Genet. Res., Camb. (1996), **68**, *pp.* 55–63 *With 3 text-figures*

Mapping QTLs for binary traits in backcross and F_2 populations

P. M. VISSCHER[1]*, C. S. HALEY[1] AND S. A. KNOTT[2]
[1] *Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland*
[2] *Institute of Cell, Animal, and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, Scotland*

(*Received 13 November 1995 and in revised form 19 February 1996*)

Summary

Mapping quantitative trait loci (QTLs) for binary traits in backcross and F_2 populations was investigated using stochastic stimulation. Data were analysed using either linear regression or a generalized linear model. Parameters which were varied in the simulations were the population size (200 and 500), heritability in the backcross or F_2 population (0·01, 0·05, 0·10), marker spacing (10 and 20 cM) and the incidence of the trait (0·50, 0·25, 0·10). The methods gave very similar results in terms of estimates of the QTL location and QTL effects and power of QTL detection, and it was concluded that in practice treating the zero–one data as continuous and using standard linear regression was efficient.

1. Introduction

Methods for mapping quantitative trait loci (QTLs) based on maximum likelihood (Lander & Botstein, 1989; Jansen, 1993, 1994; Zeng, 1993, 1994) or linear regression (Haley & Knott, 1992; Martinez & Curnow, 1992) usually assume that residual errors, i.e. residuals within QTL genotype classes, are normally distributed, although recently a general non-parametric method has been reported (Kruglyak & Lander, 1995). However, many traits of interest in human, plant or animal populations are not normally distributed. For example, survival and disease status are generally considered as binary traits, i.e. phenotypes are scored either 0 (absence) or 1 (presence). Theoretically, taking account of the distribution of residuals should enhance the power of detecting QTLs in experimental or field populations. Jansen (1992) presented a general mixture model for mapping QTLs which uses the distributional properties of the data by fitting a generalized linear model (GLM).

It has been shown that, at least for normally distributed traits in backcross and F_2 populations, linear regression (LRG) and maximum likelihood (ML) are very similar in terms of power and estimation (Haley & Knott, 1992). However, if data are not normally distributed but it is assumed they are, test statistics will not be distributed as multiples of χ^2 or F-ratios under the null hypothesis of no QTLs segregating, and the methods are likely to be less efficient.

Jansen (1992) gave an example for a trait in which the residuals were from an exponential distribution, and showed the increase in maximum likelihood by using knowledge about the appropriate distribution rather than assuming normality. However, Jansen also noted that the estimates of recombination rates and QTL effects were similar whether assuming normal or exponential distribution of residuals. The aim of this study was to investigate whether nonlinear regression methods are better, in terms of power and precision of parameter estimates, than LRG when traits are not normally distributed. In particular, we compared LRG with GLM (McCullagh & Nelder, 1989), using simulation of binary data in backcross (BC) and F_2 populations derived from inbred lines. We used the BC population to study the mapping of additive QTLs, and the F_2 population to investigate the methods of analysis when dominance effects are included.

* Corresponding author. Present address: Institute of Ecology and Resource Management, University of Edinburgh, West Mains Road, Edinburgh EH9 3JG, Scotland. Tel: +44 131 535 4052, fax: +44 131 667 2601.

2. Materials and methods

We assume that both the BC and F_2 experimental populations are derived from inbred lines. Genetic markers are fully informative and equally spaced. We assume Haldane's mapping function throughout.

The basis and theory of so-called threshold characters are well understood (e.g. Falconer & Mackay, 1996, chapter 18), and the binary trait in this study is a special (and simplest) case of a threshold character. An underlying unobservable continuous variable, often called liability, is assumed to affect the observed phenotype in a non-linear way. When the value of the continuous variable for an individual is below a certain threshold, the observed phenotype of the individual is 'normal' (in our study a score of 0), and if the value of the continuous variable is above the threshold, the phenotype is 'affected' (score of 1).

(i) *Backcross population*

Assume an additive model for a single QTL. QTL genotypes and their effects are:

Population	*Genotype*	*Value*
1	qq	$\mu-a$
2	QQ	$\mu+a$,
F_1	Qq	μ,
BC	Qq or QQ	μ or $\{\mu+a\}$.

Hence, we backcross to population 2. Let μ_c and a_c denote the overall mean and the QTL effect on an underlying continuous normal scale in residual (environmental) standard deviation units. Throughout this study we use an environmental variance ($\sigma^2_e(c)$) of unity. The heritability always refers to the underlying scale. On that scale:

$$h^2 = \sigma^2_a(c)/(\sigma^2_a(c)+\sigma^2_e(c)) = \sigma^2_a(c)/(\sigma^2_a(c)+1{\cdot}0),$$

with

$$\sigma^2_a(c) = a_c^2/4.$$

Therefore, the phenotypic variance on the continuous scale is the sum of the additive genetic and environmental variance ($=\sigma^2_a(c)+\sigma^2_e(c) = \sigma^2_a(c)+1.0$). The relationship between a_c and h^2 is shown in Table 1.

(a) *Threshold determination for binary trait*

Given the means and variances for the two genotype populations, and given the overall required incidence (P, i.e. the probability of an observation being 1 in the BC population), the threshold (T) was determined using the algorithm of Ducrocq and Quaas (1988). In Table 1 the thresholds are shown for different heritabilities and different incidences.

Table 1. *Relationship between parameters on the Normal and binary scale for backcross populations. T and a_c are in environmental standard deviation units*

h^2	P	a_c	T	$P(Qq)$	P(QQ)	a_{01}
0·01	0·50	0·2010	0·1005	0·4600	0·5400	0·0800
	0·25		0·7784	0·2182	0·2818	0·0637
	0·10		1·3885	0·0825	0·1175	0·0350
0·05	0·50	0·4588	0·2294	0·4093	0·5907	0·1815
	0·25		0·9218	0·1783	0·3217	0·1434
	0·10		1·5446	0·0612	0·1388	0·0776
0·10	0·50	0·6667	0·3333	0·3694	0·6306	0·2611
	0·25		1·0460	0·1478	0·3522	0·2045
	0·10		1·6857	0·0459	0·1541	0·1082

h^2 is the heritability, T is the threshold on the underlying scale corresponding to incidence P, and $P(Qq)$, $P(QQ)$ and $P(qq)$ are the incidences for QTL genotypes Qq, QQ and qq, respectively, a_c is the additive QTL effect on the underlying scale, and a_{01} is the additive QTL effect on the observed scale.

(b) *Transformation to other scales*

On the observed 0/1 scale, the population means for the two genotypes are

$$P(Qq) = 1-\Phi(T-\mu_c), \quad (1)$$

$$P(QQ) = 1-\Phi(T-\mu_c-a_c) \quad (2)$$

and

$$a_{01} = P(QQ)-P(Qq), \quad (3)$$

where Φ is the Normal cumulative density function and $P(Qq)$ and $P(QQ)$ are the incidences for genotypes Qq and QQ, respectively.

From 0/1 scale to underlying scale

We estimate μ_{01} and a_{01} (estimates of the mean and additive effect on the observed scale), and wish to estimate a_c:

$$E(\mu_{01}) = P(Qq),$$

$$E(a_{01}) = P(QQ)-p(Qq).$$

Using eqns (1) and (2).

$$E[\Phi^{-1}(1-\mu_{01})] = T-\mu_c,$$

$$E[\Phi^{-1}(1-\mu_{01}-a_{01})] = T-\mu_c-a_c,$$

and the parameter of interest, a_c, is obtained by difference:

$$a_c = \Phi^{-1}(1-\mu_{01})-\Phi^{-1}(1-\mu_{01}-a_{01}).$$

From GLM probit analyses to 0/1 scale

Estimates on the probit scale (subscript p) are related to the underlying Normal scale by

$$E(\mu_p) = (T-\mu_c),$$

$$E(a_p) = a_c.$$

https://doi.org/10.1017/S0016672300033887 Published online by Cambridge University Press

Table 2. *Relationship between parameters on the Normal and binary scale for* F_2 *populations, for a QTL with a heritability of 0·10. Models: additive (add;* $a_c = 0{\cdot}4714$, $d_c = 0$*), dominant (dom;* $a_c = 0{\cdot}4851$, $d_c = 0{\cdot}4851$*) and recessive (rec;* $a_c = 0{\cdot}4851$, $d_c = -0{\cdot}4851$*)*

Model	P	T	$P(qq)$	$P(Qq)$	$P(QQ)$	a_{01}	d_{01}
add	0·50	0·0	0·3187	0·5000	0·6813	0·1813	0·0
	0·25	0·7106	0·1184	0·2383	0·4050	0·1433	−0·0234
	0·10	1·3511	0·0342	0·0882	0·1894	0·0776	−0·0235
dom	0·50	0·2568	0·2291	0·5903	0·5903	0·1806	0·1806
	0·25	0·9821	0·0712	0·3096	0·3096	0·1192	0·1192
	0·10	1·6233	0·0175	0·1275	0·1275	0·0550	0·0550
rec	0·50	−0·2568	0·4097	0·4097	0·7709	0·1806	−0·1806
	0·25	0·4833	0·1664	0·1664	0·5005	0·1671	−0·1671
	0·10	1·1597	0·0500	0·0500	0·2500	0·1000	−0·1000

T is the threshold on the underlying scale corresponding to incidence P, and $P(Qq)$, $P(QQ)$ and $P(qq)$ are the incidences for QTL genotypes Qq, QQ and qq, respectively. a_{01} and d_{01} are the additive and dominance QTL effect on the observed scale.

We wish to estimate a_{01} from μ_p and a_p:

$$E[\Phi(\mu_p)] = 1 - P(Qq),$$

$$E[\Phi(\mu_p + a_p)] = 1 - P(QQ),$$

so

$$a_{01} = \Phi(\mu_p) - \Phi(\mu_p + a_p).$$

From GLM logit analyses to 0/1 scale

On the logistic scale (subscript g),

$$E[\exp(\mu_g)] = P(Qq)/(1 + P(Qq)),$$

$$E[\exp(\mu_g + a_g)] = P(QQ)/(1 + P(QQ)).$$

Hence,

$$P(Qq) = \exp(\mu_g)/[1 + \exp(\mu_g)],$$

$$P(QQ) = \exp(\mu_g + a_g)/[1 + \exp(\mu_g + a_g)],$$

$$a_{01} = P(QQ) - P(Qq).$$

The parameters on the logistic scale are difficult to compare directly with the underlying scale. To make a comparison possible, parameters on the logistic scale can be transformed to the underlying Normal scale using the approximation (e.g. Mood *et al.* 1974)

$$\mu_c = \mu_g/\sigma_g,$$

$$a_c = a_g/\sigma_g,$$

with

$$\sigma_g^2 = \pi^2/3.$$

Note that the transformation to obtain a_{01} was performed with the estimates on the logistic scale.

(ii) F_2 *population*

Extension from a BC to an F_2 population is relatively straightforward. We now have an additional genotype (qq) with effect $\{\mu_c - a_c\}$ on the underlying Normal scale, and the heterozygote may be different from the average value of the homozygotes because of dominance:

F_2 *genotype*	*Value*
qq	$\mu - a$,
QQ	$\mu + a$,
Qq	$\mu + d$.

Estimation and transformation of parameters is straightforward. For example, the relationship between parameters on the observed and underlying scale is:

$$P(qq) = 1 - \Phi(T - \mu_c + a_c), \quad (4)$$

$$P(Qq) = 1 - \Phi(T - \mu_c - d_c), \quad (5)$$

$$P(QQ) = 1 - \Phi(T - \mu_c - a_c), \quad (6)$$

and

$$a_{01} = (P(QQ) - P(qq))/2, \quad (7)$$

$$d_{01} = p(Qq) - ((P(qq) + P(QQ))/2. \quad (8)$$

Under an additive model, the expectation of d_{01} is not zero, because of the non-linearity of the frequencies on the observed scale. Hence, we expect a dominance effect on the observed scale, even if there is no dominance effect on the underlying scale. Relationships between population parameters on the observed and underlying scale are presented in Table 2.

(iii) *Simulation*

A single chromosome of 100 cM with 6 or 11 evenly spaced fully informative markers was simulated. Parameters were estimated either using the linear regression method of Haley & Knott (1992), or using a GLM (Numerical Algorithms Group, 1990) with either a probit or a logit link. For any location on the chromosome, the values of the explanatory variables

https://doi.org/10.1017/S0016672300033887 Published online by Cambridge University Press

were exactly the same for both methods. Effects fitted were an overall mean and the expectation, in terms of either an additive effect or both an additive and dominance effect, for the mean genotypic effect of a putative QTL given its flanking markers (Haley & Knott, 1992). The probit model is the more appropriate one given the model of simulation, but both probit and logit models are widely used in GLM analyses of binary data and they tend to give similar results (McCullagh & Nelder, 1989). Data were simulated on an underlying Normal scale, and then transformed to a 0/1 (binary) scale using the appropriate threshold from Tables 1 and 2.

For the LRG method, the test statistic used was an approximate likelihood ratio, i.e.

$$\text{Test statistic} = N \log\left(\frac{\text{Residual SS reduced model}}{\text{Residual SS full model}}\right),$$

where N is the number of observations (Haley & Knott, 1992). The full model contains both a mean effect and QTL effects (an additive effect in a BC population, and both an additive and a dominance effect in F_2 populations), and in the reduced model only an overall mean is fitted. The GLM produced an average deviance between fitted and observed values. For a single location on the chromosome, the difference in the deviance for the full and reduced model is asymptotically distributed as a χ^2 with degrees of freedom equal to the difference in the number of parameters fitted (i.e. 1 D.F. for BC populations, and 2 D.F. for F_2 populations when both additive and dominance effects are fitted).

By chance, estimates of the mean or additive QTL effect on the observed scale can be negative (for low values of P and small population sizes) when using linear regression. This occurs when the incidence pertaining to one of the marker genotype classes is either 0 or 1, i.e. no variation within a genotype class. In that case, a transformed threshold on the underlying scale was chosen so that the probability of obtaining the observed incidence for a population size equal to that particular genotype class was 0·5. For example, a negative estimate for μ_p from a BC population of $N =$ 200 results from $P(Qq) = 0$. The frequency, say $P^*(Qq)$, which would give $P(Qq) = 0$ in 50% of samples of 100 individuals from that genotype class is calculated from $[1 - P^*(Qq)]^{100} = 0{\cdot}5$, which gives $P^*(Qq) = 0{\cdot}0069$. Finally, the corresponding parameter on the underlying scale is calculated as $(T-\mu_c)^* = \Phi^{-1}(1 - P^*(Qq))$. For this example, $(T-\mu_c)^* =$ 2·46. For F_2 populations ($N = 500$), the corresponding value for $(T-\mu_c+a_c)^*$ was 2·54.

To investigate the power of the different methods, 5% significance thresholds were simulated from 10000 replicate populations.

Parameters which were varied for the additive model in BC populations were population size (200 or 500, which are population sizes corresponding to real experiments), incidence (0·10, 0·25 and 0·50), marker spacing (10 or 20 cM), and the effect of the QTL (h^2 of 0·01, 0·05 and 0·10).

Table 3. *Simulated (10000 replicates) 5% significance thresholds in a backcross population for different incidences* (P), population sizes (N), and marker spacing (Δ, *in cM*), *for linear regression (LRG), GLM with probit link (GLM(p)), and GLM with logit link (GLM(g)). The same data were used for the different methods of analysis*

P	N	Δ	LRG	GLM(p)	GLM(g)
0·50	200	10	7·5	7·4	7·4
		20	7·0	6·9	6·9
	500	10	7·4	7·4	7·4
		20	6·8	6·8	6·8
0·25	200	10	7·3	7·3	7·3
		20	6·9	6·9	6·9
	500	10	7·4	7·4	7·4
		20	6·9	6·9	6·9
0·10	200	10	7·2	7·5	7·5
		20	6·9	7·1	7·1
	500	10	7·3	7·4	7·4
		20	6·8	6·8	6·8

F_2 populations were simulated to investigate the estimation of dominance effects for different models. For all simulations, a population size of 500 and a marker spacing of 20 cM were used, and incidences were varied as before. A QTL was simulated which was either additive ($a_c = 0{\cdot}4714$, $d_c = 0$), completely dominant ($a_c = 0{\cdot}4851$, $d_c = 0{\cdot}4851$) or completely recessive ($a_c = 0{\cdot}4851$, $d_c = -0{\cdot}4851$). These three genetic models correspond to a narrow sense heritability of 10% in the F_2 population.

3. Results

(ii) *Backcross population*

The 5% significance thresholds for BC populations are presented in Table 3. For a particular population size and incidence, all significance thresholds were determined using the same data, i.e. the same data were analysed with all three models. For each combination of incidence and population size, the 5% thresholds are very similar for all models. There is no difference in thresholds to the accuracy shown between the two GLMs. In general, threshold values do not change much with different values of the proportion affected or population size. This suggests that the likelihood ratio approximation used by Haley & Knott (1992) is robust to departures from normality.

A comparison between GLM using either the probit or logit link was made for a marker spacing of 20 cM, based on 1000 replicate populations. To calculate the power of the models, the 5% significance threshold for GLM(p) and GLM(g) from Table 3 were used. The results (which are not shown elsewhere) showed

https://doi.org/10.1017/S0016672300033887 Published online by Cambridge University Press

Table 4. *Comparison between analyses for backcross populations using linear regression (LRG) and GLM(p) on the same data for P = 0·50, 0·25 and 0·10, and a marker spacing of 20 cM. Results for 1000 replicates*

			LRG			GLM(p)		
P	N	h^2	a_c	a_{01}	Power (%)	a_c	a_{01}	Power (%)
0·50	200	0·01	0·24	0·094	12	0·24	0·094	12
		0·05	0·51	0·199	52	0·51	0·199	52
		0·10	0·71	0·274	86	0·71	0·274	86
	500	0·01	0·22	0·089	27	0·23	0·089	27
		0·05	0·48	0·189	93	0·48	0·189	93
		0·10	0·68	0·267	100	0·68	0·267	100
0·25	200	0·01	0·22	0·068	12	0·22	0·068	12
		0·05	0·52	0·159	46	0·52	0·160	46
		0·10	0·72	0·218	80	0·72	0·219	80
	500	0·01	0·23	0·071	22	0·22	0·071	22
		0·05	0·48	0·151	87	0·48	0·151	87
		0·10	0·68	0·209	99	0·68	0·209	99
0·10	200	0·01	0·25	0·041	7	0·25	0·041	8
		0·05	0·56	0·086	28	0·57	0·086	28
		0·10	0·78	0·116	53	0·79	0·117	53
	500	0·01	0·24	0·041	16	0·24	0·041	17
		0·05	0·51	0·084	66	0·50	0·084	66
		0·10	0·71	0·113	93	0·70	0·113	93

Symbols are explained in previous tables.

that in terms of the average test statistic, power and the estimate of the additive effect on the observed scale, the models gave the same averages up to at least two significant digits. Therefore, we use only the probit link in further analyses.

Table 4 shows the comparison between LRG and GLM(p) for different values of P, N and h^2. The similarity between the estimates on both scales and the power for both models of analysis is striking. Estimates of the QTL effect on both the observed (a_{01}) and underlying scale (a_c) were usually biased upwards for both models (compared with the population values from Table 1). However, for QTL of large effects (h^2 = 5% or 10%) and N = 500, the biases were not very large. Additional to the simulations with a marker spacing of 20 cM (Table 4), simulations were performed with a marker spacing of 10 cM. However, there was little difference in power for different marker spacings, and results for a marker spacing of 10 cM are not shown in Table 4. Typically, the power for the reduced marker spacing was 1–2% higher. Power decreased with decreasing heritability and incidence and smaller population size, and for low powers the average estimated position of the QTL tended towards the middle of the chromosome (results not shown). For example, the power and average location for a QTL explaining 10% of the variation for N = 200 was 89% and 27 cM for an incidence of 0·50 (Table 4), and 56% and 31 cM for P = 0·10 (Table 4).

Table 5. *Comparison between analyses using linear regression (LRG) and a generalized linear model (GLM) on the same data for* F_2 *populations. Results for 1000 replicates. Pos., average location of QTL in cM. Dominance effects were fitted but not simulated. N = 500, Δ = 20 cM, h² = 0·10*

P	Method	a_c	d_c	a_{01}	d_{01}	Pos.	Power (%)
0·50	LRG	0·48	0·01	0·183	0·002	24	99
	GLM	0·48	0·01	0·182	0·002	24	99
0·25	LRG	0·49	0·00	0·145	−0·026	24	98
	GLM	0·48	0·01	0·144	−0·023	24	98
0·10	LRG	0·50	0·00	0·079	−0·029	26	87
	GLM	0·52	0·05	0·079	−0·023	26	87

Symbols are explained in previous tables.

(ii) F_2 *population*

Simulated thresholds for an F_2 of 500 individuals for the LRG and GLM(p) models were 10·2 and 10·2 (P = 0·50), 9·9 and 9·9 (P = 0·25) and 10·0 and 10·2 (P = 0·10), respectively. As expected, these values are larger than the threshold values from the BC populations, because an additional parameter was estimated. Again, threshold values were similar when using either LRG or a GLM.

Results for a simulated additive model are presented in Table 5. Both models of analysis produce similar results. For incidences of 0·25 and 0·10, the average estimated dominance is significantly different from zero for both models. Hence, although no dominance

https://doi.org/10.1017/S0016672300033887 Published online by Cambridge University Press

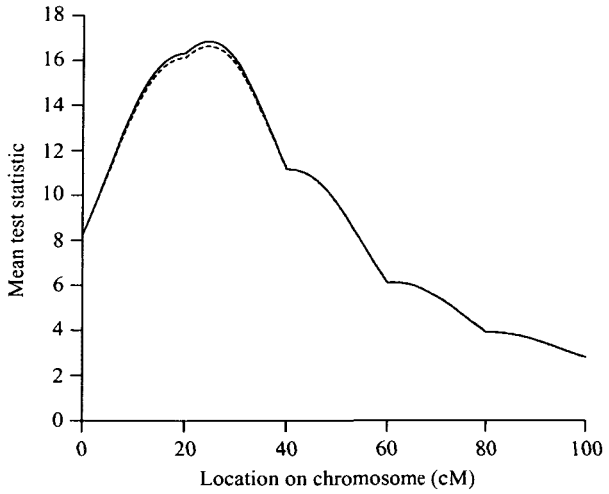

Fig. 1. Average test statistic (over 1000 replicates) per chromosome location for an F_2 population ($N = 500$) under an additive QTL model ($a_c = 0{\cdot}4714$, $d_c = 0$) and incidence of 10%. Continuous line, LRG; dashed line, GLM.

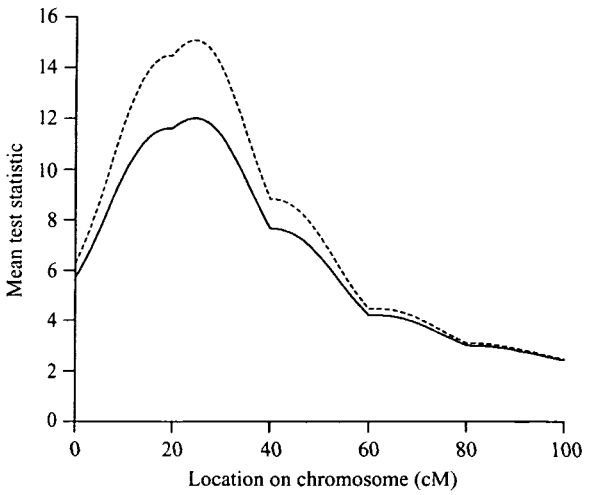

Fig. 2. Average test statistic (over 1000 replicates) per chromosome location for an F_2 population ($N = 500$) under a dominant QTL model ($a_c = 0{\cdot}4851$, $d_c = 0{\cdot}4851$) and incidence of 10%. Continuous line, LRG; dashed line, GLM.

was simulated, a small effect (about $-0{\cdot}02$ on the observed scale, equivalent to a dominance ratio (d/a) of around $-0{\cdot}3$ for 10% incidence data) was estimated. In Fig. 1, the average test statistic along the chromosome is plotted for both models for an incidence of 10%. This confirms that the methods are very similar in power.

Results for a completely dominant and recessive QTL are shown in Table 6. Estimates and power for the two methods are similar, except for an incidence of $P = 0{\cdot}10$ for the dominant genetic model (the positive QTL allele is completely dominant). In that case, the GLM method appears better, in that the power is significantly larger than the power obtained from linear regression (87% *v.* 78%, respectively). The average test statistic along the chromosome for both methods is shown in Fig. 2. For the recessive QTL, although there is no difference in power at the 5% level (both methods having 100% power), an investigation of the test statistics revealed that the average test statistic for LRG was larger than that from GLM. This is shown in Fig. 3.

4. Discussion

We have shown that mapping QTL for binary traits on the observed 0/1 scale using a linear model gives very similar results to more sophisticated GLMs. Jansen (1992) noted for an example with exponential residuals that although the maximum likelihood was much larger under GLM, parameter estimates were similar.

Table 6. *Comparison between analyses using linear regression (LRG) and a generalized linear model (GLM) on the same data for* F_2 *populations. Results for 1000 replicates. Pos., average location of QTL in cM.* $N = 500$, $\Delta = 20$ *cM*, $h^2 = 0{\cdot}10$. *The QTL was either completely dominant (dom) or completely recessive (rec)*

Gene action	P	Method	a_c	d_c	a_{01}	d_{01}	Pos.	Power (%)
dom	0·50	LRG	0·49	0·49	0·182	0·179	24	100
		GLM	0·49	0·48	0·180	0·179	24	100
	0·25	LRG	0·52	0·51	0·121	0·120	24	99
		GLM	0·48	0·50	0·118	0·122	24	100
	0·10	LRG	0·53	0·53	0·057	0·055	26	78
		GLM	0·55	0·58	0·054	0·059	25	87
rec	0·50	LRG	0·50	−0·49	0·183	−0·182	24	100
		GLM	0·49	−0·49	0·181	−0·182	24	100
	0·25	LRG	0·49	−0·50	0·167	−0·171	24	100
		GLM	0·50	−0·49	0·170	−0·170	24	100
	0·10	LRG	0·51	−0·50	0·101	−0·103	24	100
		GLM	0·53	−0·45	0·105	−0·101	24	100

Symbols are explained in previous tables.

https://doi.org/10.1017/S0016672300033887 Published online by Cambridge University Press

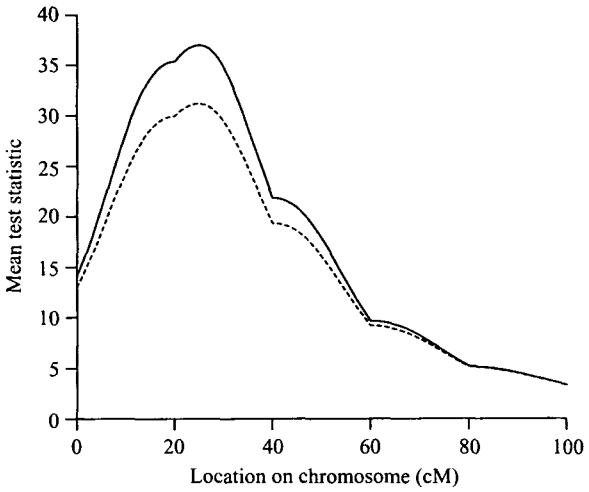

Fig. 3. Average test statistic (over 1000 replicates) per chromosome location for an F_2 population ($N = 500$) under a recessive QTL model ($a_c = 0{\cdot}4851$, $d_c = -0{\cdot}4851$) and incidence of 10%. Continuous line, LRG; dashed line, GLM.

Theoretically, using all information on the data, including their distribution, should be better, so that we would expect the GLM model to do better. However, in terms of estimates of QTL effects and power, no differences were found. Even for more extreme incidences, differences between the models were very small. For example, for a BC population with $P = 0{\cdot}01$ and $N = 1000$, powers for QTL explaining 5% or 10% of the variance were 29% and 46% for LRG, and 28% and 45% for GLM (results not shown in tables). For smaller population sizes and/or more extreme incidences the models may produce different results, but the power for such populations would be very small. Several studies in the animal breeding literature which compare linear models with non-linear models for estimation genetic parameters such as heritabilities and genetic correlations and for prediction of breeding values have come to similar conclusions, i.e. that linear models are robust to departures from normality (e.g. Meijerink & Gianola, 1985; Perez-Enciso *et al.* 1993).

The few cases for which there was a difference between LRG and GLM in terms of power are interesting, and need explanation. For example, a difference of 9% in power was observed for the dominant QTL (Table 6), with GLM being the more powerful model. From simulation we found that the distributions of the test statistic for LRG and GLM under the null hypothesis were very similar (results not shown). Therefore, the cause of the greater power must solely be the higher average test statistic for GLM when the dominant QTL is present (Fig. 2). We calculated the value of the test statistics for LRG and GLM assuming that the QTL was at a single marker (in that case regression is equivalent to maximum likelihood for normally distributed data) and using the population values for parameters in the likelihood equations, and compared these expected test statistic values with simulation results (also for a single marker). We found that the GLM test statistic was larger than the one from LRG, and that predicted and simulated results were similar. For example, LRG and GLM were compared in an F_2 population of 500 individuals, with a mean incidence of 0·25, and test statistics were averaged over 1000 replicate populations. The expected difference in the likelihood ratio test statistic between the linear and generalized linear model, i.e. (test LRG − test GLM), were 0·4, −4·8 and 7·8 for an additive, dominant and recessive model, respectively, while the observed values from simulation were 0·5, −5·3 and 8·2, respectively. Details of these calculations are shown in the Appendix. For a recessive QTL it was found by simulation that the average test statistic for LRG is larger than that for GLM (Fig. 3) (therefore, where power is intermediate, that for LRG is likely to be higher than that for GLM). Again, likelihood calculations for a single marker (see example above and the Appendix) reveal that for the set of parameters we used for the recessive QTL the test statistic for LRG is larger than the test statistic for GLM, as observed by simulation.

When the power is low, both methods apparently give estimates of the QTL location which are biased towards the centre of the chromosome. However, the average location of a QTL which is presented in the tables is an average over all replicates, irrespective of the strength of evidence for a QTL in any particular sample. In some of the replicate populations the location with the largest test statistic will represent a type I error, and, on average, the corresponding estimated QTL location pertaining to these samples will be in the centre of the chromosome. Hence the average QTL location over all replicate samples will be biased towards the centre of the chromosome. If the size of the test statistic for each replicate population is taken into account, for example by summing the test statistic for each chromosome location over replicates, the location with the largest average test statistic corresponds to the location of simulated QTL (results not shown), so that the location estimate is unbiased when using the criterion of average test statistic per chromosome location.

In practice, LRG has the advantages that it is easier and quicker to use, for example, facilitating more detailed analyses such as bootstrapping and permutation tests. Our study suggests that LRG may be used in practice for binary traits, and perhaps for traits with other non-Normal distributions as well. There are no apparent drawbacks in using LRG, with the direction of any differences in power from GLM being dictated by the underlying genetic model, which is unlikely to be known in advance.

P.M.V. was funded by the Marker Assisted Selection Consortium of the UK pig industry (Cotswold Pig Development Company Ltd, J.S.R. Farms Ltd, National Pig Development Company, Newsham Hybrid Pigs Ltd, Pig

Improvement Company, and the Meat and Livestock Commission) and by MAFF, DTI and the BBSRC. C.S.H. acknowledges support from MAFF, BBSRC and the European Commission. S.A.K. was supported by the BBSRC. We thank Bill Hill, Robin Thompson, John Webb and Sijne van de Beek for many useful comments on an earlier version of the manuscript.

Appendix

(i) *Expected test statistics for LRG and GLM in* F_2 *populations*

We consider a simple case where the QTL genotypes are known, i.e. we consider a single marker, and the QTL is at the marker. Notation is as follows:

$p(1) = E(P(qq))$ = expected incidence of genotype qq,

$p(2) = E(P(Qq))$ = expected incidence of genotype Qq,

$p(3) = E(P(QQ))$ = expected incidence of genotype QQ,

$n(1) = E(n(qq)) = N/4$,

$n(2) = E(n(Qq)) = N/2$,

$n(3) = E(n(QQ)) = N/4$,

P = overall incidence in F_2 population.

ML estimates for linear and generalized linear model

For both models, assume that, under the null hypothesis (i.e. no QTL, fit just an overall mean), the expected parameter estimates on the observed scale are

$$\mu = [\Sigma n_i p_i]/N, \tag{A 1}$$

$$\sigma^2 = [\Sigma n_i p_i(1-p_i)]/N. \tag{A 2}$$

Hence, when the null hypothesis H_0 (no QTL) is true:

$\mu = P$,

$\sigma^2 = P(1-P)$.

(ii) *Linear model*

The log-likelihood can be written as

$$L \propto -\tfrac{1}{2}[N\log\sigma^2 + \Sigma_i\Sigma_j(y_{ij}-\mu_i)^2/\sigma^2] \quad (i = 1,3; j = 1,n_i), \tag{A 3}$$

$$E(\mathrm{ML}) \approx -\tfrac{1}{2}[N + N\log(\sigma^2)].$$

The (expected) maximum likelihood (ML) values are obtained by substituting the expected parameter estimates under the full and reduced model into eqn (A 3), i.e.

$\sigma^2_{\text{full}} = [\Sigma n_i p_i(1-p_i)]/N$,

$\sigma^2_{\text{red}} = \mu(1-\mu)$

$= ([\Sigma n_i p_i]/N)(1-[\Sigma n_i p_i]/N)$

The test statistic, $t(\mathrm{LRG}) = 2(\mathrm{ML(full)} - \mathrm{ML(reduced)})$, then becomes

$$t(\mathrm{LRG}) = N[(\log(\mu(1-\mu)) - \log(\sigma^2_{\text{full}})]. \tag{A 4}$$

(iii) *Generalized linear model*

The log-likelihood and maximum log-likelihood equation are (from McCullagh & Nelder, 1989), again using expected parameter estimates,

$L(\mathrm{GLM}) = \Sigma_i\Sigma_j[y_{ij}\log(p_i) + (1-y_{ij})\log(1-p_i)]$,

$E(\mathrm{ML(GLM)} \approx \Sigma[n_i p_i \log(p_i) + n_i(1-p_i)\log(1-p_i)]$.

For the full model, p_i is $P(qq)$, $P(Qq)$ and $P(QQ)$, respectively. For the reduced model, p_1 (only a single probability fitted) is $\mu = [\Sigma n_i p_i]/N$.

The likelihood ratio test for GLM then becomes

$$t(\mathrm{GLM}) = 2\{\Sigma[n_i p_i \log(p_i) + n_i(1-p_i)\log(1-p_i)] - N\mu\log(\mu) - N(1-\mu)\log(1-\mu)\}. \tag{A 5}$$

References

Ducrocq, V. & Quaas, R. L. (1988). Prediction of genetic response to truncation selection across generation. *Journal of Dairy Science* **71**, 2543–2553.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. Harlow, England: Longman.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 299–309.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Jansen, R. C. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**, 252–260.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Jansen, R. C. (1994). Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics* **138**, 871–881.

Kruglyak, L. & Lander, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

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

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.

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

Meijerink, A. & Gianola, D. (1985). Linear versus nonlinear methods of sire evaluation for categorical traits: a simulation study. *Genetics, Selection, Evolution* **17**, 115–132.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974). *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Numerical Algorithms Group (1990). *The NAG Fortran Library Manual*, mark 14. Oxford: NAG Ltd.

Perez-Enciso, M., Tempelman, R. J. & Gianola, D. (1993). A comparison between linear and Poisson mixed models for litter size in Iberian pigs. *Livestock Production Science* **35**, 303–316.

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.