# Mapping multiple linked quantitative trait loci in non-obese diabetic mice using a stepwise regression strategy

HEATHER J. CORDELL*, JOHN A. TODD AND G. MARK LATHROP

*The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK*

(*Received 5 March 1997 and in revised form 11 August 1997 and 10 November 1997*)

## Summary

A simple regression strategy for mapping multiple linked quantitative trait loci (QTLs) in inbred populations is proposed and applied to data from a non-obese diabetic (NOD) mouse backcross. The method involves adding and deleting markers from a linear model in a stepwise manner, allowing the association with a particular marker to be examined once associations with other (in particular neighbouring) markers have been taken into account. This approach has the advantage of using programs available in standard statistical packages while still allowing adequate separation of possible multiple linked effects. For the mouse backcross, using these methods, at least two and possibly three diabetogenic loci are detected on each of chromosomes 1 and 3. Some evidence for epistasis is seen between the loci on chromosome 1, with a possible additional epistatic interaction between the loci on chromosome 3. Congenic strain analysis of the chromosome regions in NOD diabetes suggests that although the true type I error rate may be larger than that suggested by the nominal *P* values, our results nevertheless correspond well with those disease loci and interactions detected using a congenic approach, indicating that the regression method may be a powerful strategy for the detection and characterization of QTLs in inbred populations.

## 1. Introduction

An important aim in genetics and breeding is the identification of those genes that contribute significantly to quantitative variation within and between populations or species. With the recent rapid development of molecular markers, detection and mapping of quantitative trait loci (QTLs) in experimental organisms has been greatly facilitated (Paterson *et al.*, 1988; O'Brien, 1993; Wang *et al.*, 1994), and interest has now focussed on the development of appropriate statistical methodology for the analysis and interpretation of experimental data. In response to the shortcomings of traditional QTL analysis (Soller *et al.*, 1976), Lander & Botstein (1989) proposed an interval mapping method in which the strength of evidence for QTLs at various positions along the genome is assessed. The same techniques have been used in outbred populations (Lathrop *et al.*, 1985). Recently Kruglyak & Lander (1995) proposed a non-parametric version of the interval method. There are still problems with this method, however, particularly in distinguishing multiple linked QTL effects. Zeng (1993, 1994) has addressed these problems by proposing a method that combines interval mapping with multiple regression, producing test statistics in an interval that are independent of the effects of QTLs at other regions of the chromosome. Moreno-Gonzalez (1992) and Jansen (1994) have also proposed using regression methods, which have some similarities and also some differences compared with the method proposed by Zeng (1993) (see Zeng (1993) for a comparison of the methods).

The method proposed by Zeng (1993, 1994) involves fitting a model of the form

$$y = \beta_0 + \beta^* x^* + \sum_{j \neq i, i+1} \beta_j x_j + \epsilon \qquad (1)$$

to test for a QTL in a marker interval $(i, i+1)$, where $y$ is the phenotypic value of the trait, $x_i$ is a binary variable corresponding to the genotype (homozygous

[*] Corresponding author. Present address: Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, 2500 MetroHealth Drive, Cleveland, OH 44109, USA.

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

or heterozygous for a backcross) at a given marker $i$, $\beta^*$ is the effect of the putative QTL expressed as the difference in effects between the homozygote and heterozygote, $x^*$ is an indicator variable taking a value 0 or 1 depending on the genotypes of the markers $i$ and $i+1$ and the testing position of the putative QTL, and $\epsilon$ is assumed to be normally distributed with mean 0. Maximization of the likelihood requires specific programming via the ECM algorithm (Meng & Rubin, 1993), although recently a suite of programs (QTL Cartographer) has been developed (Basten *et al.*, 1994) to fit the model in the case of specific crosses. This model contrasts with the standard model for multiple regression analysis, which is of the form

$$y = \beta_0 + \sum_i \beta_i x_i + \epsilon \qquad (2)$$

and which in the case where only one marker $x_i$ is included corresponds to traditional QTL methods (Soller *et al.*, 1976). The main reason suggested for *not* using this model given by Zeng (1993) is the fact that the partial regression coefficient for a marker will be a biased estimate of the true QTL effect. However, since the bias is downwards, this problem is less serious if an effect is detected, as we are more (rather than less) confident that it reflects a true effect. While there are still implications for the power of the method to detect linkage, we would not expect to inflate the type I error using this approach. We therefore propose using (2) as the basis of a genetic-model-free method for assessing the effect and position of QTLs, while making no assumptions as to the underlying genetic mechanism behind the disease. By including different combinations of markers in the regression equation, we can examine the effects associated with a particular marker once effects at other regions of the genome have been taken into account. This allows the identification and positioning of putative disease loci (see Section 2 for details). The regression method effectively becomes a genetic-model-free method for inferring the location of disease loci in relation to a fixed set of markers, since although the specific models fitted are parametric in terms of the error distribution and linear predictors, the genetic inference we shall make from them is not. This approach has the advantage of allowing detection and separation of possible multiple linked QTLs using a simple yet flexible model whose properties are well understood and which is available in any standard statistical analysis package.

## 2. Methods

The advantage of the regression approach that we will exploit is that we are not limited to including a single marker at a time in the model, nor to including all available markers, but rather may include any number of linked and unlinked loci simultaneously. Comparing the fit of different models, including different combinations of markers, allows us to examine the effect associated with a particular marker once associations at other markers have already been taken into account. The idea is to fit linear models of the form $y = \beta_0 + \Sigma_i \beta_i x_i + \epsilon$, where the linear predictors $x_i$ may be interaction terms (in a statistical sense) as well as factors referring to the genotype at the markers typed. By including predictors $x_i$ in a stepwise manner, adding or deleting them once other (in particular neighbouring marker) predictors have already been included and noting the increase or decrease in residual deviance, some indication of the position of potential QTLs may be obtained.

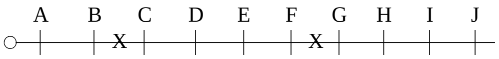

Fig. 1. Example of positions of marker and disease loci on a typical chromosome that might be analysed using stepwise regression methods.

As an example, suppose we have disease loci positioned as shown (marked X) in Fig. 1 in relation to markers A–J, assuming a marker spacing of about 10 cM. Using single-locus methods we are quite likely to see a broad spread of significance, with the effects from the two disease loci combining to give the strongest association at marker E, which in fact corresponds to a ‘ghost’ QTL lying between the two disease loci (Martinez & Curnow, 1992). If markers are entered in a stepwise manner, however, then we may move across the chromosome testing whether, for instance, marker B gives a significantly better fit (in terms of the reduction in deviance) when added to the model including just marker A, whether C gives a significantly better fit when added to the model including just B, and so on. In this case, we would expect that B will be significant when added to the model including just A, since B is closer to the first disease locus. Similarly we would expect C to be significant when added to the model including A and B, or in fact including just B, since C is on the other side of the disease locus from B, and so information about the genotype at C should increase the information about the genotype at the disease locus. We would not expect to see a significant improvement in fit when adding marker D to models including C, or E to models including D, since in this case the disease locus is separated from the marker by another marker that has already been included in the model and, as Zeng (1993) has shown, the partial regression coefficient of a marker in a multiple regression analysis will be independent of any QTLs located outside the two adjacent markers, assuming there is no epistasis. If F is added to the model just E or markers to the left of E, however, we would hope to see a significant improvement in fit since we are beginning to gain information about the second disease locus. Similarly if G is added after F we expect significance whereas if H is added to G we expect no significance.

This reasoning forms the basis for a regression strategy where we move back and forth across a chromosome fitting models where one marker or two adjacent markers are included at a time. Consider the region B–C–D. If C is significant after B (i.e. the model including both B and C fits significantly better than the model including B only) but D not significant after C, this suggests that a QTL lies between B and the midpoint of C–D: to the right of B since C adds information when added to B, but closer to C than to D since D does not improve the fit when added to C. Similarly if C is significant when added to D but B not significant when added to C, a QTL between D and the midpoint of B–C is suggested. If all are significant once the neighbour has been included, or all are non-significant, we cannot position a QTL in relation to the three markers: it is the pattern of significance followed by non-significance that allows us to position the QTL in this way.

Clearly the power of such a strategy will depend critically on the marker spacing and position of any QTLs in relation to the marker map, as well as on the magnitude and relative magnitudes of any QTL effects. Also this strategy makes no use of the inter-marker distances in the modelling process. Under the situation shown in Fig. 1, if distances are small, it may be that all stepwise tests (B after A, C after B, etc.) will show significance, although assuming a reasonable marker spacing in relation to the sample size (so that both recombinants and non-recombinants will occur between any two markers) we would still expect to swap from significance to non-significance when testing H after G as opposed to G after F, and similarly for A after B as opposed to B after C. This strategy would therefore still allow us to hypothesize the existence of two separate effects in the F–G– and –B–C region. If the QTLs are closely linked to particular markers, it may be that a simpler strategy such as forward selection or backward deletion of these individual markers is sufficient to deduce the existence of the QTLs or model the data. To detect a QTL in a region while allowing for epistasis between the QTL and QTLs in the same or neighbouring regions, a forward selection strategy where two neighbouring markers together with their interaction term are entered simultaneously into the model (i.e. three new terms) may be adopted (see Section 5). Having deduced the position of the QTLs by whichever strategy or combination of strategies is preferred, epistatic interactions between the disease loci can be assessed by fitting interaction terms between the closest linked marker or markers to each QTL.

## 3. Experimental details

The experimental data to be analysed were generated from the mouse backcross between the non-obese diabetic (NOD) strain and the diabetes-resistant strain C57BL/10-NOD.$H2^{g7}$ (B10.$H2^{g7}$), described in Todd *et al.* (1991). The NOD mouse is a widely used genetic model for type 1 (insulin-dependent) diabetes melitus (IDDM) because of its spontaneous development of the disease, which shares many immunopathological and genetic features with the human disorder. Following this previous report (in which 61 marker loci were analysed on 97 spontaneous diabetic and non-diabetic BC1 progeny) up to 106 diabetic and 190 non-diabetic BC1 progeny were typed using a total of 123 marker loci throughout the mouse genome. The markers and phenotypes measured are described in Ghosh *et al.* (1993). Animals in the first backcross generation were classified as homozygous (1) or heterozygous (0) at each locus typed, and were monitored for the development of diabetes and insulitis. Spontaneous diabetics had ages of onset from 94 to 436 days. The histology of the pancreas in terms of the extent of insulitis was assessed in the non-diabetic progeny by grading histology into seven categories of ascending severity: 0, no evidence of lymphocytes in the pancreas; 1, some periductal lymphocyte infiltration; 2, peri-islet infiltration, no insulitis; 3, very mild insulitis in some islets with no reduction in islet cell mass; 4, extensive insulitis with significant islet cell mass remaining; 5, extensive insulitis with significant reduction in islet cell mass; 6, as for 5 but only residual islets remaining.

From previous studies (Ghosh *et al.*, 1993; Wicker *et al.*, 1995; McAleer *et al.*, 1995) at least 14 loci in addition to the major locus *Idd1* on chromosome 17 have been shown to contribute to the development of diabetes or insulitis. For the regression analyses performed here we focussed on those regions on chromosomes 1 and 3 that were of particular interest because of their linkage to insulitis as well as diabetes, and because of the spread of significance across a wide region, suggesting that they may each contain more than one disease locus. The regions of interest on chromosomes 1 and 3 are shown in Fig. 2. Data were not available at *D1Nds2* (marker G on chromosome 1) or *D3Mit19* (marker J on chromosome 3) for a substantial number of animals; therefore these markers were excluded from the analyses. Strictly speaking this is not necessary as any standard method (e.g. the EM algorithm; Dempster *et al.*, 1977) for dealing with missing values in a regression problem could be used. However, we wanted to keep the simplicity of analysing the data using standard procedures available in any statistical programming package, and therefore opted for exclusion. This left available for analysis a total of 296 animals typed

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

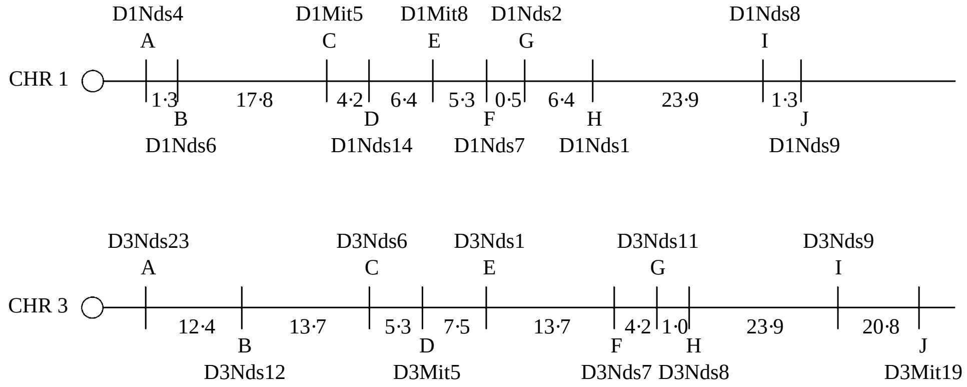

Fig. 2. Positions of loci typed on chromosomes 1 and 3 in the NOD backcross. Inter-marker distances are given in centiMorgans.

at the chromosome 1 markers and 287 animals typed at the chromosome 3 markers.

## 4. Statistical analysis of IDDM data

For the statistical analysis, the binary disease response was modelled using logistic regression, fitting a model of the form

$$\ln\frac{p}{1-p}=\alpha+\boldsymbol{\beta}^{T}\mathbf{x} \qquad (3)$$

(where $p$ is the probability of contracting the disease, $\mathbf{x}$ is the vector of binary genotypes at the loci, and $\alpha$ and the vector $\boldsymbol{\beta}$ coefficients to be estimated). Animals were then classified into eight response categories according to the degree of insulitis observed (diabetics being placed in the highest category), and the response modelled as a continuous variable assuming normal errors, as in (2), and then using the polychotomous logistic model (McCullagh & Nelder, 1989), which allows for the ordinal nature of the response through a series of parallel regressions that model the probability of being placed higher than the $j$th category as:

$$P(\text{histology} > j)=\frac{e^{\alpha_j+\boldsymbol{\beta}^{T}\mathbf{x}}}{1+e^{\alpha_j+\boldsymbol{B}^{T}\mathbf{x}}} \quad \text{for } j=0,1,\ldots,6$$

$$P(\text{histology} > 7)=0. \qquad (4)$$

Neither model is particularly proposed as the true genetic basis for diabetes and insulitis, but rather forms the parametric basis for a genetic-model-free regression strategy as described earlier.

The strategy for detection and modelling of the QTL effects was as follows: First the effect of including a single marker in the regression equation was investigated, for all markers (method 1). Next we moved across a chromosome in both directions looking at pairs of adjacent markers, and identifying those regions where the second marker is significant when added to the model including only the first marker, but a third marker is not significant when added to the model including only the second (i.e. regions showing the pattern of significance followed by non-significance described earlier). This we will denote method 2. Next we used a stepwise forward selection strategy (method 3) where at each stage three terms were entered simultaneously, corresponding to two neighbouring markers and their interaction, denoted $M_i^{**}M_{i+1}$ (equivalent to a model containing $M_i+M_{i+1}+M_i{*}M_{i+1}$) for markers $M_i$ and $M_{i+1}$. At each stage all possible new models corresponding to the previous model plus a term of the form $M_i^{**}M_{i+1}$ were considered, and that model giving the most significant reduction in deviance (i.e. the most significant increase in fit to the data) was selected, the interpretation being that inclusion of a term $M_i^{**}M_{i+1}$ implied the existence of a disease locus in the $–M_i––M_{i+1}–$ region. Finally, having determined the approximate locations of disease loci, the model with single-marker terms giving the best overall fit to the data was determined, and interactions between the markers were tested to see whether there was significant evidence for epistasis.

## 5. Results

The results for chromosome 1 were as follows. Using a binary logistic regression model, (3), only weak evidence ($P$ value $> 0{\cdot}02$, results not shown) was obtained for the significance of any loci. The results from using a normal errors model are given in Table 1. Adding single-locus terms to the mean gives a broad region of strong significance from A to F, with some positive evidence of linkage also seen at I and J. Using a stepwise approach this situation is separated into two effects on either side of locus H (i.e. in the –F–H and H–I– regions), since F is very significant when added to the model including H but E is much less significant when added to F, and I is very significant after H but J is not significant after I. There is also a borderline effect at B–C–. No terms were

Table 1. *Chromosome 1: significance of terms added stepwise to the regression model*

| | Normal errors model | | Polychotomous model | |
|---|---|---|---|---|
| Term | *F* statistic | *P* value | $\chi^2$ statistic | *P* value |
| A | 18·89 | 0·00002 | 16·89 | 0·00004 |
| B | 19·57 | 0·00001 | 17·28 | 0·00003 |
| C | 20·25 | 0·00001 | 17·68 | 0·00003 |
| D | 16·18 | 0·00007 | 14·02 | 0·0002 |
| E | 12·38 | 0·0005 | 10·09 | 0·0015 |
| F | 7·24 | 0·0075 | 5·83 | 0·0157 |
| H | 0·91 | — | 0·68 | — |
| I | 4·74 | 0·030 | 4·66 | 0·0308 |
| J | 4·58 | 0·033 | 4·42 | 0·0355 |
| B after A | 0·64 | — | 0·425 | — |
| C after B | 4·22 | 0·041 | 3·652 | 0·056 |
| D after C | 0·00 | — | 0·016 | — |
| E after D | 0·00 | — | 0·305 | — |
| F after E | 1·10 | — | 0·796 | — |
| H after F | 8·05 | 0·0050 | 7·351 | 0·007 |
| I after H | 12·54 | 0·0005 | 12·502 | 0·0004 |
| J after I | 0·00 | — | 0·002 | — |
| I after J | 0·15 | — | 0·242 | — |
| H after I | 8·62 | 0·0036 | 8·517 | 0·004 |
| F after H | 14·52 | 0·00017 | 12·507 | 0·0004 |
| E after F | 6·11 | 0·014 | 5·054 | 0·025 |
| D after E | 3·64 | — | 4·234 | 0·040 |
| C after D | 3·84 | — | 3·680 | 0·055 |
| B after C | 3·57 | — | 3·244 | 0·072 |
| A after B | 0·00 | — | 0·035 | — |

Results are given for the significance of including terms either individually, or after neighbouring terms have already been included, in a normal errors model (2) or a polychotomous logistic model (4).

significant when deleted from the full main effects model. Entering neighbouring terms and interactions in a forward stepwise manner gave significance for C**D ($F = 8{\cdot}64$, $P$ value $1{\cdot}6 \times 10^{-5}$) followed by I**J ($F = 5{\cdot}15$, $P$ values 0·0019). These results are almost identical (in terms of significance) to those obtained using the polychotomous logistic regression model given by (4).

For chromosome 3, the significance of adding individual markers to various models is shown in Table 2. For the binary logistic model, adding terms to the mean, a broad region of significance is seen across the chromosome, most notably in the D–F region. Taking a stepwise approach, there is evidence for the presence of disease loci in the –H–I and D–E– regions. When deleted from the full main effects model locus E is significant ($F = 14{\cdot}01$, $P$ value 0·0017) with borderline significance at B and G ($P$ values $\approx 0{\cdot}07$). Entering neighbouring terms and interactions in a forward stepwise manner gave significance for D**E ($F = 13{\cdot}69$, $P$ value $8{\cdot}4 \times 10^{-6}$) followed by B**C ($F = 4{\cdot}22$, $P$ value 0·014). Using a normal errors model, with histology as the dependent variable, all loci are highly significant when added to the mean. Using a stepwise approach we see significant effects located at A–B–, D–E–, –H–I and –E–F, since B is significant when added after A but C is not significant when added after B, and similarly for E after D, H after I and E after F. When deleted from the full main effects model, loci B ($F = 4{\cdot}04$, $P$ value 0·045) and E ($F = 4{\cdot}96$, $P$ value 0·027) are significant. Entering neighbouring terms and interactions in a forward stepwise manner gave high significance for D**E ($F = 23{\cdot}67$, $P$ value $< 5 \times 10^{-8}$) followed by A**B ($F = 3{\cdot}53$, $P$ value 0·015). These results are again extremely similar to those obtained using the polychotomous logistic regression model, although for the polychotomous model there is additional borderline significance in the region –B–C.

As might be expected, the results of the analyses above are considerably clearer (in terms of patterns and degrees of significance) for models that consider the phenotype as a quantitative variable rather than as a binary trait. It may, however, not always be possible to use the quantitative models, either because quantitative data are not available or because the quantitative measure may not be an appropriate measure of disease status. It is interesting that the normal error models and the polychotomous models produce very similar results. This illustrates the fact that we are not really attempting to parameterize the disease model by the regression equation, but rather hope to use the regression results as a non-parametric or model-free method to identify and position disease loci.

For chromosome 1, taking the analyses as a whole, there are two regions of linkage detected. These are difficult to localize precisely, but appear to be linked to *D1Mit5* and *D1Nds1* (markers C and H respectively). There is some evidence for the existence of two effects, on either side of H (*D1Nds1*). When entered in a forward stepwise manner, the markers that produced the best fit to the data (assuming the normal errors model) were C, I and E (*D1Mit5*, *D1Nds8* and *D1Mit8*), supporting the hypothesis of two separate effects on either side of *D1Nds1*. This is broadly supported by the results in Table 1, although one might have expected to see significance of B after C, F after E and I after J, if the effects are real. However, non-significance is a less conclusive result than significance (or a pattern of significance followed by non-significance) as one would not necessarily see significance of neighbouring terms included after C and E if the QTLs are tightly linked to markers C and E. In addition our data for markers I and J were highly correlated, with only 4 of the 296 animals typed being recombinant between these loci, which would explain why neither I nor J is significant once the other has been entered. Having fitted terms for loci C, I and E, the interaction term C*E was also significant ($F =$

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

Table 2. *Chromosome 3: significance of terms added stepwise to the regression model*

| | Binomial errors model | | Normal errors model | | Polychotomous model | |
|---|---|---|---|---|---|---|
| Term | *F* statistic | *P* value | *F* statistic | *P* value | $\chi^2$ statistic | *P* value |
| A | 5·63 | 0·024 | 26·43 | $5 \times 10^{-7}$ | 23·86 | $1 \times 10^{-6}$ |
| B | 10·04 | 0·0034 | 40·05 | $< 5 \times 10^{-8}$ | 35·15 | $< 5 \times 10^{-8}$ |
| C | 9·54 | 0·0041 | 40·43 | $< 5 \times 10^{-8}$ | 34·58 | $< 5 \times 10^{-8}$ |
| D | 13·52 | 0·00086 | 52·24 | $< 5 \times 10^{-8}$ | 44·43 | $< 5 \times 10^{-8}$ |
| E | 39·10 | $5 \times 10^{-7}$ | 65·60 | $< 5 \times 10^{-8}$ | 56·96 | $< 5 \times 10^{-8}$ |
| F | 18·45 | 0·00015 | 51·83 | $< 5 \times 10^{-8}$ | 46·21 | $< 5 \times 10^{-8}$ |
| G | 11·73 | 0·0017 | 46·03 | $< 5 \times 10^{-8}$ | 40·28 | $< 5 \times 10^{-8}$ |
| H | 13·53 | 0·00086 | 45·64 | $< 5 \times 10^{-8}$ | 40·49 | $< 5 \times 10^{-8}$ |
| I | 6·25 | 0·018 | 20·12 | $1 \times 10^{-5}$ | 19·57 | $9{\cdot}7 \times 10^{-6}$ |
| B after A | 3·84 | 0·059 | 12·59 | 0·00045 | 11·333 | 0·001 |
| C after B | 0·54 | — | 3·01 | 0·09 | 3·030 | 0·082 |
| D after C | 3·00 | 0·093 | 10·49 | 0·0013 | 9·997 | 0·002 |
| E after D | 17·61 | 0·00021 | 14·25 | 0·0002 | 13·635 | 0·00022 |
| F after E | 0·29 | — | 3·07 | 0·08 | 2·557 | — |
| G after F | 0·70 | — | 0·17 | — | 0·082 | — |
| H after G | 1·30 | 0·08 | 0·85 | — | 1·186 | — |
| I after H | 0·22 | — | 0·00 | — | 0·098 | — |
| H after I | 6·16 | 0·019 | 23·75 | $1{\cdot}8 \times 10^{-6}$ | 21·023 | $4{\cdot}5 \times 10^{-6}$ |
| G after H | 0·03 | — | 1·18 | — | 0·972 | — |
| F after G | 5·57 | 0·025 | 5·15 | 0·024 | 6·011 | 0·014 |
| E after F | 13·10 | 0·001 | 14·80 | 0·00015 | 13·308 | 0·00026 |
| D after E | 0·11 | — | 2·89 | — | 1·104 | — |
| C after D | 0·03 | — | 0·17 | — | 0·147 | — |
| B after C | 0·92 | — | 2·68 | — | 3·605 | 0·058 |
| A after B | 0·18 | — | 0·17 | — | 0·046 | — |

Results are given for the significance of including terms either individually, or after neighbouring terms have already been included, in a normal errors model (2) or a polychotomous logistic model (4).

Table 3. *Coefficients of chromosome 1 interaction model*

| Term | Coefficient | Standard error | *F* to enter | *P* value |
|---|---|---|---|---|
| Mean | $\beta_0 = 4{\cdot}105$ | 0·2343 | | |
| C | $\beta_1 = -0{\cdot}1355$ | 0·6829 | 20·25 | $9{\cdot}8 \times 10^{-6}$ |
| I | $\beta_2 = -1{\cdot}190$ | 0·3111 | 12·49 | $4{\cdot}75 \times 10^{-4}$ |
| E | $\beta_3 = -0{\cdot}7713$ | 0·5858 | 2·50 | 0·11 |
| C*E | $\beta_4 = 1{\cdot}987$ | 0·8820 | 5·08 | 0·025 |

The coefficients are the estimated regression coefficients $\beta_0$, $\beta_1, \beta_2, \beta_3, \beta_4$ obtained from fitting the model of (2) with $x_1 = 1$ for a homozygote and 0 for a heterozygote at marker C, $x_2 = 1$ for a homozygote and 0 for a heterozygote at marker I, $x_3 = 1$ for a homozygote and 0 for a heterozygote at marker E and $x_4 = 1$ for an animal homozygous at both C and E, and 0 otherwise. Also given are the standard errors of the regression coefficients and the significance in terms of an *F* statistic (*F* to enter) for the significance of adding the term to the regression equation in the order given.

5·08, *P* value 0·025), suggesting that there may be an epistatic interaction between C and E (*D1Mit5* and *D1Mit8*). The coefficients for this model are given in Table 3. Here the data have been coded so that the coefficients are given for homozygotes, so that a positive coefficient indicates an NOD susceptibility, or B10 protective, locus. We see that homozygosity at either C or E produces a slight decrease in histology score if an animal is heterozygous at the other locus, but a stronger increase in histology score if the animal is homozygous at the other locus, due to the interaction. This resembles an epistatic model at C and E (*D1Mit5* and *D1Mit8*), where both disease loci are required for disease expression. The coefficient for I (*D1Nds8*) indicates a homozygous protective effect at this locus. It is interesting to note that our results for chromosome 1 correspond quite well to those effects found at G and A (*D1Nds2* and *D1Nds4*) by Garchon *et al.* (1994). Our results are also consistent with those of Cornall *et al.* (1991) in which the susceptibility gene *Idd5* was mapped to a broad region of chromosome 1 distal to *D1Nds4* (marker A).

For chromosome 3, the most significant effect is seen in the D–F region, closely linked to *D3Nds1* (marker E). There is also a strong effect linked to *D3Nds12* (marker B). Another strong effect, though only seen using method 2, occurs in the H–I region. It is interesting to compare these results with those obtained using experiments in congenic mice (Wicker *et al.*, 1995). *D3Nds1* is a microsatellite locus adjacent to the *Il2* gene which is a strong candidate for *Idd3*. From congenic analysis, the *Idd3* interval has now been fine-mapped to only 0·35 cM and contains the

Table 4. *Coefficients of the chromosome 3 interaction model*

| Term | Coefficient | Standard error | $F$ to enter | $P$ value |
|---|---|---|---|---|
| Mean | $\beta_0 = 2{\cdot}315$ | 0·2605 | | |
| B | $\beta_1 = 1{\cdot}547$ | 0·4698 | 65·74 | $< 1 \times 10^{-7}$ |
| H | $\beta_2 = 1{\cdot}472$ | 0·4944 | 6·03 | 0·015 |
| E | $\beta_3 = 1{\cdot}068$ | 0·4466 | 5·46 | 0·020 |
| B*H | $\beta_4 = -1{\cdot}058$ | 0·5778 | 3·36 | 0·068 |

The coefficients are the estimated regression coefficients $\beta_0$, $\beta_1, \beta_2, \beta_3, \beta_4$ obtained from fitting the model of (4) with $x_1 = 1$ for a homozygote and 0 for a heterozygote at marker B, $x_2 = 1$ for a homozygote and 0 for a heterozygote at marker H, $x_3 = 1$ for a homozygote and 0 for a heterozygote at marker E and $x_4 = 1$ for an animal homozygous at both B and H, and 0 otherwise. Also given are the standard errors of the regression coefficients and the significance in terms of an $F$ statistic ($F$ to enter) for the significance of adding the term to the regression equation in the order given.

*D3Nds12* locus and the *Il2* gene (Denny *et al.*, 1997). In addition the disease locus *Idd10* an be located at *D3Nds8*, or marker H. These results locating *Idd3* and *Idd10* at *D3Nds12* and *D3Nds8* (B and H) fit very well with our results using the regression approach. Homozygosity at markers B, E and H produced an increase in histology score in each case. The only interaction term found between markers B, E and H was B*H ($F = 3{\cdot}36$, $P$ value 0·068), which suggests a possible interaction between *Idd3* and *Idd10*, although this is not significant. The coefficients of the interaction model are given in Table 4. We see that being heterozygous at *Idd3* and *Idd10* gives an increase in liability of zero (ignoring the mean and the effect at E), being homozygous at *Idd3* and heterozygous at *Idd10* gives an increase in liability of 1·547, being heterozygous at *Idd3* and homozygous at *Idd10* gives an increase in liability of 1·472, while being homozygous at both *Idd3* and *Idd10* gives an increase in liability of $1{\cdot}547 + 1{\cdot}472 - 1{\cdot}058 = 1{\cdot}961$. Taking homozygosity at the two loci as the baseline, this implies a reduction in liability of 0·414 for heterozygotes at *Idd3*, 0·489 for heterozygotes at *Idd10* and 1·961 for heterozygotes at both loci. This is consistent with the results described in Wicker *et al.* (1995), the epistatic interaction resulting in a much greater reduction in liability when resistant alleles are present at both *Idd3* and *Idd10* than that expected if the loci were acting independently.

The effect linked to E that we detected on chromosome 3 did not appear to correspond to any previously defined IDDM locus. However, the stepwise regression approach of method 2 should imply that this marker is showing a true effect even after the effects of *Idd3* and *Idd10* have been accounted for. In order to examine the effect of marker E on disease, while taking account of any effects at *Idd3* and *Idd10*, the relationship between genotype (homozygous or heterozygous) at locus E and disease status (diabetic or control) for animals whose genotype at markers B and H was fixed in advance was examined. The results are shown in Table 5. For animals heterozygous at both B and H, almost all are also heterozygous at E (as would be expected from the proximity of these markers), and there is no relationship between genotype at E and disease status. Similarly for animals homozygous at B and H, most are homozygous at E with no relationship between genotype at E and disease. For animals homozygous at B and heterozygous at H, there is some increase in the proportion of controls amongst the heterozygotes at E, but this is not statistically significant ($P$ value 0·076, Fisher's exact test). But for animals heterozygous at B and homozygous at H, there is seen to be a significant relationship between genotype at E and disease status ($P$ value 0·0019, Fisher's exact test), with homozygotes at E having a much greater probability of developing diabetes than heterozygotes. This is the effect that is being detected by our stepwise regression procedure. The third putative diabetes locus is having an important effect in those animals heterozygous at *Idd3* and homozygous at *Idd10*. Indeed there is evidence from a chromosome 3 congenic strain that a third locus may exist between *Idd3* and *Idd10* (L. Wicker and L. Peterson, unpublished data).

Table 5. *Contingency table analysis of markers B, H and E on chromosome 3*

| | He at B | | | | Ho at B | | | |
|---|---|---|---|---|---|---|---|---|
| | He at H | | Ho at H | | He at H | | Ho at H | |
| Status | He at E | Ho at E | He at E | Ho at E | He at E | Ho at E | He at E | Ho at E |
| DIAB | 9 | 0 | 2 | 18 | 6 | 10 | 0 | 57 |
| CON | 68 | 2 | 17 | 16 | 19 | 12 | 1 | 50 |

Frequencies are given for numbers of animals heterozygous (He) and homozygous (Ho) at the three markers amongst diabetics (DIAB) and controls (CON).

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

## 6. Analysis using the method of Zeng (1993)

It is of interest to compare the results from our regression analyses with those obtained using the method of Zeng (1993, 1994). A suite of programs (QTL Cartographer) to implement this method was recently made available (Basten *et al.*, 1994). We therefore additionally analysed our backcross data using these programs. Three analysis models were considered for the specification of the markers to be included as controls for the genetic background: model 1 from Zeng (1993), which uses all markers to control for the genetic background; model 3, which uses no markers to control for the genetic background and is thus equivalent to interval mapping as described in Lander & Botstein (1989); and model 6, which performs a forward stepwise regression procedure to choose the most important markers to control for the genetic background, dependent on user-specified parameters for the maximum number of such markers (nmp) and the window size (ws) for blocking out a region of the genome on either side of the markers flanking the test site. The recommended analysis models are models 3 and 6, although the optimum choice of parameters for model 6 is not yet clear, which is a problem. Also the QTL cartographer method relies on knowing or estimating genetic map distances between markers, which makes it more restricted than the regression approach.

The results of these analyses, in terms of likelihood profiles for likelihood ratio statistics, are shown in Fig. 3. A question of some importance is the

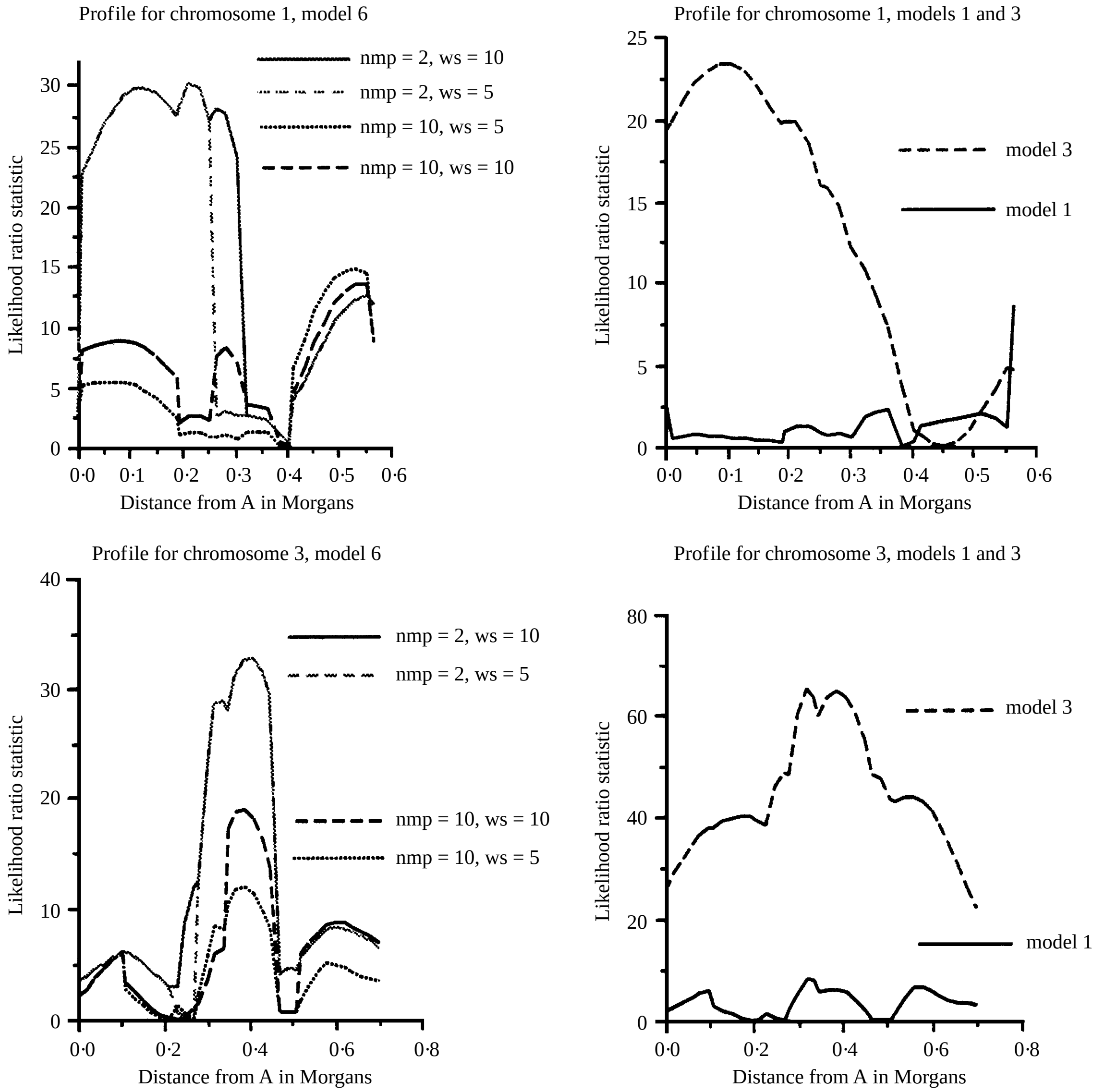

Fig. 3. Likelihood profiles for QTL Cartographer analysis of chromosomes 1 and 3. Results are given for model 6 with various values of the parameters ws (window size) and nmp (maximum number of markers controlling for genetic background), and for models 1 and 3. Model 1 uses all markers to control for genetic background, model 3 uses no markers to control for genetic background and model 6 chooses markers to control for genetic background using a forward stepwise regression procedure.

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

designation of the appropriate thresholds for declaring the presence of a QTL in an interval. From theoretical considerations and simulation, for model 1, Zeng (1994) recommends using a critical value of $\chi^2_2(\alpha/M)$ to give an approximate overall type 1 error rate of $\alpha$, where $M$ is the number of intervals tested. This would correspond to a threshold of 10·15 (for $\alpha = 0{\cdot}05$) in our analyses, with slightly smaller critical values suggested by simulation (Zeng, 1994) for models 3 and 6. These thresholds are, however, calculated for the hypothesis that no QTLs are present, and it is not clear how they relate to the error rate for detecting three QTLs when in fact only two are present, or to other possible scenarios. (Such error rates can be calculated with QTL Cartographer using bootstrap techniques but only by assuming the specific number of QTLs present and conditioning on their position.) It is therefore important to use this threshold merely as a guide to interpreting significance. We can gain some useful information concerning the position of possible QTLs simply by examining the shape of the likelihood profile over a chromosome. Some differences are seen between the various models, but broadly speaking the results are very similar to those from our non-parametric regression procedure. For chromosome 1, the effects distal to marker H (*D1Nds1* at a distance of 0·38) and at marker C (*D1Mit5* at a distance of 0·16) are seen quite clearly. The effect proximal to H, linked to marker E (*D1Mit8*), is also visible in some cases when a large number of markers are included (e.g. model 6, nmp = 10, ws = 10). For chromosome 3, although the effect at E (*D3Nds1* at a distance of 0·35) is clearly the most significant, the profiles also indicate the presence of *Idd3* and *Idd10* at distances 0·1 and 0·6. We can therefore conclude that the model-free stepwise regression procedures and the parametric models of Zeng (1993, 1994) produce very similar results, as may be expected from the similarities between the two approaches.

## 7. Simulation study

To investigate further the power and properties of the regression strategies used here, we conducted a simulation study using 1000 replicates. Four chromosomes each with 16 markers spaced at 10 cM intervals were simulated for a backcross population of sample size 300. The trait of an individual was assumed to be affected by 10 QTLs with positions and effects as given in Fig. 4. The trait value of an individual was assumed to be the sum of the effects of the QTLs possessed, plus a random environmental variable that was normally distributed with mean zero and variance scaled to give 0·7 heritability in the population. This model was chosen to be identical to that simulated by Zeng (1994), allowing easy comparison of the powers of the two approaches.

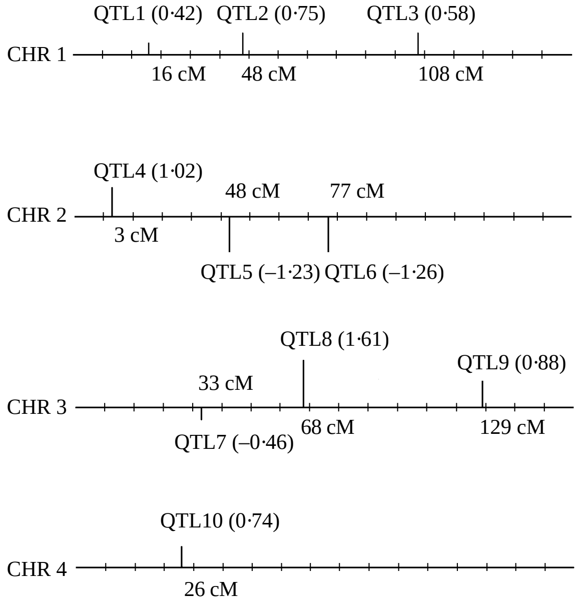

Fig. 4. Positions of markers and QTLs on the four simulated chromosomes, with marker spacing 10 cM. Distances are given from the first marker on a chromosome. The effect of a QTL is shown in brackets and by its magnitude and direction.

Each chromosome was considered to be divided into 17 chromosomal regions, with the distance between each pair of markers being divided into two ‘half regions’ for the purposes of detecting a QTL at that location. The data generated were analysed using four stepwise regression strategies. First the effect of including a single marker in the regression equation was investigated, across all markers (method 1). If a marker was significant we designated that as a detection of a QTL in the immediate vicinity of that marker (i.e. in the chromosomal region from halfway between the marker and its distal neighbouring marker, to halfway between the marker and its proximal neighbour). This method is not an interval method, but rather corresponds to traditional QTL mapping in an experimental cross. Next we used a stepwise strategy considering all pairs of adjacent markers (method 2 as described in Section 3) and moving across the chromosome looking at the pattern of significance and non-significance. For the purposes of the simulation study two versions of method 2 were considered: for method 2 (EITH) a chromosomal region was classified as containing a QTL if method 2 gave significance at that location when moving in either direction (left to right or right to left); for method 2 (BOTH) a region was only assumed to contain a QTL if method 2 gave significance in both directions. Method 3 corresponded to the stepwise forward selection strategy described in Section 3, where inclusion of a term $M_i^{**}M_{i+1}$ implied the existence of a disease locus in the $–M_i––M_{i+1}–$ region.

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

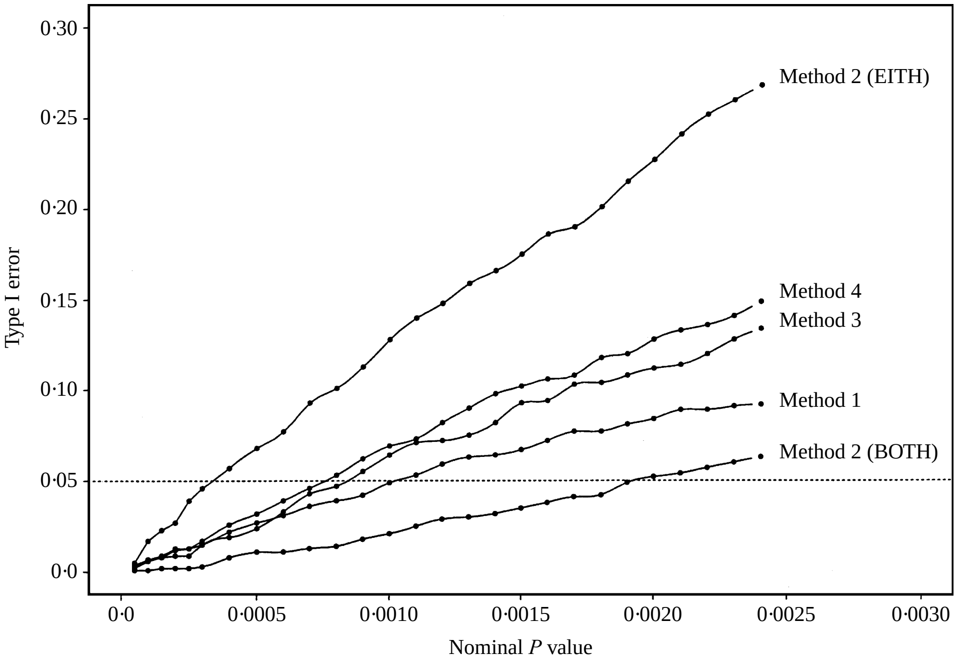

Fig. 5. Graphs of the 'genome-wide' (for four simulated chromosomes) type I errors plotted against the nominal *P* value for a single test in each of the different regression strategies. The horizontal from a type I error of 0·05 intersects the graphs to give the nominal *P* values required to give a 'true' *P* value of 0·05.

This model allows for epistasis between QTLs in a region (Cordell, 1995). Since at each stage in this strategy a large number of possible next models must be considered, the computing burden for the simulations was reduced by including a maximum of seven such terms, allowing the possible detection of seven major QTLs, but no more. Finally, method 4 was a deletion strategy, similar to method 1 but with single markers being deleted from the full model (2), and if a marker was significant we considered that to be a detection of a QTL in the immediate vicinity of the marker.

To compare the powers of the various methods, it was necessary to ensure that each method had the same type I error rate or 'true' *P* value. A nominal *P* value for each test performed in a method is given by the significance of entering or deleting the term in the regression equation, but it is not clear how that relates to the overall type I error rate for the different methods. We therefore initially simulated backcross data under the null hypothesis of no QTL effects, with the marker data simulated as described above and the quantitative trait simulated simply as a normal random variable, with variance scaled to give the same overall variance as in the QTL backcross population. Fig. 5 shows the simulated *P* values or type I errors for the different methods as a function of the nominal *P* value used for each test. We see that nominal *P* values of 0·001 (method 1), 0·00035 (method 2 (EITH)), 0·0019 (method 2 (BOTH)), 0·00085 (method 3) and 0·00075 (method 4) correspond to a genome-wide (the 'genome' here being defined as the four chromosomes simulated) type I error of about 0·05. These nominal *P* values were therefore used for each test performed as part of the appropriate regression method, when evaluating the power of the different methods. In addition it was of interest to see what effect different numbers of markers, or different marker spacing, had on the type I error. To investigate this and also to simulate situations more relevant to our IDDM data, we also simulated data under the null hypothesis of no QTL effects for a backcross of 300 individuals with a single chromosome consisting of between 4 and 16 markers, and inter-marker spacings of 5, 10, 15 and 20 cM. The results for method 2 (EITH) and method 3 are shown in Fig. 6, for a nominal *P* value of 0·01 for each test. As expected, the overall type I error increases with increasing numbers of markers (and therefore numbers of tests performed). There is no simple relationship between type I error and marker spacing. The results for all methods for a chromosome with 10 markers spaced at 10 cM intervals are shown in Fig. 7. This corresponds most closely to the chromosomes we analysed for the IDDM data. We can see that, depending on the method used, a nominal *P* value of at most 0·0025 for an individual test will be sufficient to give an overall chromosome-wide type I error of 0·05. Many of our tests in Section 3 met this criterion, and so we can consider those tests to be truly significant even taking into account the problem of multiple (albeit not independent) tests.

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

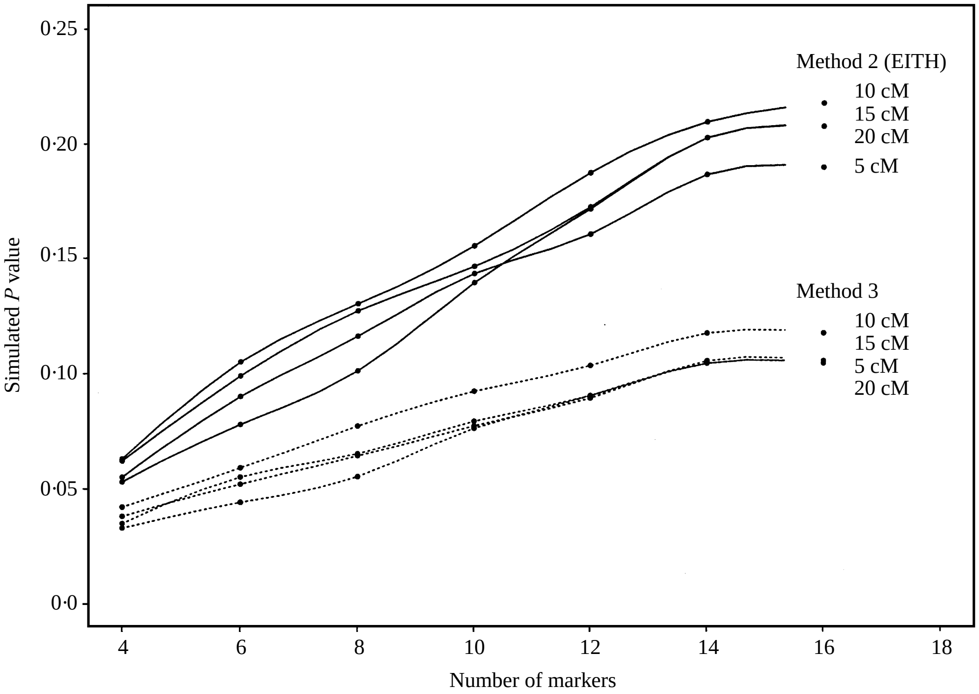

Fig. 6. Graphs of the simulated *P* values for a single chromosome with different numbers of markers and marker spacings. Results are shown for method 2 (EITH) and method 3, using a nominal *P* value for each test of 0·01.

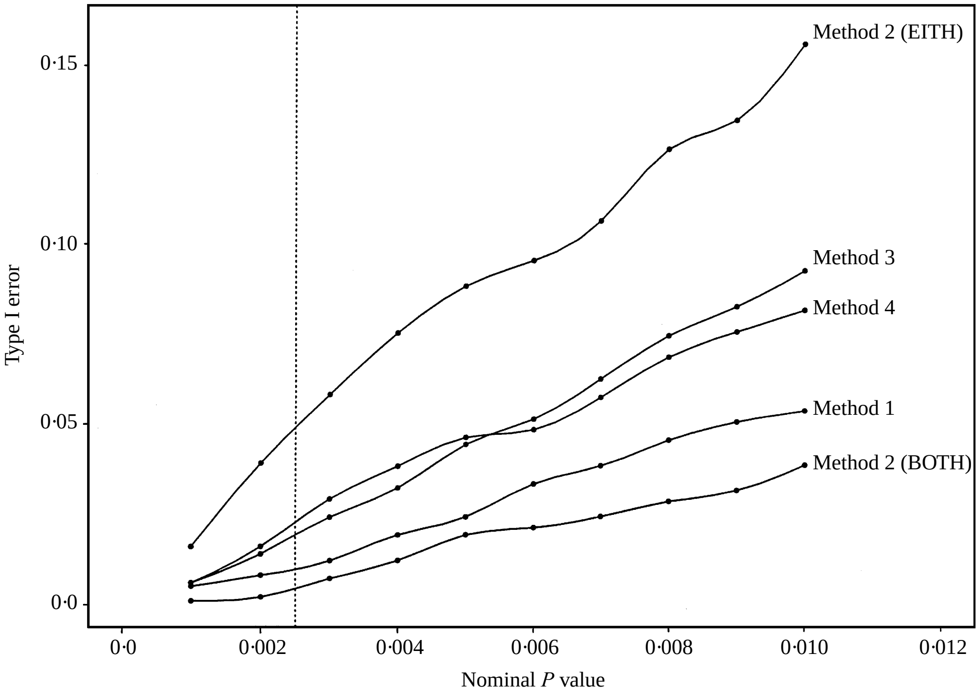

Fig. 7. Graphs of the simulated *P* values for a single chromosome with 10 markers spaced at 10 cM intervals. A nominal *P* value of 0·0025 is sufficient to ensure the simulated *P* value is less than 0·05 for all methods.

The results from the power simulations are shown in Table 6. Also given are the results from Zeng (1994) for the power of his interval method (model I) and non-interval methods (models II and III) to detect the same QTLs, although note that the methods of Zeng produce a statistic at an exact location, rather than our more imprecise positioning that locates a QTL either in the correct half, or in either half, of an inter-marker interval. We see that the traditional method (method 1), like models II and III of Zeng (1994), gives high power to detect the major QTLs, but, as noted previously, none of these methods is an interval method and so they do not necessarily have a high probability of locating a given QTL accurately. Of our

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

Table 6. *Powers to detect QTLs using regression methods from 1000 replicates of simulation, and comparison with powers from Zeng* (*1994*) (*based on 100 replicates*)

| | Chromosome 1 | | | Chromosome 2 | | | Chromosome 3 | | | Chromosome 4 |
|---|---|---|---|---|---|---|---|---|---|---|
| QTL: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Position (cM): | 16 | 48 | 108 | 3 | 43 | 77 | 33 | 68 | 129 | 26 |
| Effect: | 0·42 | 0·75 | 0·58 | 1·02 | −1·23 | −1·26 | −0·46 | 1·61 | 0·88 | 0·74 |
| Powers for detection in correct half of inter-marker interval using regression strategy | | | | | | | | | | |
| Method 1 | 0·851 | 0·981 | 0·742 | 0·006 | 0·997 | 1·0 | 0·104 | 0·998 | 1·0 | 0·442 |
| Method 2 (EITH) | 0·077 | 0·222 | 0·112 | 0·065 | 0·884 | 0·845 | 0·083 | 0·946 | 0·164 | 0·052 |
| Method 2 (BOTH) | 0·004 | 0·099 | 0·021 | 0·035 | 0.027 | 0·0299 | 0·004 | 0·0 | 0·022 | 0·005 |
| Method 3 | 0·116 | 0·761 | 0·221 | 0·701 | 0·638 | 0·809 | 0·031 | 0·409 | 0·464 | 0·552 |
| Method 4 | 0·010 | 0·083 | 0·033 | 0·289 | 0·247 | 0·250 | 0·009 | 0·008 | 0·001 | 0·027 |
| Powers for detection in either half of correct inter-marker interval using regression strategy | | | | | | | | | | |
| Method 1 | 0·879 | 0·989 | 0·810 | 0·009 | 1·0 | 1·0 | 0·440 | 1·0 | 1·0 | 0·525 |
| Method 2 (EITH) | 0·162 | 0·273 | 0·121 | 0·190 | 0·885 | 0·880 | 0·094 | 0·958 | 0·531 | 0·078 |
| Method 2 (BOTH) | 0·004 | 0·112 | 0·022 | 0·035 | 0·120 | 0·453 | 0·004 | 0·672 | 0·192 | 0·006 |
| Method 3 | 0·125 | 0·851 | 0·259 | 0·809 | 0·754 | 0·967 | 0·041 | 0·986 | 0·847 | 0·687 |
| Method 4 | 0·012 | 0·083 | 0·034 | 0·299 | 0·299 | 0·264 | 0·014 | 0·711 | 0·181 | 0·034 |
| Powers for detection from Zeng (1994) (100 replicates) | | | | | | | | | | |
| Model I | 0·0 | 0·22 | 0·0 | 0·87 | 0·83 | 0·80 | 0·0 | 0·99 | 0·58 | 0·17 |
| Model II | 0·0 | 1·0 | 0·0 | 0·0 | 1·0 | 1·0 | 0·0 | 1·0 | 1·0 | 0·99 |
| Model III | 0·0 | 1·0 | 0·0 | 0·0 | 1·0 | 1·0 | 0·0 | 1·0 | 1·0 | 0·58 |

interval methods, only method 2 (EITH) and method 3 provide powers comparable to those of model I of Zeng (1994). Method 2 (EITH) is seen to be slightly more powerful than model I of Zeng for the detection of QTLs 1, 3, 5, 6 and 7, of comparable power for QTLs 2 and 8, and less powerful for QTLs 4, 9 and 10. For QTL 4, in particular, the power is very low, perhaps because this QTL is located at the end of a chromosome, which makes the location strategy of method 2 difficult to implement. For method 3, the power is comparable and in some cases quite a bit larger than that of model I of Zeng (1994), except for detection of QTL 8.

The purpose of these comparisons is not to present our regression approach as superior to the approaches of Zeng (1994) and others (Moreno-Gonzalez, 1992; Jansen, 1994), but rather to show that in many cases the powers to detect a QTL in a correct location are not reduced using this more naive strategy. Our methods are unlikely to produce as accurate a map of QTL position as the more complicated approaches, but they do have the advantage of being computationally extremely simple and easily implemented in any standard statistical package. Such methods may therefore be useful as a first step in an analysis of data from an experimental cross, before proceeding to a more complicated analysis that may require careful consideration of the likely numbers and positions of QTLs in order to estimate significance levels accurately. Clearly further investigation of the regression approach will be required to determine the situations where it may be expected to be more, or less, powerful than other methods. Current research is being undertaken to investigate in more detail the theoretical properties of these methods in both experimental and human populations (Cordell, 1995) and will form the topic of further communications.

## 8. Discussion

The regression methods described here use a model-free approach to infer the vicinity of disease loci in relation to a fixed set of markers (although the specific models fitted are parametric in terms of the error distribution and linear predictors, the inference made from them is not). Depending on the marker spacing this positioning may be less precise than in the methods described by Zeng (1993) and by Jansen (1994). Advantages of our approach are that it is computationally extremely simple, is easily implemented in any standard statistical package, and the strategy is conceptually easier to understand than the 'black box' approach provided by some more complicated methods. In addition, no specific assumptions regarding interference are made, a straightforward test of epistasis is provided, and the methods generalize immediately to dichotomous or polychotomous traits via logistic rather than least-squares regression.

One problem with the regression strategies described here is that the non-parametric interpretation of the results can make it difficult to assign appropriate significance levels, other than by simulation. From the simulations described here and elsewhere (Cordell, 1995) we found that the true error rate depends critically on the number and magnitude of effects of disease loci present, and on the exact regression

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

method used. However, it is true to say that the simulated error rates were in almost all cases larger than the nominal $P$ values. The difference between nominal and simulated type I errors is likely to result partly from the multiple testing problem, where many significance tests, albeit not independent ones, are performed, and partly from the lack of precise correspondence between testing the hypothesis that a coefficient in a regression equation is equal to zero, and testing what is in some cases a somewhat complex linkage hypothesis, e.g. whether three disease loci are present when two have already been detected, or whether there is a disease locus present but not in the region where it was detected. Without knowing the exact scheme of testing used, the form of the phenotypic data, the number of markers available, the number and severity of the disease loci and the degree of epistasis, all of which affect the final type I error rate, it is difficult to give a definitive recommendation as to the significance level required to produce a specific genome-wide type I error. We should therefore be wary of placing too much emphasis on the nominal $P$ values achieved. Nevertheless our results correspond well to those found using congenic strain analysis, giving us some confidence that the effects we have detected are true ones, and that there are at least two and possibly three diabetogenic loci on each of chromosomes 1 and 3, with significant epistasis.

The problem of assigning appropriate significance levels is not new, nor limited to the regression methods proposed here. For any non-parametric analyses, the correspondence between the parametric $P$ value and the true false positive rate when declaring linkage, is not exact. This provides some of the motivation for the much more stringent requirements for declaring genetic linkage than for most statistical tests. From arguments based on an Orenstein–Uhlenbeck diffusion process, $P$ values as well as $2 \times 10^{-5}$ have been suggested (Lander & Botstein, 1989, 1994; Churchill & Doerge, 1994; Lander & Schork, 1994) as giving realistic false positive error rates for likelihood-based methods in human data, but this is very conservative because it assumes that all significant tests are type I errors. Recently, Churchill & Doerge (1994) and Doerge & Churchill (1996) presented a permutation method for assigning empirical threshold values to any given test of linkage to a QTL. However, their method is not directly applicable to our analyses as it relies on testing a null hypothesis of either no QTL effects in a region, or of specific (known) QTL effects according to which the data may be stratified. Also the power to detect multiple linked QTLs is low.

It may seem unrealistic to expect to achieve $P$ values of such stringency when diseases are caused by many loci, each with small effects: in our analyses only one of our stepwise results (H after I on chromosome 3) met this criterion. However, as in human linkage studies, the most conclusive way of determining true significance will be through replication. Results that occur in independent data sets, and are subsequently confirmed through congenic breeding experiments, will enable us to determine which of our initial findings are in fact true genetic effects. On this note, congenic strain analysis of mouse chromosome 3 in NOD diabetes indicates that there are indeed at least two, if not three, separate susceptibility regions, with significant epistasis.

We would like to give particular thanks to Robert Elston, Daniel Weeks, Anthony Davison and an anonymous reviewer for their helpful comments on this manuscript. Financial support was provided by the Wellcome Trust. H.C. held a Wellcome Trust Studentship in Mathematical Biology, and J.T. is a Wellcome Trust Principal Research Fellow. Current support for H.C. is provided by a National Institutes of Health grant GM-28356.

## References

Basten, C. J., Weir, B. S. & Zeng, Z.-B. (1994). Zmap: a QTL Cartographer. In *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software* (ed. C. Smith, J. S. Gavora, B. Benkel, J. Chesnais, W. Fairfull, J. P. Gibson, B. W. Kennedy & E. B. Burnside). Guelph, Ontario, Canada.

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Cordell, H. J. (1995). Statistical methods in the genetic analysis of type 1 diabetes. Unpublished D.Phil. thesis, Bodleian Library, University of Oxford, UK.

Cornall, R. J., Prins, J.-B., Todd, J. A., Pressey, A., DeLarato, N. H., Wicker, L. S. & Peterson, L. B. (1991). Type 1 diabetes in mice is linked to the interleukin-1 receptor and *Lsh/Ity/Bcg* genes on chromosome 1. *Nature* **353**, 262–265.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **B39**, 1–38.

Denny, P., Lord, C. J., Hill, N. J., Goy, J. V., Levy, E. R., Podolin, P. L., Peterson, L. B., Wicker, L. S., Todd, J. A. & Lyons, P. A. (1997). Mapping of the insulin-dependent diabetes locus, *Idd3*, to a 0·35 cM interval containing the *Interleukin-2* gene. *Diabetes* **46**, 693–700.

Doerge, R. W. & Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

Garchon, H.-J., Luan, J.-J., Eloy, L., Bedossa, P. & Bach, J.-F. (1994). Genetic analysis of immune dysfunction in non-obese diabetic (NOD) mice: mapping of a susceptibility locus close to the *Bcl-2* gene correlates with increased resistance of NOD T cells to apoptosis induction. *European Journal of Immunology* **24**, 380–384.

Ghosh, S., Palmer, S. M., Rodrigues, N. R., Cordell, H. J., Hearne, C. M., Cornall, R. J., Prins, J.-B., McShane, P., Lathrop, G., Peterson, L. B., Wicker, L. S. & Todd, J. A. (1993). Polygenic control of autoimmune diabetes in nonobese diabetic mice. *Nature Genetics* **4**, 404–409.

Jansen, R. C. & Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press

Kruglyak, L. & Lander, E. S. (1995). High-resolution genetic mapping of complex traits. *American Journal of Human Genetics* **56**, 1212–1223.

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

Lander, E. & Botstein, D. (1994). Corrigendum. *Genetics* **136**, 705.

Lander, E. S. & Schork, N. J. (1994). Genetic dissection of complex traits. *Science* **265**, 2037–2048

Lathrop, G. M., Lalouel, J. M., Julier, C. & Ott, J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics* **37**, 482–498.

Martinez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

McAleer, M. A., Reifsnyder, P., Palmer, S. M., Prochazka, M., Love, J. M., Copeman, J. B., Powell, E. E., Rodrigues, N. R., Prins, J.-B., Serreze, D. V., DeLarato, N. H., Wicker, L. S., Peterson, L. B., Schork, N. J., Todd, J. A. & Leiter, E. H. (1995). Crosses of NOD mice with the related NON strain: a polygenic model for type 1 diabetes. *Diabetes* **44**, 1186–1195.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Meng, X.-L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–268.

Moreno-Gonzalez, J. (1992). Genetic models to investigate additive and non-additive effects of marker-associated QTL using multiple regression methods. *Theoretical and Applied Genetics* **85**, 435–444.

O'Brien, S. J. (ed.) (1993). *Genetic Maps*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E. & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**, 721–726.

Soller, M., Brody, T. & Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**, 35–39.

Todd, J. A., Aitman, T. J., Cornall, R. J., Ghosh, S., Hall, J. R. S., Hearne, C. M., Knight, A. M., Love, J. M., McAleer, M. A., Prins, J.-B., Rodrigues, N., Lathrop, G. M., Pressey, A., DeLarato, N., Peterson, L. B. & Wicker, L. S. (1991). Genetic analysis of autoimmune type 1 diabetes mellitus in mice. *Nature* **351**, 542–547.

Wang, G. L., Mackill, D. J., Bonman, J. M., McCouch, S. R., Champoux, M. C. & Nelson, R. J. (1994). RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistant rice cultivar. *Genetics* **136**, 1421–1434.

Wicker, L. S., Todd, J. A. & Peterson, L. B. (1995). Genetic control of autoimmune diabetes in the NOD mouse. *Annual Review of Immunology* **13**, 179–200.

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the USA* **90**, 10972–10976.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

https://doi.org/10.1017/S0016672398003152 Published online by Cambridge University Press