of a causal SNP in the promoter (model 5 vs. 4, Tables 5 and 6). The corresponding variance component is estimated to be 0.206, compared with 0.378 for the 3′ region. These estimates are model-dependent, like the value of 0.6 for additive genetic variance at the *ACE* locus (Abecasis *et al.* 2000*b*) and none accurately describes a locus with two or more causal SNPs.

DISCUSSION

We have shown that none of the tested 3′ SNPs in *ACE* is causal, but there is one or more causal SNPs near 21.6 kb, between G2350A and 4656(CT)3/2. Substantially more information for positional cloning is provided through analysis by the Malecot model than by the TDT option in the QTDT program, which led to the inference that markers A2215G, *I*/*D*, and G2350A 'are in complete disequilibrium with the trait alleles (and could be the trait alleles themselves)' (Abecasis *et al.* 2000*b*). Zhu *et al.* (2001) have reported that G2350A is the polymorphism most significantly associated with ACE level in a large sample of Nigerian families. In random samples of Europeans and Afro-Caribbeans they found the highest association in the G2350A-4656(CT)3/2 interval, close to the more distal marker (Zhu *et al.* 2000). McKenzie *et al.* (2001) performed a study in Jamaican Afro-Caribbean families on the 10 SNPs they used on British Caucasian families (Keavney *et al.* 1998). They found that only marker G2350A showed strong but incomplete disequilibrium with the *ACE*-linked QTL, using both β and β_w in the QTDT program. This may reflect population heterogeneity, but supports our analysis in British families that the *I*/*D* polymorphism is unlikely causal. All these studies used different measures of association, but they are in reasonable agreement.

Very different results have been reported for the 5′ region, where Villard *et al.* (1996) suggested a causal SNP in French families, whereas Keavney *et al.* (1998) found no evidence in British families. In the same data Abecasis *et al.* (2000*b*) found a local maximum at T-3892C, but this was not significant when 3′ markers were considered. Zhu *et al.* (2000) relied on these results and did not test 5′ markers in Europeans and Afro-Caribbeans, but in their Nigerian families they report a highly significant 5′ effect localised near A-240T (Zhu *et al.* 2001). In McKenzie *et al.* (2001, see their Figure 1) study, there is a small peak of association in the 5′ portion of the *ACE* gene at T-93C. In our analysis of the data from Keavney *et al.* (1998) we find a causal SNP in the interval between T-2892C and A-240T significant at the 0.03 level. Estimates of ϵ do not differ significantly among the 10 SNPs considered here (Fig. 2). Weighted by the information $K_\epsilon = (1/\sigma_\epsilon)^2$, they give $\chi^2_9 = 13$, with no evidence for a linear or quadratic trend by the method of Collins *et al.* (2001). Consequently the genetic map over these 27 kb appears proportional to physical distance, which permits good resolution of one or more causal locations in the region.

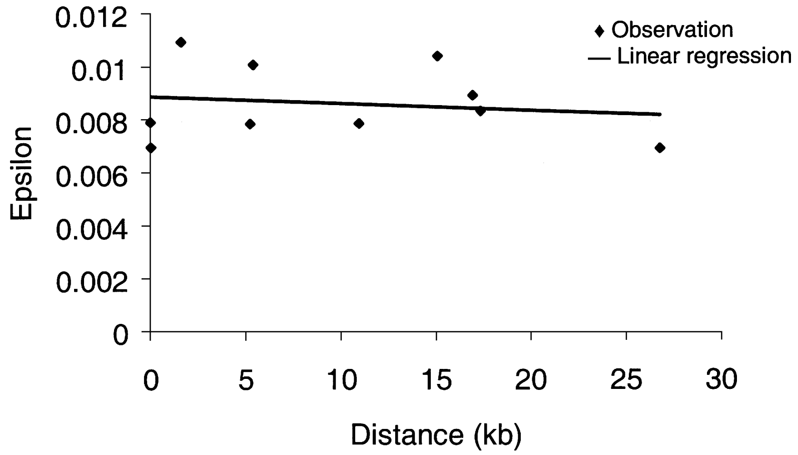

Fig. 2. Values of ϵ against physical location of markers for $M = 0.914$.

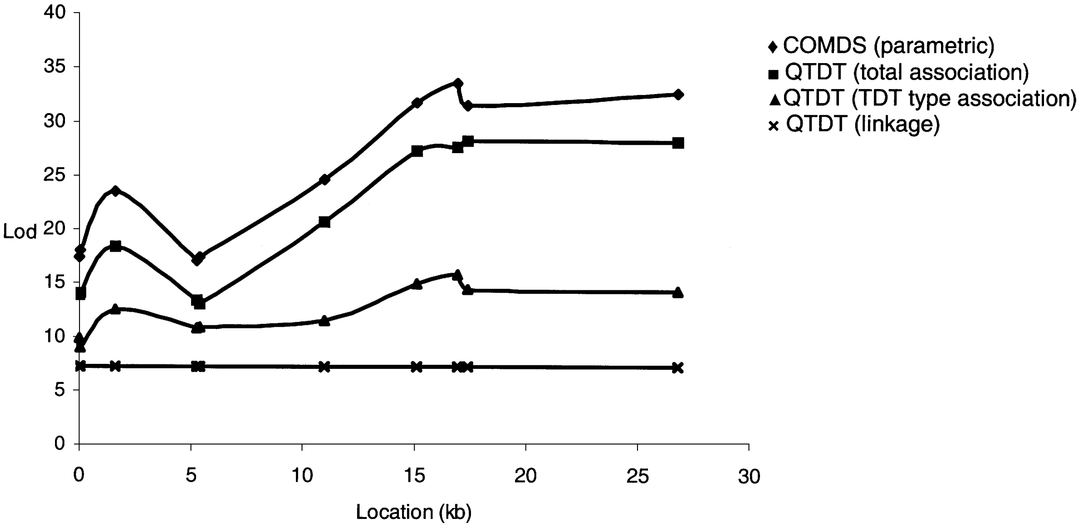

Fig. 3. Lod at each SNP calculated by COMDS (with modifier, combined association and linkage) and the QTDT program.

Identification of two or more causal sites within a small distance is not feasible by linkage. For insulin-dependent diabetes mellitus (IDDM), even at ~ 40 cM from the susceptibility locus *IDDM1*, the existence of *IDDM15* is controversial (Delepine *et al.* 1997). Allelic association greatly increases resolution of a second causal site, especially when the causal polymorphism is identified as in the hemochromatosis (*HFE*) locus, where the oligogene *H63D* has been resolved from the major *C282Y* substitution (Feder *et al.* 1996; Beutler, 1997). The existence of one or more 5′ causal SNPs in *ACE* will not be demonstrated until they have been used as markers and shown to be more closely associated with enzyme levels than nearby SNPs. Pending that demonstration, there is supporting evidence from concordance of our analysis of British data summarised in Table 6 and Fig. 3, the report of Villard *et al.* (1996) on French families, and the observations of Zhu *et al.* (2001) on Nigerians.

We are grateful to Bernard Keavney, Colin McKenzie, and colleagues who collected and genotyped the ACE families. Any requests for the data should be directed to either of them (bernard.keavney@well.ox.ac.uk or camcken@unimona.edu.jm).

REFERENCES

Abecasis, G. R., Cardon, L. R. & Cookson, W. O. C. (2000*a*). A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292.

Abecasis, G. R., Cookson, W. O. C. & Cardon, L. R. (2000*b*). Pedigree tests of transmission disequilibrium. *Eur. J. Hum. Genet.* **8**, 545–551.

Beutler, E. (1997). The significance of the 187G (H63D) mutation in hemochromatosis. *Am. J. Hum. Genet.* **61**, 762–764.

Cardon, L. R. & Abecasis, G. R. (2000). Some properties of a variance components model for fine-mapping quantitative trait loci. *Behav. Genet.* **30**, 235–243.

Collins, A., Ennis, S., Taillon-Miller, P., Kwok, P.-Y. & Morton, N. E. (2001). Allelic association with SNPs: metrics, populations and the linkage disequilibrium map. *Hum. Mutat.* **17**, 255–262.

Collins, A. & Morton, N. E. (1998). Mapping a disease locus by allelic association. *Proc. Natl. Acad. Sci. USA* **95**, 1741–1745.

Delepine, M., Pociot, F., Habita, C., Hashimoto, L., Froguel, P., Rotter, J., Cambon-Thomsen, A., Deschamps, I., Djoulah, S., Weissenbach, J., Nerup, J., Lathrop, M. & Julier, C. (1997). Evidence of a non-MHC susceptibility locus in type I diabetes linked to HLA on chromosome 6. *Am. J. Hum. Genet.* **60**, 174–187.

Devlin, B., Risch, N. & Roeder, K. (1996). Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36**, 1–16.

Ewens, W. J. & Spielman, R. S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.* **57**, 455–464.

Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A., Dormishian, F., Domingo, R. Jr., Ellis, M. C., Fullan, A., Hinton, L. M., Jones, N. L., Kimmel, B. E., Kronmal, G. S., Lauer, P., Lee, V. K., Loeb, D. B., Mapa, F. A., McClelland, E., Meyer, N. C., Mintier, G. A., Moeller, N., Moore, T., Morikang, E., Wolff, R. K. *et al.* (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* **13**, 399–408.

Jorde, L. B., Watkins, W. S., Carlson, M., Groden, J., Albertsen, H., Thliveris, A. & Leppert, M. (1994). Linkage disequilibrium predicts physical distance in the denomatous polyposis coli region. *Am. J. Hum. Genet.* **54**, 884–898.

Keavney, B., McKenzie, C. A., Connell, J. M., Julier, C., Ratcliffe, P. J., Sobel, E., Lathrop, M. & Farrall, M. (1998). Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum. Mol. Genet.* **7**, 1745–1751.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 221–239.

Lonjou, C., Collins, A., Ajioka, R. S., Jorde, L. B., Kushner, J. P. & Morton, N. E. (1998). Allelic association under map error and recombinational heterogeneity: a tale of two sites. *Proc. Natl. Acad. Sci. USA.* **95**, 11366–11370.

MacLean, C. J., Morton, N. E. & Yee, S. (1984). Combined analysis of genetic segregation and linkage under an oligogenic model. *Comput. Biomed. Res.* **17**, 471–480.

Maniatis, N., Collins, A., Xu, C.-F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X. & Morton, N. E. (2002). The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. USA* **99**, 2228–2233.

Mattei, M.-G., Hubert, C., Alhenc-Gelas, F., Roekel, N., Corvol, P. & Soubrier, F. (1989). Angiotensin I-converting enzyme gene is on chromosome 17. *Cytogenet. Cell Genet.* **51**, 1041.

McKenzie, C. A., Abecasis, G. R., Keavney, B., Forrester, T., Ratcliffe, P. J., Julier, C., Connell, J. M., Bennett, F., McFarlane-Anderson, N., Lathrop, G. M. & Cardon, L. R. (2001). Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Hum. Mol. Genet.* **10**, 1077–1084.

Morton, N. E. & Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *Proc. Natl. Acad. Sci. USA* **95**, 11389–11393.

Morton, N. E., Shields, D. C. & Collins, A. (1991). Genetic epidemiology of complex phenotypes. *Ann. Hum. Genet.* **55**, 301–314.

Morton, N. E. & Wu, D. (1988). Alternative bioassays of kinship between loci. *Am. J. Hum. Genet.* **42**, 173–177.

Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.-Y. & Collins, A. (2001). The optimal measure of allelic association. *Proc. Natl. Acad. Sci. USA* **98**, 5217–5221.

Rigat, B., Hubert, C., Alhenc-Gelas, F., Cambien, F., Corvol, P. & Soubrier, F. (1990). An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. *J. Clin. Invest.* **86**, 1343–1346.

Shields, D. C., Ratanachaiyavong, S., McGregor, A. M., Collins, A. & Morton, N. E. (1994). Combined segregation and linkage analysis of Graves disease with a thyroid autoantibody diathesis. *Am. J. Hum. Genet.* **55**, 540–554.

Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores and marker-sharing statistics. *Am. J. Hum. Genet.* **58**, 1323–1337.

Villard, E., Tiret, L., Visvikis, S., Rakotovao, R., Cambien, F. & Soubrier, F. (1996). Identification of new polymorphisms of the angiotensin I-converting enzyme (*ACE*) gene and study of their relationship to plasma ACE levels by two-QTL segregation-linkage analysis. *Am. J. Hum. Genet.* **58**, 1268–1278.

Zhu, X., Bouzekri, N., Southam, L., Cooper, R. S., Adeyemo, A., McKenzie, C. A., Luke, A., Chen, G., Elston, R. C. & Ward, R. (2001). Linkage and association analysis of Angiotensin I-converting enzyme (*ACE*) gene polymorphisms with ACE concentration and blood pressure. *Am. J. Hum. Genet.* **68**, 1139–1148.

Zhu, X., McKenzie, C. A., Forrester, T., Nickerson, D. A., Broeckel, U., Schunkert, H., Doering, A., Jacob, H. J., Cooper, R. S. & Rieder, M. J. (2000). Localization of a small genomic region associated with elevated ACE. *Am. J. Hum. Genet.* **67**, 1144–1153.