Genet. Res., Camb. (2004), **83**, *pp.* 159–168. *With 4 figures.* © 2004 Cambridge University Press
DOI: 10.1017/S0016672304006834 Printed in the United Kingdom

Mapping QTLs for traits measured as percentages

YONGCAI MAO AND SHIZHONG XU*
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521-0124, USA

(*Received 23 June 2003 and in revised form 19 September 2003 and 8 February 2004*)

Summary

Many quantitative traits are measured as percentages. As a result, the assumption of a normal distribution for the residual errors of such percentage data is often violated. However, most quantitative trait locus (QTL) mapping procedures assume normality of the residuals. Therefore, proper data transformation is often recommended before statistical analysis is conducted. We propose the probit transformation to convert percentage data into variables with a normal distribution. The advantage of the probit transformation is that it can handle measurement errors with heterogeneous variance and correlation structure in a statistically sound manner. We compared the results of this data transformation with other transformations and found that this method can substantially increase the statistical power of QTL detection. We develop the QTL mapping procedure based on the maximum likelihood methodology implemented via the expectation-maximization algorithm. The efficacy of the new method is demonstrated using Monte Carlo simulation.

1. Introduction

Many variables are discrete in nature and such discrete variables often exhibit correlation among different observations. Examples can be found in various fields like genetics, epidemiology, familial studies, pedigree analysis, teratology, toxicology, ophthalmology and sample surveys. In many experiments encountered in the biological and biomedical sciences, data are generated in the form of a ratio, n_j/N_j, or percentage, where n_j is a non-negative count of success and is bounded by the positive integer N_j, which is the number of trials (Finney, 1971; Fisler & Warden, 1997; Moody *et al.*, 1999). When N_j is assumed to be fixed and known, n_j may be modeled as a binomial variable with parameter p_j; that is, we may view n_j as the sum of N_j independent Bernoulli random variables, W_{jk} ($k=1 \ldots N_j$), with $E(W_{jk})=p_j$. If there is some correlation among the W_{jk} values then n_j would no longer follow a binomial distribution. This situation is not uncommon (Garren *et al.*, 2001) and, in certain applications, the basic assumption of a binomial model in which individuals are responding independently of each other might not be defensible. The lack of independence among the individual respondents will result in a larger variability than can be explained by the binomial model. Count data coming from such studies have a larger variance than the variance of a binomial variable and are said to exhibit overdispersion (Moore, 1986; Sudhir & Islam, 1995). For example, in the analyses of littermate data from biological or toxicological experiments, it is often of interest to study the intraclass correlation as a means of investigating the heritability of a certain trait. Although there is an extensive literature, summarized in reviews by Donner (1986) and Muller & Buttner (1994), for the statistical analysis of intraclass correlation for continuous response variables, techniques are less developed for proportional data, which are also of practical importance in many medical and biostatistical applications (Ahmed *et al.*, 2000). As Donner (1986) remarked, the application of continuous theory to proportional variables has severe limitations because the associated methods of inferences are not strictly valid.

Typical percentage traits in biological experiments include the percentage of deformed seeds in plants,

* Corresponding author. Department of Botany and Plant Sciences, University of California, Riverside, CA 92521-0124, USA. Tel: +1 (909) 787 5898. e-mail: xu@genetics.ucr.edu

the mortality of litters in pigs and so on. Binary traits are special forms of the percentage traits. Traditional methods for mapping quantitative trait loci (QTLs) responsible for the variation of percentage traits have not taken into account how the percentage values are measured. The same percentage value (say 50%) measured from different sample sizes (say 2/4 and 20/40) should have different residual error variance. Yet this heterogeneous residual variance is rarely used in an attempt at QTL mapping for percentage traits. Rather than being treated as continuous characters, traits measured as percentage are better analysed as binomial variables based on the theory of discrete data analysis.

The simplest and most naive approach to analysing binomial proportional data is to ignore the indications that the data might be binomial in nature and to perform the standard analysis of variance. The two most obvious problems with this approach are: (1) that the predicted values are not necessarily between 0 and 1; and (2) that the equal-variance assumption is not necessarily satisfied. Regarding the second problem, the assumption that the variances are equal implies that the mean is not related to the variance, which is contrary to the binomial model, in which the variance is a function of the mean. Of major significance in hypothesis testing is the fact that the standard errors of the estimated proportions from the standard analysis of variance do not reflect the nature of the binomial variance of the response variable. Yet this method is still used at times because of the wide availability of least-square software, relying on asymptotic theory to justify the use of the normal distribution (Brooks *et al*., 1997). In practice, such an analysis might produce reasonable results when the treatment groups have similar binomial variances and little extra-binomial variability occurs. Collett (1991) gave a detailed discussion of this issue.

Various transformations on the proportions have been used in an attempt to minimize the effects of these two major problems. An approach found in many standard statistical text books is to use the logistic transformation of the proportional variable $T_j = \ln(n_j/N_j) - \ln(1 - n_j/N_j)$ as the response variable. The probit model has been the dominant model in biometrics. The logit and probit models give similar predictions except for extreme values of the dose. There is no compelling biological reason, however, to adopt either the logit or the probit specification (Neter *et al*., 1996).

The arcsine transformation is a useful transformation for proportions and percentages. The proportion can be made nearly normal if the square root of each proportion is used with the arcsine transformation, $A_j = \arcsin[(n_j/N_j)^{0.5}]$. However, the transformation is not very good at the extreme ends of the data (near 0% and near 100%). A discussion of the arcsine transformation can be found in Snedecor & Cochran (1989) or Zar (1996).

For the correlated percentage data, we have to deal with the intraclass correlation. There are many different estimators of intraclass correlation that have been proposed for binary data. Ridout *et al*. (1999) gave an excellent review; see also Mak (1988).

Here, we present a method for analysing correlated binomial proportion data using the correlated probit model. One advantage of our approach is that the standard errors of the probit model can be computed in a straightforward way. Some earlier papers have addressed the main issue of this paper but the results have not led to suggestions that are easy to use in practical applications. Examples of these papers include Ochi & Prentice (1984), Poirier & Ruud (1988), Throne *et al*. (1995) and Gueorguieva & Agresti (2001). Here, we use the maximum likelihood (ML) method implemented via the expectation-maximization (EM) algorithm to estimate genetic effects.

We also provide a simulation study to investigate the performance of our method and compare it with three other methods that do not take the correlation into account. The study indicates that our method performs better than other methods, particularly in small samples. Initially, we construct a model that incorporates the key features of the applications that would benefit from our approach. Finally, we use simulations to obtain numerical estimates of the parameters.

2. Model and methods

(i) *Genetic model*

We consider a single, large, full-sib family with m sibs. QTL mapping in full-sibs is important in forest trees and laboratory animals (Knott *et al*., 1996; Xu, 1996, 1998). Let $\{(j, 1), (j, 2), \ldots, (j, N_j)\}$ be the labels of N_j trials of sib j. For ease of presentation, the responses are arbitrarily named as 'normal' and 'deformed', and the $\{0, 1\}$ metric is imposed with 0 for normal and 1 for deformed. We observe the number of deformed n_j out of the N_j trials for sib j (N_j is the size of the trials). Let W_{jk} be a random variable taking a value 1 if trial (j, k) is deformed and 0 if trial (j, k) is normal. The observed values w_{jk} are defined such that

$$n_j = \sum_{k=1}^{N_j} w_{jk},$$

and the observations from different individuals are assumed to be independent.

Often, an underlying continuous scale called liability has been assumed; trials are scored 1 if they exceed a certain threshold value t. Let Z_{jk} represent the underlying random variable associated with trial (j, k)

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

such that

$$W_{jk}=1 \Leftrightarrow Z_{jk}>t.$$

In the scale of liability, we assume that Z_{jk} can be described by the following mixed linear model

$$Z_{jk}=\mu+q_j\alpha^p+u_j\alpha^m+r_j\delta+e_{jk}, \quad (1)$$

where μ is the population mean, α^p is the average effect of allelic substitution of the paternal parent (i.e. the difference between the genetic values of the two alleles carried by the father), α^m is the average effect of allelic substitution of the maternal parent, δ is the dominance effect (interaction between the two allelic substitution effects) and e_{jk} is the environmental error assumed to be normally distributed with mean 0. It is then postulated that the correlation between any pair Z_{jk}, Z_{jl} has the same value ρ for any j and $k\neq l$. The coefficients of the genetic effects, q_j, u_j and r_j, are defined as follows. Let the genotypes of the father and mother be $A_1^pA_2^p$ and $A_1^mA_2^m$, respectively. There are four possible genotypes in the progeny ($A_1^pA_1^m$, $A_1^pA_2^m$, $A_2^pA_1^m$ and $A_2^pA_2^m$). Notice that these genotypes are ordered with paternal allele followed by the maternal allele. Variables q_j, u_j and r_j depend on the genotype of j, and are defined as follows: $(q_j, u_j, r_j)=(1, 1, 1)$ if individual j takes the first genotype, $A_1^pA_1^m$; $(q_j, u_j, r_j)=(1, -1, -1)$ if j takes the second genotype, $A_1^pA_2^m$; $(q_j, u_j, r_j)=(-1, 1, -1)$ if j takes the third genotype, $A_2^pA_1^m$; and $(q_j, u_j, r_j)=(-1, -1, 1)$ if j takes the last genotype, $A_2^pA_2^m$. These variables are collectively called the design matrix in a general linear model. Under Mendelian segregation, the four genotypes will take an equal frequency in the full-sib family. Therefore, the three genetic effects defined this way are orthogonal. If there is no segregation distortion, we expect that

$$\mathrm{E}\begin{bmatrix}q\\u\\r\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix} \quad \text{and} \quad \mathrm{Var}\begin{bmatrix}q\\u\\r\end{bmatrix}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}.$$

Let $\mathrm{Var}(Z_{jk})=\sigma_e^2$ and ρ be the correlation between Z_{jk} and Z_{jl}, then

$$\begin{aligned}\mathrm{var}(\bar{Z}_j)&=\mathrm{var}\left(\frac{1}{N_j}\sum_{k=1}^{N_j}Z_{jk}\right)\\&=\frac{1}{N_j^2}\left[\sum_{j=1}^{N}\mathrm{var}(Z_{jk})+2\sum_{k<l}\mathrm{cov}(Z_{jk},Z_{jl})\right]\\&=\frac{1+(N_j-1)\rho}{N_j}\sigma_e^2. \quad (2)\end{aligned}$$

The genetic model described above assumes that the genotype of j at the locus of interest is observed (i.e. variables q_j, u_j and r_j are known). In practice, however, we only observe marker genotypes and the three variables q_j, u_j and r_j actually represent the genotypes of markers. If a marker is linked to a QTL, the marker genotypes can be used to formulate the genetic model, from which we can estimate genetic parameters and perform statistical tests.

(ii) *Correlated probit model*

The threshold t cannot be estimated and thus it must be treated as a constant. There is no loss of generality in setting $t=0$, and we adopt this convention regardless of the distribution of e_{jk}. If the actual t is not zero, the population mean μ will be shifted, but μ is only a nuisance parameter in the model whose value does not change the estimates and tests of the QTL effects. Model 1 for the complete data $\{Z_{jk}\}$ can be translated into the following model for the observed binomial proportional data: Conditional on q_j, u_j and r_j,

$$\Phi^{-1}(p_j)=\mu+q_j\alpha^p+u_j\alpha^m+r_j\delta, \quad (3)$$

where p_j is the mean of the observed variable and Φ denotes the standardized cumulative normal distribution function.

If we use the sample proportions n_j/N_j as the estimates of the mean p_j then the probit model will be invalid when $n_j=0$ or $n_j=N_j$. We use the following Bayesian estimate of p_j to overcome this problem (Press, 1989). Let N_j be the number of independent trials of an experiment in which there are two possible outcomes on each trial, 'deformed' or 'normal'. Let n_j denote the number of deformed during the N_j trials, and let p_j be the probability of deformed on a single trial. The probability mass function for n_j is given by

$$f(n_j|p_j)=\binom{N_j}{n_j}p_j^{n_j}(1-p_j)^{N_j-n_j},$$

$$0<p_j<1, \quad n_j=0,1,\ldots,N_j.$$

Assume that the prior distribution for p_j is uniform (uninformative prior),

$$g(p_j)=\begin{cases}1 & 0<p_j<1\\0 & \text{otherwise,}\end{cases}$$

then using Bayes' theorem, the posterior density is

$$\begin{aligned}h(p_j|n_j)&=\frac{p_j^{n_j}(1-p_j)^{N_j-n_j}g(p_j)}{\int_0^1 p_j^{n_j}(1-p_j)^{N_j-n_j}g(p_j)dp_j}\\&=\frac{p_j^{n_j}(1-p_j)^{N_j-n_j}}{B(n_j+1,N_j-n_j+1)}.\end{aligned}$$

That is, the posterior distribution of p_j given n_j is a β distribution with parameters n_j+1 and N_j-n_j+1; therefore, the mean of p_j given n_j is

$$E(p_j|n_j)=\frac{n_j+1}{N_j+2}.$$

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

In this paper, we use

$$\hat{p}_j = \frac{n_j + 1}{N_j + 2}, \tag{4}$$

as the estimates of p_j.

We now formulate the correlated probit model for the binomial proportion data. Let

$$\mathbf{H} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \equiv \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \mathbf{h}_3 \\ \mathbf{h}_4 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mu \\ \alpha^p \\ \alpha^m \\ \delta \end{pmatrix}, \tag{5}$$

then our correlated probit model is (Gueorguieva & Agresti, 2001)

$$y_j = \mathbf{x}_j \mathbf{b} + \sigma_j e_j, \tag{6}$$

where $y_j = \Phi^{-1}[(n_j+1)/(N_j+2)]$, $\mathbf{x}_j = \mathbf{h}_1$ if individual j takes the first genotype $(A_1^p A_1^m)$, $\mathbf{x}_j = \mathbf{h}_2$ if j takes the second genotype $(A_1^p A_2^m)$ and so on, e_j is the residual error distributed as $N(0, \sigma_e^2)$, and $\sigma_j = \{[1 + (N_j - 1)\rho]/N_j\}^{0.5}$.

(iii) *Parameter estimation using the EM algorithm*

Let $A_1 = A_1^p A_1^m$, $A_2 = A_1^p A_2^m$, $A_3 = A_2^p A_1^m$ and $A_4 = A_2^p A_2^m$. When the genotype is A_i, the distribution of y_j is

$$\begin{aligned} f(y_j|A_i) = f_j(i) &= \frac{1}{\sqrt{2\pi\sigma_j^2\sigma_e^2}} \exp\left[-\frac{1}{2\sigma_j^2\sigma_e^2}(y_j - \mathbf{h}_i\mathbf{b})^2 \right] \\ &= \sqrt{\frac{N_j}{2\pi[1 + (N_j - 1)\rho]\sigma_e^2}} \\ &\quad \times \exp\left[-\frac{N_j}{2[1 + (N_j - 1)\rho]\sigma_e^2}(y_j - \mathbf{h}_i\mathbf{b})^2 \right]. \end{aligned} \tag{7}$$

Using the multipoint method (Rao & Xu, 1998), we can infer the probabilities of QTL genotypes conditional on the marker information, $p_j(i) = \Pr(\mathbf{x}_j = \mathbf{h}_i | I_M)$ for $i = 1, \ldots, 4$, where I_M represents marker information. Therefore, the mixture of the four distributions is

$$f(y_j) = \sum_{i=1}^{4} p_j(i) f_j(i),$$

and the log likelihood function is

$$L(\mathbf{b}, \sigma_e^2, \rho | \mathbf{y}, \mathbf{M}) = \sum_{j=1}^{m} \ln\left[\sum_{i=1}^{4} p_j(i) f_j(i) \right]. \tag{8}$$

The EM steps are described as follows.

(1) Take the probit transformation, obtaining the data $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$ with $y_j = \Phi^{-1}(\hat{p}_j) = \Phi^{-1}[(n_j+1)/(N_j+2)]$, where Φ represents the standardized cumulative normal distribution function.

(2) Choose initial values of the parameters $\theta^{(0)} = (\mathbf{b}^{(0)}\ \sigma_e^{2(0)}\ \rho^{(0)})$.

(3) Calculate the posterior probabilities of QTL genotype given the initial values of the parameters and y_j

$$p_j^*(i) = \frac{p_j(i) f_j(i)}{\sum_{k=1}^{4} p_j(k) f_j(k)}, \quad i = 1, \ldots, 4,$$

where $f_j(i)$ $(i = 1, \ldots, 4)$ are evaluated at the initial values of parameters.

(4) Expectation step: calculate the following expectations using the posterior distribution of y_j:

$$E(\mathbf{x}_j^T \mathbf{x}_j) = \sum_{i=1}^{4} p_j^*(i) \mathbf{h}_i^T \mathbf{h}_i,$$

$$E(\mathbf{x}_j^T y_j) = \sum_{i=1}^{4} p_j^*(i) \mathbf{h}_i^T y_j,$$

$$E(y_j - \mathbf{x}_j \mathbf{b}^{(0)})^2 = \sum_{i=1}^{4} p_j^*(i)(y_j - \mathbf{h}_i \mathbf{b}^{(0)})^2.$$

(5) Maximization step: having obtained the above expectations, we use the generalized weighted least-squares method to calculate the maximum-likelihood estimates of the parameters $\mathbf{b}$ and σ_e^2:

$$\begin{aligned} \mathbf{b}^{(1)} &= [E(\mathbf{X}^T \mathbf{W} \mathbf{X})]^{-1} E(\mathbf{X}^T \mathbf{W} \mathbf{y}) \\ &= \left\{ \sum_{j=1}^{m} \frac{N_j}{[1 + (N_j - 1)\rho^{(0)}]\sigma_e^{2(0)}} E(\mathbf{x}_j^T \mathbf{x}_j) \right\}^{-1} \\ &\quad \times \left\{ \sum_{j=1}^{m} \frac{N_j}{[1 + (N_j - 1)\rho^{(0)}]\sigma_e^{2(0)}} E(\mathbf{x}_j^T y_j) \right\}, \end{aligned}$$

where

$$\mathbf{X} = (\mathbf{x}_1^T\ \mathbf{x}_2^T, \ldots, \mathbf{x}_m^T),$$

$$\mathbf{W} = \operatorname{diag}\left(\frac{N_1}{[1 + \rho^{(0)}(N_1 - 1)]\sigma_e^{2(0)}}, \frac{N_2}{[1 + \rho^{(0)}(N_2 - 1)]\sigma_e^{2(0)}}, \ldots, \frac{N_m}{[1 + \rho^{(0)}(N_m - 1)]\sigma_e^{2(0)}} \right).$$

$$\sigma_e^{2(1)} = \frac{1}{m} \sum_{j=1}^{m} E(y_j - \mathbf{x}_j \mathbf{b}^{(0)})^2.$$

The estimate of ρ is obtained by the solution of $E(\partial L/\partial \rho) = 0$ with

$$L = \sum_{j=1}^{m} \left\{ \frac{1}{2} \ln \frac{N_j}{2\pi\sigma_e^2} - \frac{1}{2} \ln[1 + (N_j - 1)\rho] - \frac{N_j[y_j - \mathbf{x}_j \mathbf{b}]^2}{2[1 + (N_j - 1)\rho]\sigma_e^2} \right\},$$

that is,

$$\sum_{j=1}^{m} \left\{ \frac{N_j - 1}{1 + (N_j - 1)\rho} - \frac{N_j(N_j - 1)E[(y_j - \mathbf{x}_j \mathbf{b}^{(0)})^2]}{\sigma_e^{2(0)}[1 + (N_j - 1)\rho]^2} \right\} = 0.$$

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

Table 1. *Mean estimates and standard deviations (in parentheses) of the position (cM) and effects of the QTLs calculated from 100 replicated simulations. The standard deviations among the 100 repeated simulations represent standard errors of the estimated parameters. Empirical estimates of the statistical power at a Type I rate of 0·05 are given in the last column*

Heritability (h^2)	Sample size (N_j)	Method	QTL location (cM)	Mean μ	Paternal allelic effect (α^p)	Maternal allelic effect (α^m)	Dominance effect (δ)	Statistical power (%) ($\alpha = 0{\cdot}05$)
3·03 %		True	45	0	0·09	0·09	0·13	
	π (5)	Model A	45·80 (24·06)	0·00 (0·05)	0·07 (0·06)	0·08 (0·05)	0·09 (0·07)	42
		Model B	46·90 (24·42)	0·00 (0·06)	0·07 (0·08)	0·07 (0·07)	0·09 (0·10)	28
		Model C	44·18 (24·24)	0·50 (0·02)	0·03 (0·03)	0·03 (0·03)	0·03 (0·04)	28
		Model D	46·17 (24·78)	45·08 (1·28)	1·61 (1·70)	1·58 (1·57)	1·87 (2·29)	28
	π (10)	Model A	46·39 (18·30)	0·00 (0·05)	0·10 (0·06)	0·08 (0·05)	0·12 (0·06)	50
		Model B	48·62 (20·68)	0·00 (0·06)	0·09 (0·07)	0·07 (0·07)	0·12 (0·08)	43
		Model C	50·91 (22·56)	0·50 (0·02)	0·03 (0·03)	0·03 (0·03)	0·04 (0·03)	44
		Model D	48·67 (20·61)	45·05 (1·25)	1·97 (1·60)	1·45 (1·70)	2·61 (1·78)	43
	π (50)	Model A	45·67 (15·50)	0·00 (0·05)	0·10 (0·06)	0·10 (0·05)	0·14 (0·06)	75
		Model B	46·44 (16·17)	0·00 (0·06)	0·10 (0·07)	0·10 (0·06)	0·14 (0·07)	67
		Model C	45·73 (16·70)	0·50 (0·02)	0·04 (0·02)	0·04 (0·02)	0·05 (0·02)	70
		Model D	46·15 (16·31)	45·03 (1·21)	2·24 (1·43)	2·21 (1·27)	3·06 (1·51)	69
5·88 %		True	45	0	0·13	0·13	0·18	
	π (5)	Model A	45·27 (19·19)	−0·01 (0·05)	0·10 (0·06)	0·09 (0·06)	0·12 (0·08)	66
		Model B	45·73 (19·73)	−0·01 (0·06)	0·10 (0·07)	0·08 (0·07)	0·12 (0·08)	51
		Model C	45·37 (19·46)	0·50 (0·02)	0·04 (0·02)	0·03 (0·03)	0·04 (0·03)	53
		Model D	45·86 (19·83)	44·71 (1·26)	2·11 (1·47)	1·91 (1·62)	2·58 (1·86)	51
	π (10)	Model A	44·28 (14·15)	0·00 (0·04)	0·11 (0·05)	0·12 (0·05)	0·17 (0·05)	75
		Model B	43·27 (14·30)	0·00 (0·05)	0·11 (0·07)	0·12 (0·06)	0·16 (0·06)	67
		Model C	43·41 (14·26)	0·50 (0·02)	0·04 (0·02)	0·04 (0·02)	0·06 (0·02)	67
		Model D	43·31 (14·28)	45·05 (1·23)	2·29 (1·43)	2·51 (1·43)	3·58 (1·42)	68
	π (50)	Model A	45·82 (7·20)	0·00 (0·04)	0·13 (0·05)	0·13 (0·05)	0·19 (0·05)	97
		Model B	46·28 (8·92)	0·00 (0·05)	0·14 (0·06)	0·14 (0·06)	0·19 (0·06)	94
		Model C	46·32 (8·13)	0·50 (0·02)	0·05 (0·02)	0·05 (0·02)	0·07 (0·02)	95
		Model D	46·33 (8·17)	45·02 (1·10)	3·07 (1·28)	3·02 (1·34)	4·11 (1·25)	94
11·11 %		True	45	0	0·18	0·18	0·25	
	π (5)	Model A	44·10 (7·83)	0·00 (0·05)	0·13 (0·05)	0·13 (0·05)	0·19 (0·06)	95
		Model B	43·84 (8·97)	0·00 (0·05)	0·13 (0·06)	0·13 (0·06)	0·19 (0·07)	89
		Model C	43·77 (8·29)	0·50 (0·02)	0·05 (0·02)	0·05 (0·02)	0·07 (0·03)	90
		Model D	43·75 (8·00)	45·09 (1·11)	2·91 (1·27)	2·80 (1·28)	4·07 (1·61)	89
	π (10)	Model A	44·89 (5·52)	0·00 (0·04)	0·15 (0·05)	0·16 (0·05)	0·23 (0·04)	100
		Model B	44·59 (6·56)	0·00 (0·05)	0·15 (0·06)	0·16 (0·06)	0·23 (0·06)	99
		Model C	44·55 (6·30)	0·50 (0·02)	0·05 (0·02)	0·06 (0·02)	0·08 (0·02)	98
		Model D	44·57 (6·56)	44·94 (1·21)	3·19 (1·38)	3·47 (1·23)	4·95 (1·25)	99
	π (50)	Model A	45·19 (4·35)	0·00 (0·04)	0·18 (0·04)	0·18 (0·05)	0·26 (0·05)	100
		Model B	45·28 (5·35)	0·00 (0·06)	0·19 (0·06)	0·19 (0·06)	0·27 (0·06)	100
		Model C	45·28 (5·15)	0·50 (0·02)	0·06 (0·02)	0·06 (0·02)	0·09 (0·02)	100
		Model D	45·24 (5·14)	44·97 (1·20)	4·02 (1·23)	4·11 (1·29)	5·77 (1·34)	100

The estimate $\rho^{(1)}$ of unknown parameter ρ can be solved numerically using the bisection procedure according to the above equation.

(6) Replace the initial parameters $\theta^{(0)}$ by $\theta^{(1)}$ and go back to step 2 for the next iteration. Continue the iterations until a criterion of convergence is reached. At the convergence, the values of the parameters are the maximum likelihood solutions.

(iv) *Likelihood ratio test*

To test the significance of the QTL effect, a likelihood ratio statistic is used. We first evaluate the log likelihood function with the parameters substituted by their ML estimate under the full model, denoted by

$$L_1 = L(\hat{\mathbf{b}}, \hat{\sigma}_e^2, \hat{\rho} | \mathbf{y}, \mathbf{M}).$$

We then evaluate the log likelihood function under the null model so that $\mathbf{b} = (\mu, 0, 0, 0)^T$ is used in place of $\mathbf{b}$, denoted by

$$L_0 = L((\hat{\mu}, 0, 0, 0)^{\mathrm{T}}, \hat{\sigma}_e^2, \hat{\rho} | \mathbf{y}, \mathbf{M}).$$

Notice that here $\hat{\mu}, \hat{\sigma}_e^2, \hat{\rho}$ are obtained by maximizing the log likelihood function under the

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

Table 2. *Mean estimates and standard deviations (in parentheses) of the intraclass correlation in Model A calculated from 100 replicated simulations. The standard deviations among the 100 repeated simulations represent standard errors of the estimated parameters. The true value of the intraclass correlation is 0·2*

Heritability (h^2)	Sample size (N_j)	Intraclass correlation (ρ)
3·03 %	π (5)	0·1606 (0·0429)
	π (10)	0·1658 (0·03)
	π (50)	0·1747 (0·0183)
5·88 %	π (5)	0·1635 (0·0337)
	π (10)	0·1702 (0·0209)
	π (50)	0·1793 (0·0133)
11·11 %	π (5)	0·1957 (0·0357)
	π (10)	0·1983 (0·0214)
	π (50)	0·1994 (0·0166)

reduced model

$$L((\mu,0,0,0)^{\mathrm{T}},\sigma_e^2,\rho|\mathbf{y},\mathbf{M}) = \sum_{j=1}^{m}\ln\left\{\sqrt{\frac{N_j}{2\pi[1+(N_j-1)\rho]\sigma_e^2}} \times \exp\left[-\frac{N_j}{2[1+(N_j-1)\rho]\sigma_e^2}(y_j-\mu)^2\right]\right\}$$

and are different from those in $\hat{\mathbf{b}}, \hat{\sigma}_e^2, \hat{\rho}$. The likelihood ratio test statistic is defined as

$$\lambda = -2(L_0 - L_1). \quad (9)$$

3. Simulation studies

This section reports the results of experiments that test the accuracy of the approaches in applications with small sample sizes. We focus on the performance of our method and three alternative methods. We call our model Model A (Eqn 6); the other three models include the probit model, $y_j = \mathbf{x}_j\mathbf{b} + e_j$ where $y_j = \Phi^{-1}((n_j+1)/(N_j+2))$ with homogeneous residual variance (Model B); $y_j = \mathbf{x}_j\mathbf{b} + e_j$ where $y_j = (n_j+1)/(N_j+2)$ (Model C); and the arcsine transformation model $y_j = \mathbf{x}_j\mathbf{b} + e_j$ with $y_j = \arcsin[(n_j+1)/(N_j+2)]^{0.5}$ (Model D). Model C simply treats the percentage data as a regular quantitative trait without any transformation.

We simulated a single chromosome of 11 markers with 10 cM between consecutive markers. A single QTL was simulated at position 45 cM (between markers 5 and 6). The parental marker genotype of each locus comprised two alleles randomly sampled from five unique alleles.

The residual error was assumed to be normally distributed, with a variance set at $\sigma_e^2 = 1{\cdot}0$, and the intraclass correlation was set at $\rho = 0{\cdot}2$. The total phenotypic variance explained by the QTL was simulated at three levels:

(1) $\sigma_a^2 = 0{\cdot}03125$ and $\sigma_d^2 = 0{\cdot}0625$, where the QTL explains $h^2 = (2\times 0{\cdot}03125 + 2\times 0{\cdot}0625)/(2\times 0{\cdot}03125 + 0{\cdot}0625 + 1) = 11{\cdot}11\,\%$ of the trait variance. The corresponding allelic effects that generate this set of QTL variances are $\alpha^p = 0{\cdot}1768$, $\alpha^m = 0{\cdot}1768$ and $\delta = 0{\cdot}25$.
(2) $\sigma_a^2 = 0{\cdot}015625$ and $\sigma_d^2 = 0{\cdot}03125$, where the QTL explains $h^2 = 5{\cdot}88\,\%$ of the trait variance. The corresponding allelic effects that generate this set of QTL variances are $\alpha^p = 0{\cdot}125$, $\alpha^m = 0{\cdot}125$ and $\delta = 0{\cdot}1768$.
(3) $\sigma_a^2 = 0{\cdot}0078125$ and $\sigma_d^2 = 0{\cdot}015625$, where the QTL explains $h^2 = 3{\cdot}03\,\%$ of the trait variance. The corresponding allelic effects that generate this set of QTL variances are $\alpha^p = 0{\cdot}0884$, $\alpha^m = 0{\cdot}0884$ and $\delta = 0{\cdot}125$.

The population mean of the liability was 0. We simulated an outbred full-sib family with 100 individuals. The number of trials per individual, N_j, was simulated according to a Poisson distribution with parameter 5, 10 or 50. The likelihood ratio (LR) test statistic profile was calculated across the chromosome with 1 cM increments.

Table 1 shows the means and standard deviations of estimates of location as well as effects of the QTL and the empirical power calculated from 100 repeated simulations. For large numbers of trials and high QTL heritability h^2, the three methods tend to produce an unbiased estimate of the QTL position and small estimation errors. For low h^2, especially with a small number of trials, the estimated position of QTL is biased towards the center of the chromosome. Our correlated probit model (model A) and the probit model (model B) give estimates of the paternal and maternal allelic substitution effects and the dominance deviation that are reasonably close to the true values. Our method has higher power than the other two models when the number of trials is small and h^2 is low.

The means and standard deviations of the estimated intraclass correlation ρ over 100 replicates are given in Table 2. We can see that the estimate of ρ gets more accurate as h^2 and N_j increase.

Figures 1–4 show the results of the LR test statistic profiles for the four models against the map position for the cases of heritability equal to 11·11 % and average N_j equal to 10, heritability 11·11 % and average N_j 50, heritability 5·88 % and average N_j 10, and heritability 5·88 % and average N_j 50, respectively. In each figure, parts a–d show the results for Models

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

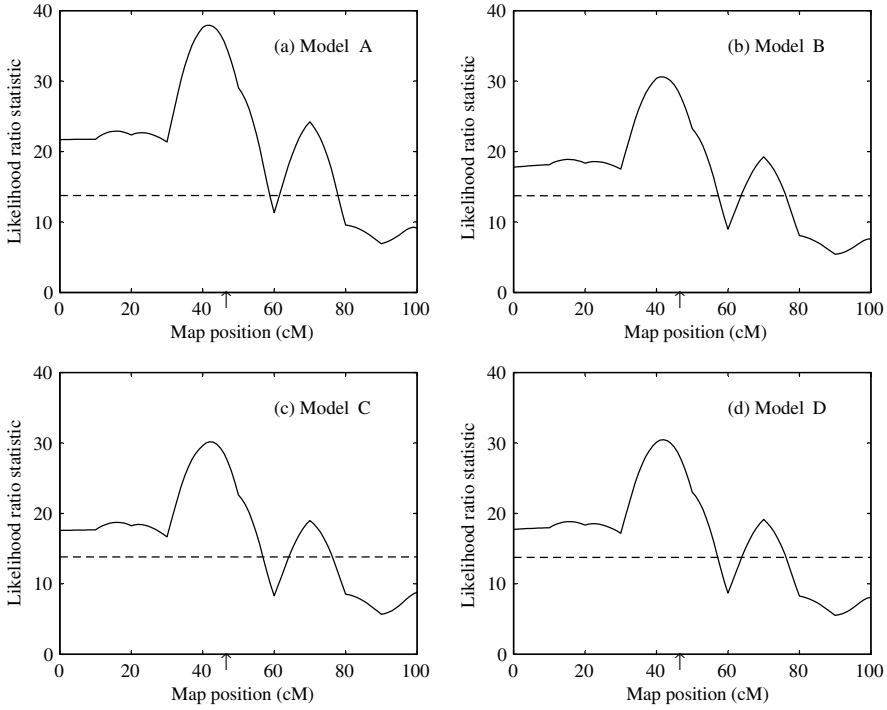


Fig. 1. Comparison of the likelihood ratio test statistic profiles of the four models for the case where the variation explained by the QTL is 11·11 % and the distribution parameter for the trial number of each individual is 10. For this test, 11 codominant markers are equally spaced along a chromosome of 100 cM and a single QTL resides at position 45 cM. (a) Model A. (b) Model B. (c) Model C. (d) Model D. The labels of the horizontal axis indicate the marker positions measured in centiMorgans (cM) counted from the left-hand end of the chromosome. The simulated true QTL location is indicated with an arrow.

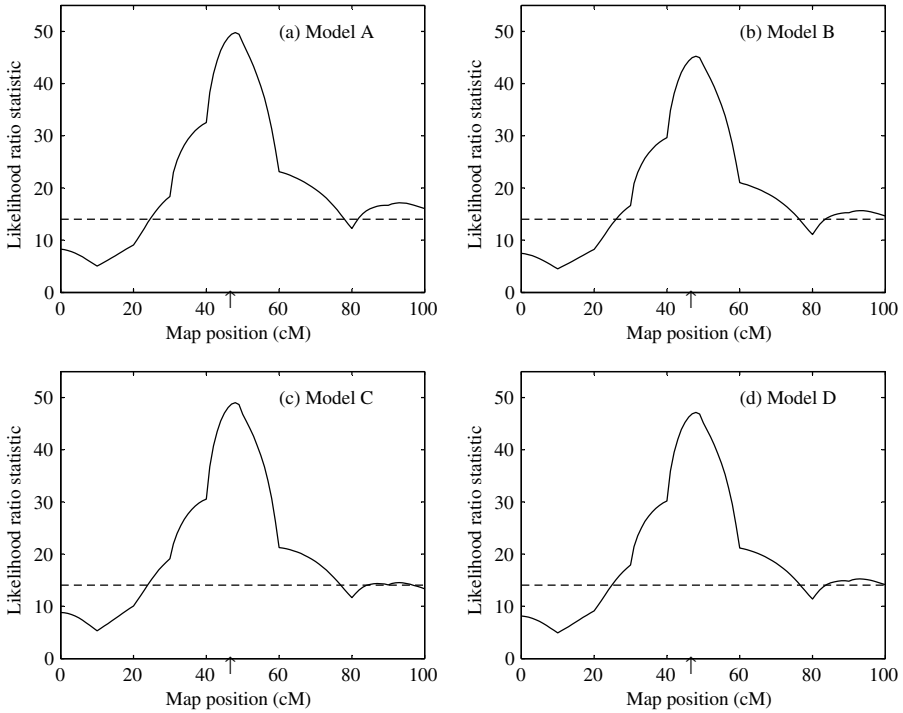


Fig. 2. Comparison of the likelihood ratio test statistic profiles of the four models for the case where the variation explained by the QTL is 11·11 % and the distribution parameter for the trial number of each individual is 50. For this test, 11 codominant markers are equally spaced along a chromosome of 100 cM and a single QTL resides at position 45 cM. (a) Model A. (b) Model B. (c) Model C. (d) Model D. The labels of the horizontal axis indicate the marker positions measured in centiMorgans (cM) counted from the left-hand end of the chromosome. The simulated true QTL location is indicated with an arrow.

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

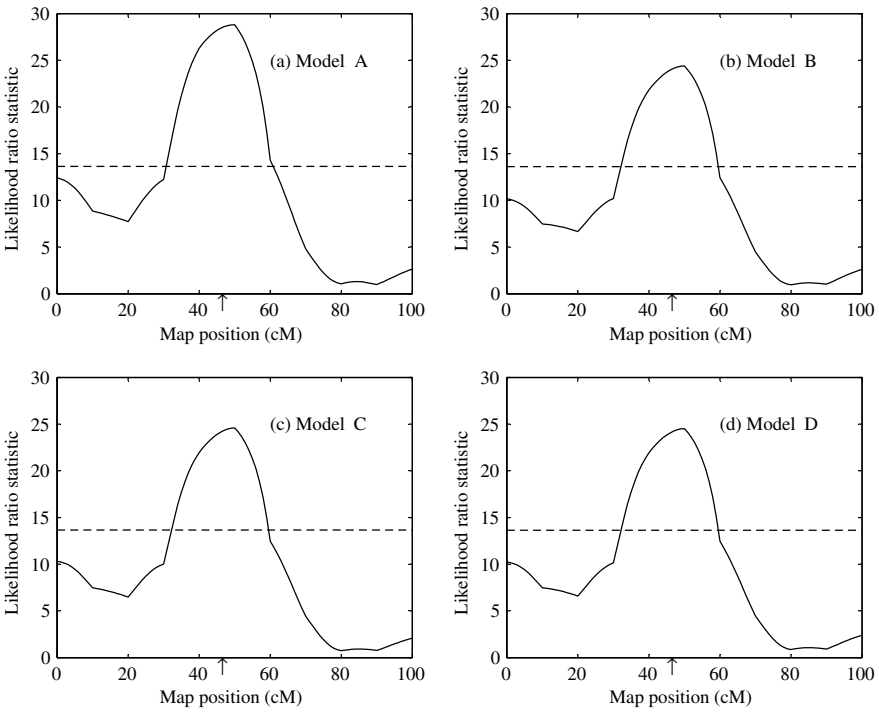


Fig. 3. Comparison of the likelihood ratio test statistic profiles of the four models for the case where the variation explained by the QTL is 5·88 % and the distribution parameter for the trial number of each individual is 10. For this test, 11 codominant markers are equally spaced along a chromosome of 100 cM and a single QTL resides at position 45 cM. (a) Model A. (b) Model B. (c) Model C. (d) Model D. The labels of the horizontal axis indicate the marker positions measured in centiMorgans (cM) counted from the left-hand end of the chromosome. The simulated true QTL location is indicated with an arrow.

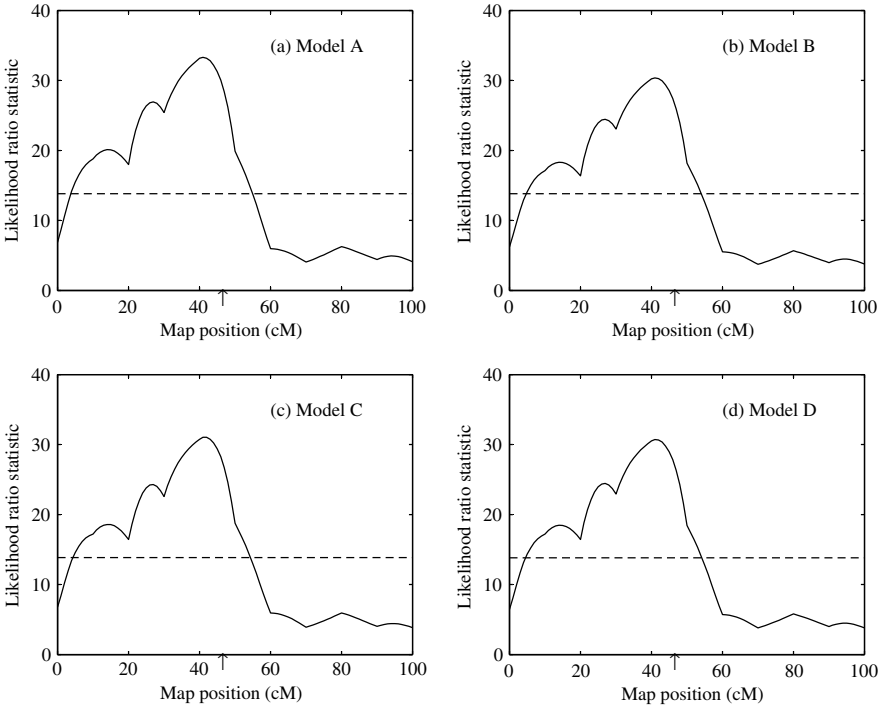


Fig. 4. Comparison of the likelihood ratio test statistic profiles of the four models for the case where the variation explained by the QTL is 5·88 % and the distribution parameter for the trial number of each individual is 50. For this test, 11 codominant markers are equally spaced along a chromosome of 100 cM and a single QTL resides at position 45 cM. (a) Model A. (b) Model B. (c) Model C. (d) Model D. The labels of the horizontal axis indicate the marker positions measured in centiMorgans (cM) counted from the left-hand end of the chromosome. The simulated true QTL location is indicated with an arrow.

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

A–D, respectively. As stated earlier, the intraclass correlation of the simulated data was 0·2.

These figures show clearly that our method performs better than the other methods for small N_j values. As N_j increases, all methods perform equally well. Also, all curves peak near the true location (45 cM) of the QTL. In each graph, the dashed line is the approximate threshold for QTL detection computed using the method of Piepho (2001).

4. Discussion

In this study, we developed a correlated probit model for mapping binomial proportional data. The correlation can reduce the effect of error variance and therefore make it easier to detect QTLs (i.e. make the test more powerful). Not only is the power of QTL detection increased but also the precision of the estimated QTL position is improved. Binary traits are special cases of binomial characters in which the number of trials always equals one ($N_j=1$). As a result, the algorithm developed here can be applied to binary trait mapping except that the intraclass correlation is irrelevant here. The estimated probability of success for individual j becomes $\hat{p}_j=(n_j+1)/(N_j+2)=2/3$ and the probability of failure is $1-\hat{\rho}_j=1/3$. However, we do not recommend using our algorithm for binary data analysis because algorithms specialized for binary trait mapping have been developed. One of the earliest works on binary trait mapping can be found in Visscher *et al.* (1996), who treated binary characters (defined as 0 or 1) as normally distributed variables so that a least-squares method can be applied. McIntyre (2001) developed a probabilistic approach to mapping QTLs for binary traits. Hackett & Weller (1995), Xu & Atchley (1996), Rao & Xu (1998) and Yi & Xu (2000) described binary traits using a threshold model so that the QTL effects are estimated in the scale of liability. More recently, Xu *et al.* (2003) proposed an EM algorithm to map QTLs for binary traits in a four-way cross experiment. The EM algorithm has unified QTL mapping for discrete traits with that for continuous traits. All the aforementioned methods were designed for mapping binary or ordinal traits rather than for mapping traits measured as percentages.

The main advantage of this method is the simplicity of converting the percentage data into (approximately) normally distributed data and thus we can use the EM algorithm straightforwardly. The weighted regression analogy for estimation of QTL parameters makes the method easy to implement in writing computer programs. Our simulation results suggest that the probit model can be used for the binomial proportional data. However, the usual probit model is not always suitable, particularly if the number of trials is small and there is a correlation structure in the data. Our correlated probit model has solved the problem.

Chib & Greenberg (1998) developed a method of simulated maximum likelihood for the multinomial probit model in which estimates are obtained using a Monte Carlo version of the EM algorithm. However, it is well known that, for the multinomial probit model, the full-information simulation estimation methods, at their current state of development, are subject to numerous computational difficulties in finding an optimal solution for all but the simplest models. Our method, however, takes advantage of its special form and does not have this problem.

Although we demonstrate the statistical method of QTL mapping using full-sib families as an example, families from other types of mating designs can in principle be readily incorporated by simplifying the full-sib family model. The model considered here assumes only one QTL on the chromosome. In reality, complex binomial proportional traits might be controlled by multiple loci. If there are multiple QTLs in the same chromosome, the estimator tends to be biased because of interference caused by QTLs located on the same chromosome but outside the tested region (Zeng, 1994). This problem can be solved by resorting to the concept of composite interval mapping (Jansen, 1994; Zeng, 1994).

The method described here is not intended to replace the standard QTL mapping procedure for percentage data. If the number of trials per individual is sufficiently large, the usual probit model and other methods would provide correct estimates of the location and effects of QTLs. However, when the number of trials is small, especially when the heritability of the QTL is low, the method presented in this paper will allow correct analysis of the binomial proportional data.

The presented method has been implemented in a Matlab program, which is available on request from the authors.

We thank two anonymous reviewers for their helpful comments on the original submission. The work was supported by the National Institutes of Health Grant R01-GM55321 and the USDA National Research Initiative Grants Program 00-35300-9245 to S.X.

References

Ahmed, S. E., Gupta, A. K., Khan, S. M. & Nicol, C. (2000). Simultaneous estimation of several intraclass correlation coefficients. *Annals of the Institute of Statistical Mathematics* **53**, 354–369.

Brooks, S. P., Morgan, B. J. T., Ridout, M. S. & Pack, S. E. (1997). Finite mixture models for proportions. *Biometrics* **53**, 1097–1115.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

Collett, D. (1991). *Modelling Binary Data*. Chapman & Hall.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* **54**, 67–82.

Finney, D. J. (1971). *Probit Analysis*, 3rd edn. Cambridge: Cambridge University Press.

Fisler, J. S. & Warden, C. H. (1997). Mapping of mouse obesity genes: a generic approach to a complex trait. *The Journal of Nutrition* **127**, 1909S–1916S.

Garren, S. T., Smith, R. L. & Piegorsch, W. W. (2001). Bootstrap goodness-of-fit test for the beta-binomial model. *Journal of Applied Statistics* **28**, 561–571.

Gueorguieva, R. V. & Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association* **96**, 1102–1112.

Hackett, C. A. & Weller, J. I. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1252–1263.

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

Knott, S. A., Neale, D. B., Sewell, M. M. & Haley, C. S. (1997). Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. *Theoretical and Applied Genetics* **94**, 810–820.

McIntyre, L. M., Coffman, C. J. & Doerge, R. W. (2001). Detecting and localization of a single binary trait locus in experimental populations. *Genetical Research* **78**, 79–92.

Mak, T. K. (1988). Analysing intraclass correlation for dichotomous variables. *Applied Statistics* **37**, 344–352.

Moody, D. E., Pomp, D., Nielsen, M. K. & Van Vleck L. D. (1999). Identification of quantitative trait loci influencing traits related to energy balance in selection and inbred lines of mice. *Genetics* **152**, 699–711.

Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika* **73**, 583–588.

Muller, R. & Buttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine* **13**, 2465–2476.

Neter, J., Kutner, M. H., Nachtsheim, C. J. & Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th edn. Homewood, IL: R.D. Irwin.

Ochi, Y. & Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71**, 531–543.

Piepho, H. P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* **157**, 425–432.

Poirier, D. J. & Ruud, P. A. (1988). Probit with dependent observations. *Review of Economic Studies* **55**, 593–614.

Press, J. (1989). *Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons.

Rao, S. & Xu, S. (1998). Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity* **81**, 214–224.

Ridout, M. S., Demetrio, G. B. & Firth D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55**, 137–148.

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods*, 8th edn. Ames, IA: Iowa State University Press.

Sudhir, R. P. & Islam, A. S. (1995). Analysis of proportions in the presence of over-/under-dispersion. *Biometrics* **51**, 1400–1410.

Throne, J. E., Weaver, D. K., Chew, V. & Baker, J. E. (1995). Probit analysis of correlated data: multiple observations over time at one concentration. *Journal of Economic Entomology* **88**, 1510–1512.

Visscher, P. M., Haley, C. S. & Knott, S. A. (1996). Mapping QTLs for binary traits in backcross and F2 populations. *Genetical Research* **68**, 55–63.

Xu, S. (1996). Mapping quantitative trait loci using four-way crosses. *Genetical Research* **68**, 175–181.

Xu, S. (1998). Iteratively reweighted least squares mapping of quantitative trait loci. *Behavior Genetics* **28**, 341–355.

Xu, S. & Atchley, W. R. (1996). Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* **143**, 1417–1424.

Xu, S., Yi, N., Burke, D., Galecki, A. & Miller, R. A. (2003). An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. *Genetical Research* **82**, 127–138.

Yi, N. & Xu, S. (2000). Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* **155**, 1391–1403.

Zar, J. H. (1996). *Biostatistical Analysis*, 3rd edn. Prentice Hall.

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

Genet. Res., Camb. (2004), **83**, *pp.* 169–176. *With 3 figures.* © 2004 Cambridge University Press
DOI: 10.1017/S0016672304006858 Printed in the United Kingdom

Sex-specific selection on the human X chromosome?

PATRICIA BALARESQUE[1,2]*, BRUNO TOUPANCE[1], QUINTANA-MURCI LLUIS[3], BRIGITTE CROUAU-ROY[2] AND EVELYNE HEYER[1]*
[1] *Unité Eco-Anthropologie MNHN/CNRS/P7 UMR5145, Paris, France*
[2] *Laboratoire Evolution et Diversité Biologique, Toulouse, France*
[3] *CNRS URA1961, Institut Pasteur, Paris, France*

(*Received 12 August 2003 and in revised form 12 March 2004*)

Summary

Genes involved in major biological functions, such as reproductive or cognitive functions, are choice targets for natural selection. However, the extent to which these genes are affected by selective pressures remains undefined. The apparent clustering of these genes on sex chromosomes makes this genomic region an attractive model system to study the effects of evolutionary forces. In the present study, we analysed the genetic diversity of a X-linked microsatellite in 1410 X-chromosomes from 10 different human populations. Allelic frequency distributions revealed an unexpected discrepancy between the sexes. By evaluating the different scenarios that could have led to this pattern, we show that sex-specific selection on the tightly linked VCX gene could be the most likely cause of such a distortion.

1. Introduction

Natural selection is one of the main forces shaping the patterns of genetic variability in the human genome, although its role has been often neglected in most population genetics studies. Indeed, most genetic polymorphisms used in population genetic studies are assumed to be neutral and affected mainly by both mutation and genetic drift. Interestingly, the recent results from human genome sequencing have revealed that each category of repeated sequences possesses a specific dynamics in space and time. This suggests a combined and complex action of different evolutionary forces. Furthermore, these results revealed that microsatellites, which are among the most used markers in population genetics studies, displayed a non-random distribution through the human genome: there are fewer polymorphic loci on the X-chromosome compared with autosomes (International Sequencing Human Genome Consortium (ISHGC), 2001). The X-chromosome, known to harbour a number of genes involved in human fertility (Wang *et al.*, 2001; Saifi & Chandra, 1999) and in cognitive functions (Gécz & Mulley, 2000; Hurst & Randerson, 1999; Graves & Delbridge, 2001) may be potential target for natural selection. However, the precise extent to which these genes are affected by selective pressures, or a possible variation in selective pressures acting between both sexes, remains undefined. In addition, the precise extent to which these genes have an influence on surrounding sequences either through direct or background selection remains poorly studied. We investigated the allelic diversity of an X-linked dinucleotide microsatellite (DXS8175) located in the Xp22.3 region, a gene-rich region, in 10 human populations and focused on the allelic distributions. We found strikingly different allelic frequency distributions between males and females. In this study, we investigated the likely demographic and selective scenarios as the bases of these observations.

* Corresponding authors: Equipe de Génétiques des populations, Musée de l'Homme, 17, place du Trocadéro, 75116 Paris, France. Tel: +33 1 44057253. Fax: +33 1 44057241. e-mail: heyer@mnhn.fr or balares@mnhn.fr

2. Materials and methods

(i) *Samples and PCR amplification*

We genotyped 951 unrelated subjects, for a total of 1410 chromosomes belonging to 10 different human populations from Africa (Akan and Yacouba from

Table 1. *Distribution of alleles frequencies in males (M) and females (F), $\Delta = p_f - p_m$, and statistical test of differentiation between the two genders*

	DXS 8175[c] alleles frequencies											Exact test differentiation (Raymond & Rousset, 1995)	
Populations	10	11	12	13	14	15	16	17	18	19	20	χ^2	*P* value
European													
Corsican													
M[a] 63[b]				6	10	46	37	2				1·75	0·417
F 60				3	13	32	48	3					
$\Delta = p_f - p_m$				**−3**	**4**	**−14**	**12**	**2**					
Sardinian													
M 36			11		11	47	22	8				10·65	0·005
F 94			1		18	51	30						
$\Delta = p_f - p_m$			**−10**		**7**	**4**	**8**	**−8**					
Orcadian													
M 32						34	59	6				3·14	0·208
F 76					11	38	46	5					
$\Delta = p_f - p_m$					**11**	**4**	**−13**	**−1**					
African													
Akan													
M 59	5		8		36	17	17	14	2	2		15·05	0·000
F 166	5		3		18	16	45	11		1	1		
$\Delta = p_f - p_m$	**−0**		**−6**		**−18**	**−1**	**28**	**−2**	**−2**	**−1**	**1**		
Yacouba													
M 62			3		19	27	37	6	3	3		1·21	0·546
F 46	2		2		28	13	39	11	2	2			
$\Delta = p_f - p_m$	**2**		**−1**		**9**	**−14**	**2**	**4**	**−1**	**−1**			
Amhara													
M 31				7	26	19	32	13	3			0·35	0·838
F 56		5	5	5	20	14	36	13	2				
$\Delta = p_f - p_m$		**5**	**5**	**−1**	**−6**	**−5**	**3**	**−0**	**−1**				
Oromo													
M 31			7	3	13	13	55	10				1·97	0·373
F 80	3		6		23	21	38	6		4			
$\Delta = p_f - p_m$	**3**		**−0**		**10**	**8**	**−17**	**−4**		**4**			
Moroccan Berber													
M 38					11	47	32	5		5		0·63	0·728
F 136	2		1	2	14	43	33	5		1			
$\Delta = p_f - p_m$	**2**		**1**	**2**	**4**	**−4**	**2**	**−0**		**−4**			
Mozabit Berber													
M 85			1	1	5	34	48	9	1			2·25	0·324
F 42	2		2		14	24	50	7					
$\Delta = p_f - p_m$	**2**		**1**	**−1**	**10**	**−10**	**2**	**−2**	**−1**				
South American													
Bolivian													
M 55					9	60	29	2				0·24	0·889
F 162				1	7	57	33	1					
$\Delta = p_f - p_m$				**1**	**−2**	**−3**	**4**	**−1**					

[a] Genders.
[b] Number of chromosomes analysed.
[c] Alleles are named according to Scozzari *et al*. (1997).

Ivory Coast, Amhara and Oromo from Ethiopia, Algerian Mozabits and Moroccan Berbers from North Africa), Europe (Sardinian, Corsican and Orcadian) and South-America (Bolivians) (see Table 1). DXS8175 microsatellite amplifications were performed according to Malaspina *et al*. (1997) and Scozzari *et al*. (1997). PCR primers were fluorescently labelled and the PCR products were run in a standard 6% denaturing gel and detected using an ABI 373A automated sequencer. GeneScan software (ABI) and Genotyper software package (ABI) were used to size the amplified alleles. In addition, we sequenced a total

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

of 50 microsatellites randomly chosen from the 10 populations to control for possible indel events in flanking sequences (Grimaldi & Crouau-Roy, 1997) and showed that no indels have caused length homoplasy. Moreover, no null alleles have been detected as previously reported by Scozzari *et al.* (1997).

(ii) *Statistical analysis*

The distributions of allele frequencies in the two sexes were compared using the exact test of population differentiation, implemented in GENEPOP software (Raymond & Rousset, 1995), well adapted for allele frequencies comparisons (Goudet *et al.*, 1996). The probability of type I error for the test was set at 0·05 and a Bonferroni correction for multiple tests used following the method of Dunn and Sidak (Ury, 1976). The analytical study of the evolution at X-linked loci under selection was performed with Mathematica 4.1 (Wolfram, 2001). Figures were obtained with R software (Ihaka & Gentleman, 1996).

3. Results and Discussion

The DXS8175 displays a total of 10 alleles, ranging from 10 to 20 repeats in all populations studied, a common feature for a dinucleotide microsatellite (Zhivotovsky *et al.*, 2003; Renwick *et al.*, 2001). Allelic frequencies in the 10 populations revealed similar distributions to those observed by Scozzari *et al.* (1997) in 30 populations from Europe, Africa, Asia and the Americas for the same marker. However, we observed a discrepancy in allele frequencies between males and females in all populations. We found that two populations (Akan and Sardinian) show a significant difference in allele frequencies (alleles 12, 14, 16, 17) between the sexes (Table 1). This difference remained significant after Bonferonni correction for multiple tests.

By evaluating the different scenarios that could have led to this pattern, it appeared that two main possible and testable hypotheses could explain the discrepancy in allele frequencies: an admixture event or sex-specific selection acting on a gene located on the X chromosome affecting the DXS8175 marker by a hitch-hiking effect.

(i) *Admixture*

To test whether admixture is the putative cause for the observed differences, we need to take into consideration the parental populations. If a population results from the admixture of two different founder groups with significant differences in allele frequencies, discrepancies are expected in the offspring generation (F1). The sex difference in the F1 generation is equal to half the difference among parental generation (with an opposite sign). In the more extreme case, where all males and females come from two different single populations (corresponding to 100% admixture), differences in allele frequencies between sexes in the parental populations have to be twice that seen in the F1 generation. In a moderate case, and so more realistic for human populations, if admixture is less than 100%, the frequency difference between sexes in the parental populations has to be much higher than twice the F1 difference. The necessary frequencies under this hypothesis were not observed in our study.

For example, the highest discrepancy in allele frequencies was observed in the Akan population, in which allele 16 showed the highest male to female frequency divergence: $\Delta = p_f - p_m = +0{\cdot}28$. In order to explain such a result, we could consider a fictitious case (100% admixture) where the most divergent populations in allele frequencies found in our sample would represent the parental populations. Considering males of the Orcadian and females of the Sardinian populations, the maximal value which could be reached in the F1 generation would be $\Delta = p_f - p_m = 0{\cdot}30/2 = 0{\cdot}15$ (allele 16_m: 0·59 and allele 16_f: 0·29), a value much lower than 0·28.

Others pairs of parental populations would result in lower expected divergence or in male frequencies being higher than female frequencies. Furthermore, the expected frequency patterns for others alleles would never match the observed Akan pattern. Using similar calculations, it can easily be shown that the discrepancy in sex frequency for the Sardinian population (for allele 12 or 17) cannot be explained by admixture. The admixture hypothesis seems not adapted to explain the sex frequency discrepancies in the Akan or the Sardinian population. The historical and demographic parameters necessary to achieve such a situation (100% admixture) from very distant populations are very unrealistic and an admixture event leading to the observed pattern of sex-allele discrepancies is unlikely. Such a difference in allele frequencies between males and females was never observed when we compared allele frequencies for other microsatellites on others parts of the genome (data not shown).

This argument can be extended to a multiple allele approach by noticing that the distance between sexes in the parental populations, as measured by F_{ST}, has to be approximately 4 times greater than the distance between the two sexes in the F1 generation. Such a distance was never achieved (data not shown).

(ii) *Selection*

(a) *Nearness of* VCX/Y *genes*

If the DXS8175 microsatellite is in the vicinity of a gene under selective pressures, this could explain such

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

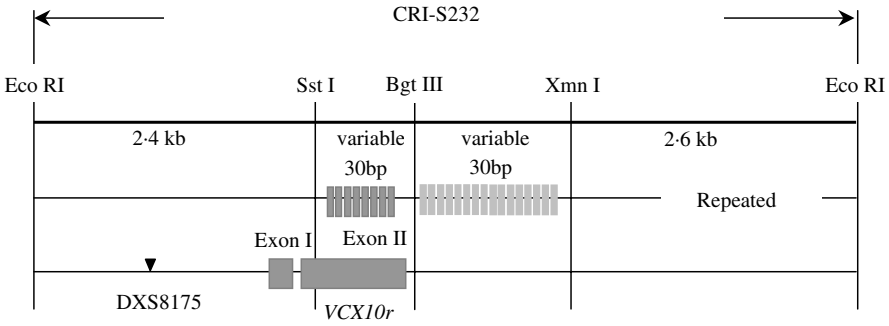


Fig. 1. Relative positions of DXS8175 microsatellite and *VCX10r* gene on the CRI-S232 genomic element (adopted from Lahn & Page, 2000; Fukami *et al.*, 2000).

a discrepancy in allele frequencies. By performing *in silico* investigation, we located the DXS8175 microsatellite ~2 kb upstream of the *VCX10r* gene (Fukami *et al.*, 2000; Lahn & Page, 2000; Balaresque *et al.*, 2003; Fig. 1) within a CRI-S232 duplicated element (Ballabio *et al.*, 1990; Li *et al.*, 1992), also called segmental duplications. The recombination fraction expected between the DX8175 microsatellite and the *VCX10r* gene is approximately 0·002%, confirming the association between the microsatellite and the gene.

(b) *Antagonistic allelic sex-specific selection on the X-chromosome*

Differential sex selection on the human X-chromosomes is an alternative explanation for the discrepancy in allele frequencies distributions in males and females. Population genetic theory shows that stable polymorphisms can be maintained by selection at sex-linked loci when alleles are antagonistically selected in the two sexes (see for example Crow & Kimura, 1970, p. 278). Here we will show that in such cases large differences between allele frequencies in males and females can be obtained at equilibrium. Consider a single locus on the X-chromosome with two alleles (A1 and A2) under zygotic selection with one allele (A1) being deleterious in males (heterogametic sex) but advantageous in females (homogametic sex). Selective effects of A1/A2 alleles were parameterized by a selective coefficient s reducing YA1 male fitness to $w1=1-s$ (YA2 male fitness being $w2=1$), a selective coefficient t reducing homozygous A2A2 female fitness to $w22=1-t$ (homozygous A1A1 female fitness being $w11=1$), and a dominance parameter h (i.e. heterozygous A1A2 female fitness was $w12=1-ht$). We supposed that the fitness of heterozygous females was intermediate between those of homozygous females (i.e. h varying between 0 and 1), thus excluding under- or over-dominance phenomena. This fitness model was a slight modification of the one used by Rice (1984) and was adopted because of its symmetry, all parameters (s, t and h) varying between 0 and 1.

The precise analysis of the evolution at X-linked loci under selection has been done by many authors (see for example Cannings, 1967; Crow & Kimura, 1970; Rice, 1984). The existence of a stable polymorphic equilibrium depended on the three parameters (s, t, h) as shown in Fig. 2. Globally, the region for a stable polymorphism was reduced with increase in the dominance parameter h.

Using our fitness model, equilibrium A1 frequencies in females ($\hat{p}_f$) and in males ($\hat{p}_m$) were:

$$\hat{p}_f = \frac{t[2-h(2-s)]-s}{2t[1-h(2-s)]}, \tag{1}$$

$$\hat{p}_m = \frac{(1-s)[t(2-h(2-s))-s]}{s^2+t[2(1-s)-h(2-s)^2]}. \tag{2}$$

Combining these equations (which are equivalent to equations 7 and 8 of Rice, 1984), the difference in allelic frequencies, $\Delta=\hat{p}_f-\hat{p}_m$, was:

$$D = \frac{s[t(2-h(2-s))-s][s-(2-s)ht]}{2t[1-h(2-s)][s^2+t\{2(1-s)-h(2-s)^2\}]}. \tag{3}$$

This difference was strictly positive in the stable polymorphic equilibrium region, i.e. equilibrium A1 female frequency $\hat{p}_f$ was always higher than equilibrium A1 male frequency $\hat{p}_m$. The dominance parameter h had little influence on the difference between $\hat{p}_f$ and $\hat{p}_m$. As a result, we restricted our analysis to the case of a complete dominance of A1 over A2 in females (i.e. $h=0$). In this situation, the discrepancy between female and male A1 frequencies at equilibrium reduced to:

$$D = \frac{s^2(2t-s)}{2t[s^2+2t(1-s)]}. \tag{4}$$

The exact value of the excess of A1 allele in females (as measured by Δ) obviously depended on the selective coefficients s and t, and can be visualized in the (s, t) parameters space as lines with equal Δ values

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

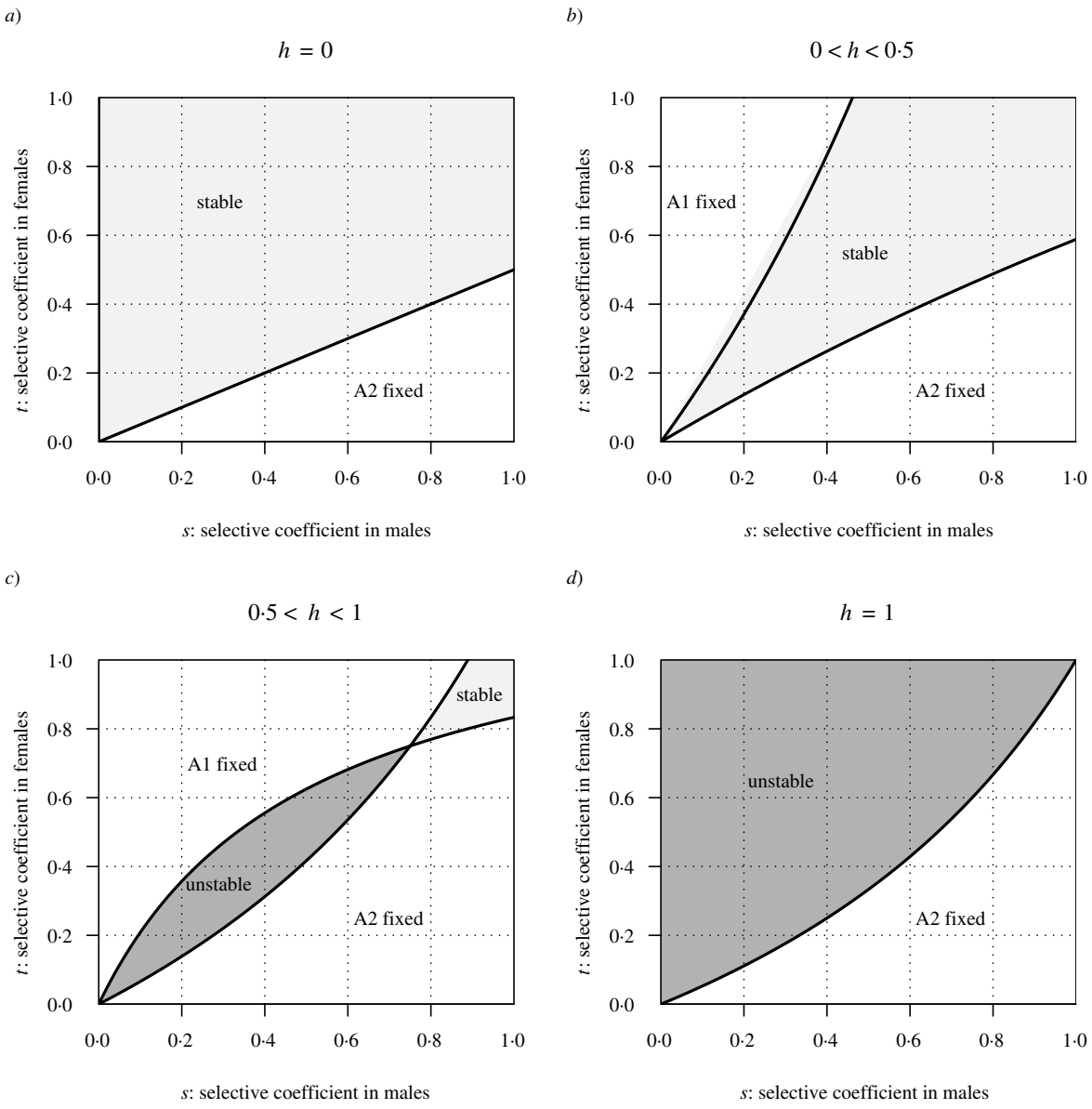


Fig. 2. Equilibrium at X-linked loci under antagonistic selection in the two sexes (allele A1 being deterious in males and advantageous in females). White areas correspond to monomorphic equilibrium (A1 fixed or A2 fixed), light grey areas correspond to stable polymorphic equilibrium, and dark grey areas correspond to unstable polymorphic equilibrium. Stability is possible when $(1-ht)(2-s)>2(1-s)$ and $(1-ht)(2-s)>2(1-t)$. (*a*) When A1 is dominant in females ($h=0$), two equilibrium states are possible: a stable polymorphic state and a monomorphic state (A2 fixed). (*b*) When A1 is partially dominant in females ($0<h<0{\cdot}5$), there are two monomorphic states (A1 fixed or A2 fixed) and a reduced area for the polymorphic state. (*c*) When A1 is partially recessive in females ($0{\cdot}5<h<1$), an unstable polymorphic state appears. (*d*) When A1 is partially recessive in females ($h=1$), polymorphism is unstable, thus A2 always becomes fixed.

(Fig. 3). Globally, Δ increased with increasing s and t (i.e. under high selection) and tended to 0·5 with males being only YA2 and females being only A1A2 heterozygotes. For more realistic selective parameters (i.e. lower s and t values), Δ was strongly decreased and would be impossible to detect for selective coefficients s and t lower than 0·2 ($\Delta<0{\cdot}05$).

With this model, we showed that sexual antagonism is sufficient to create differences in allele frequencies in males and females. The more extreme pattern we observed (Akan population) can be explained in this framework, although it requires very high selective coefficients both in males and in females ($s=0{\cdot}748$ and $t=0{\cdot}675$ for $\hat{p}_m=0{\cdot}169$ and $\Delta=0{\cdot}277$). We must, however, note (1) that the true Δ value in the Akan population might be lower than that observed in our sample, and (2) that the observed Δ value might not be the equilibrium one. Classical population genetics results show that an initial difference in allele frequencies in males and females for a neutral locus on the X-chromosome needs some generations to disappear with fluctuations around 0 from generation to

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

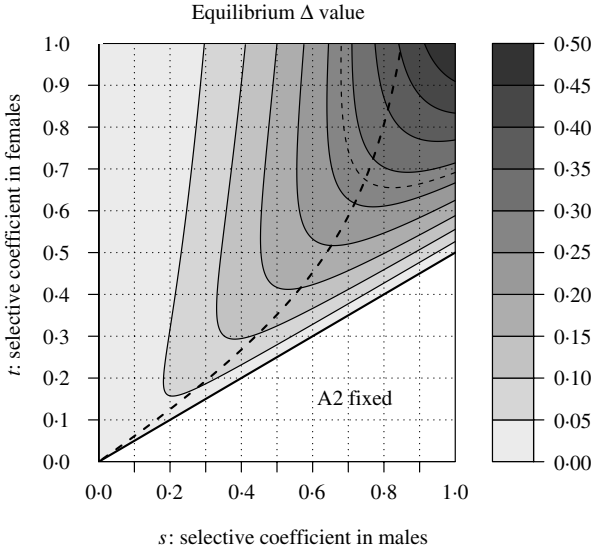


Fig. 3. Equilibrium $\Delta = \hat{p}_f - \hat{p}_m$ value in selective coefficients (s, t) space with complete dominance of A1 in females ($h = 0$). The stable polymorphic equilibrium region is delimited by the thick plain line: A2 gets fixed for parameters values below the line. Thin plain lines represent (s, t) values yielding a given Δ values (shown values from 0·05 to 0·45). The thin dashed line represents $\Delta = 0{\cdot}277$. The thick dashed line represents an equilibrium A1 frequency in males $\hat{p}_m = 0{\cdot}169$. The dashed line intersect at $s = 0{\cdot}748$ and $t = 0{\cdot}675$.

generation (being successively positive and negative). Adding sexual antagonistic selection (as in our model) results in a translation of the equilibrium value from 0 (in the neutral case) to a positive value. In the non-equilibrium phase, fluctuations may well transiently increase Δ to a value much higher than its final equilibrium value. Random genetic drift or moderate admixture could be recurrent sources of displacement from the equilibrium. In some situations (especially when the discrepancy is reduced by such phenomena), selection could transiently drive the system to higher values than expected at equilibrium.

The model with an advantage in the homogametic sex could explain the pattern observed in the Akan population. Similarly, the reverse frequency pattern observed in the Sardinian population (higher frequency of allele 12 or 17 in males than in females) could be explained by the reverse model where A1 is advantageous in males and deleterious in females.

(c) VCX/Y *genes: target for selection*

We showed that a differential selection between the two sexes on a gene in the vicinity of the DXS8175 microsatellite would create such a difference in allelic frequency distribution at equilibrium in a given population. The *VCX10r* gene, a member of the *VXC/Y* gene family, is a good candidate as expression analysis showed that all copies of the *VCX/Y* gene family have a testis-specific expression, probably in the germ cells (Fukami *et al.*, 2000; Lahn & Page, 2000). Their involvement in female reproductive functions remains to be defined but, to our knowledge, no expression studies of the *VCX/Y* gene family have been performed in fetal ovary tissue and therefore a role of *VCX* members during oogenesis cannot be ruled out. A similar sex-specific selective process has been observed in *Drosophila* for sexually antagonistic genes (Rice, 1992; Chippindale *et al.*, 2001) in which some genes are advantageous in the heterogametic sex whereas they are disadvantageous in the homogametic sex. Genes located on the human X-chromosome constitute potential and interesting targets on which antagonistic selective pressures between both sexes could be acting (Gibson *et al.*, 2002).

A second class of model could involve selection acting at the gametic level rather than at the zygotic levels, resulting in similar sex frequency discrepancies provided that selection acts antagonistically in the two sexes (data not shown).

A third class of model with alleles acting as a segregation distorter in males but being deleterious in females would probably also result in a difference in allele frequency at equilibrium. As we have shown, the DXS8175 microsatellite is located near the *VCX* gene within the duplicated element CRI-S232. Interestingly, in a recent study Lahn & Page (2000) and Lahn *et al.* (2001) reported that the *VCX/Y* genes could act as meiotic distorters. Their statement is mainly based on two observations: (i) the molecular characteristics of *VCX/Y* genes that recalled those of the fruitfly X-linked Stellate (Ste) and Y-linked crystal, which are meiotic drive elements in *Drosophila melanogaster* (Belloni *et al.*, 2002); (ii) recombination between CRI-S232 elements is known to cause frequent deletions in the X-chromosome short arm, resulting in steroid sulfatase deficiency (X-ichthyosis). This could be a satisfactory explanation for an old speculation of male bias among the offspring of ichthyosis carrier females reported in some human populations (Filippi & Meera Khan, 1968; Gladstein *et al.*, 1979).

(iii) *Differences among populations*

The observed differences in allele frequencies between sexes, dependent on the allele or the population, could be due to several causes including differential demographic histories of some populations associated with variation in linkage disequilibrium (LD) levels (Ardlie *et al.*, 2002): if the mutation arrived more recently in one population, there is a higher LD between the gene under selection and the microsatellite, and therefore it can be detected through the microsatellite polymorphism in this population. In a population where the mutation arrived earlier, LD has been reduced through recombination between the selected gene

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

and the microsatellite and it is very difficult to detect such an effect. This could explain why such an association is only detectable in 2 populations in 10. Among these, the Sardinian population is known to show high levels of linkage disequilibrium (Taillon-Miller *et al*., 2000; Angius *et al*., 2002).

Moreover, local selection on the *VCX* gene could also explain differences between populations. The instability of this region through misalignment between duplicated elements could lead to a copy number polymorphism of *VCX* genes: these differences among populations have been documented for other gene families (Trask *et al*., 1998), suggesting that variable selective patterns may be expected across populations.

4. Conclusions and perspectives

We have shown that discrepancies in allele frequencies between males and females are probably due to sex-specific selective pressures. The model of Rice (1984) seems well adapted to illustrate intra-locus antagonistic pressures acting on sex-specific-linked alleles. Although this model with two alleles is clearly an over-simplification of the reality, it provides an interesting framework to explain our data. The concept of antagonistic intra- or inter-locus selective pressures becomes especially relevant when the candidate loci are polymorphic and are part of a multigenic family in which different members act in synergy. Recent results on the human genome reported that segmental duplications constitute approximately 5% of the human genome and that all copies of each family share about 90–100% similarity (ISHGC, 2001; Samonte & Eichler, 2002). These large blocks of sequence similarity provide the substrate for aberrant recombination leading to variation in copy number among individuals or populations (Menashe *et al*., 2003). This observation underlines the fact that duplicated sequences are part of an ongoing process that results in a novel form of large-scale variation in the human genome (Eichler, 2001), which may be subject to complex selective patterns.

We would like to thank the following collaborators for DNA samples: A. Chaventre, G. F. De Stefano, A. Baali, M. Cherkaoui, J. M. Dugoujon, G. Bellis, P. Kouamé, A. Pitte, M. S. Isaad, G. Larrouy and A. Sevin. Thanks go also to R. Jambou-Clerc for his technical assistance; and to P. H. Gouyon and A. Sibert for fruitful discussions. This work was supported by the Research group 'interaction génomique' (CNRS).

References

Angius, A., Bebbere, D., Petretto, E., Falchi, M., Forabosco, P., Maestrale, B., *et al*. (2002). Not all isolates are equal: linkage disequilibrium analysis on Xq13.3 reveals different patterns in Sardinian sub-populations. *Human Genetics* **111**, 9–15.

Ardlie, K. G., Kruglyak, L. & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Review Genetics* **3**, 299–309.

Balaresque, P., Toupance, B., Heyer, E. & Crouau-Roy, B. (2003). Evolutionary dynamics of duplicated microsatellites shared by sex chromosomes. *Journal of Molecular Evolution* **57**, S128–S137.

Ballabio, A., Bardoni, B., Guioli, S., Basler, E. & Camerino, G. (1990). Two families of low-copy-number repeats are interspersed on Xp22.3: implications for the high frequency of deletions in this region. *Genomics* **8**, 263–270.

Belloni, M., Tritto, P., Bozzetti, M. P., Palumbo, G. & Robbins, L. G. (2002). Does Stellate cause meiotic drive in *Drosophila melanogaster*? *Genetics* **161**, 1551–1559.

Cannings, C. (1967). Equilibrium, convergence and stability at a sex-linked locus under natural selection. *Genetics* **56**, 613–618.

Chippindale, A. K., Gibson, J. R. & Rice, W. R. (2001). Negative genetic correlation for adult fitness between sexes reveals ontogenetic conflict in Drosophila. *Proceedings of the National Academy of Sciences of the USA* **98**, 1671–1675.

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York.

Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics* **17**, 661–669.

Filippi, G. & Meera Khan, P. (1968). Linkage studies on X-linked ichthyosis in Sardinia. *American Journal of Human Genetics* **20**, 564–569.

Fukami, M., Kirsch, S., Schiller, S., Richter, A., Benes, V., Franco, B., *et al*. (2000). A member of a gene family on Xp22.3, VCX-A, is deleted in patients with X-linked nonspecific mental retardation. *American Journal of Human Genetics* **67**, 563–573.

Gécz, J. & Mulley, J. (2000). Genes for cognitive function: developments on the X. *Genome Research* **10**, 157–163.

Gibson, J. R., Chippindale, A. K. & Rice, W. R. (2002). The X chromosome is a hot spot for sexually antagonistic fitness variation. *Proceedings of the Royal Society of London, Series B* **269**, 499–505.

Gladstein, K., Shapiro, L. J. & Spence, M. A. (1979). Estimating sex-ratio biases in X-linked disorders: is there an excess of males in families with X-linked ichthyosis? *American Journal of Human Genetics* **31**, 1091–1094.

Goudet, J., Raymond, M., de Meeus, T. & Rousset, F. (1996). Testing differentiation in diploid populations. *Genetics* **144**, 1933–1940.

Graves, J. A. M. & Delbridge, M. L. (2001). The X: a sexy chromosome. *Bioessays* **23**, 1091–1094.

Grimaldi, M. C. & Crouau-Roy, B. (1997). Microsatellite allelic homoplasy due to variable flanking sequences. *Journal of Molecular Evolution* **44**, 336–340.

Hurst, L. D. & Randerson, J. P. (1999). An eXceptional chromosome. *Trends in Genetics* **15**, 383–385.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

International Sequencing Human Genome Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Lahn, B. T. & Page, D. C. (2000). A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. *Human Molecular Genetics* **9**, 311–319.

Lahn, B. T., Pearson, N. M. & Jegalian, K. (2001). The human Y chromosome, in the light of evolution. *Nature Review Genetics* **2**, 207–216.

Li, X. M., Yen, P. H. & Shapiro, L. J. (1992). Characterization of a low copy repetitive element S232 involved in the generation of frequent deletions of the distal short arm of the human X chromosome. *Nucleic Acids Research* **20**, 1117–1122.

Malaspina, P., Ciminelli, B. M., Viggiano, L., Jodice, C., Cruciani, F., Santolamazza, P., *et al*. (1997). Characterization of a small family (CAIII) of microsatellite-containing sequences with X-Y homology. *Journal of Molecular Evolution* **44**, 652–659.

Menashe, I., Man, O., Lancet, D. & Gilad, Y. (2003). Different noses for different people. *Nature Genetics* **34**, 143–144.

Raymond, M. & Rousset, F. (1995). GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248–249.

Renwick, A., Davison, L., Spratt, H., King, J. P. & Kimmel, M. (2001). DNA dinucleotide evolution in humans: fitting theory to facts. *Genetics* **159**, 737–747.

Rice, W. R. (1984). Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38**, 735–742.

Rice, W. R. (1992). Sexually antagonistic genes: experimental evidence. *Science* **256**, 1436–1439.

Saifi, G. M. & Chandra, H. S. (1999). An apparent excess of sex- and reproduction-related genes on the human X chromosome. *Proceedings of the Royal Society of London, Series B* **266**, 203–209.

Samonte, R. V. & Eichler, E. E. (2002). Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics* **3**, 65–72.

Scozzari, R., Cruciani, F., Malaspina, P., Santolamazza, P., Ciminelli, B. M., Torroni, A., *et al*. (1997). Differential structuring of human populations for homologous X and Y microsatellite loci. *American Journal of Human Genetics* **61**, 719–733.

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., *et al*. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics* **25**, 324–328.

Trask, B. J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J., *et al*. (1998). Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Human Molecular Genetics* **7**, 13–26.

Ury, H. K. (1976). A comparison of four procedures for multiple comparisons among means (pairwise contrasts) for arbitrary sample sizes. *Technometrics* **18**, 89–97.

Wang, P. J., McCarrey, J. R., Yang, F. & Page, D. C. (2001). An abundance of X-linked genes expressed in spermatogonia. *Nature Genetics* **27**, 422–426.

Wolfram, S. (2001). *Mathematica* (*version 4.1*)*: A System for doing Mathematics by Computer*. Reading, MA: Addison-Wesley.

Zhivotovsky, L. A., Rosenberg, N. A. & Feldman, M. W. (2003). Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *American Journal of Human Genetics* **72**, 1171–1186.

Genet. Res., Camb. (2004), **83**, *pp.* 177–188. *With* 4 *figures.*
DOI: 10.1017/S0016672304006780 Printed in the United Kingdom

Use of regression methods to identify motifs that modulate germline transcription in *Drosophila melanogaster*

EMILY HONEYCUTT[1] AND GREG GIBSON[1,2]*
[1] *Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695-7566, USA*
[2] *Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614, USA*

(*Received 18 July 2003 and in revised form 29 December 2003*)

Summary

Identification of *cis*-regulatory motifs has been difficult due to the short and variable length of the sequences that bind transcription factors. Using both sequence and microarray expression data, we present a method for identifying *cis*-regulatory motifs that uses regression trees to refine results from simple linear regression of expression levels on motif counts. Analysis of expression patterns from two separate datasets for genes showing significant differences in expression between the sexes in *Drosophila melanogaster* resulted in a model that identified known binding sites upstream of genes that are differentially expressed in the germline. We obtained a strong result for motif TCGATA, part of the larger, characterized binding site of *dGATAb* protein. We also identified an uncharacterized motif that is positively associated with sex-biased expression and was assembled from smaller motifs grouped by our model. A regression tree model provides a grouping of independent variables into multiple linear models, an advantage over a single multivariate model. In our case, this grouping of motifs suggests binding sites for cooperating factors in sex-specific expression, as well as a way of combining smaller motifs into larger binding sites.

1. Introduction

The search for DNA regulatory motifs has been the focus of much recent research, with various methods being employed in motif discovery. Detection of transcriptional regulatory motifs in the upstream region of genes has presented a real challenge because transcription factor binding regions tend to be short, discontinuous, and quite variable. *Saccharomyces cerevisiae* has frequently been the organism of choice for development of methods that identify transcription factor binding sites since many binding motifs have already been experimentally characterized in this organism. Nevertheless, most methods of motif detection have found limited success, often resulting in a high rate of false positives (Werner, 2002). In higher eukaryotes, the structure of regulatory motifs is more complex and less well defined, making the development of new methods and verification of results even more difficult.

Before access to the sequence of multiple whole genomes, many motif-detection methods involved statistical approaches to the creation of weight matrices. A weight matrix is derived from a number of short sequences known to be bound by a given transcription factor, and then the matrix is used to search a sequence or a set of sequences for a match to that motif. Examples include MatInd and MatInspector (Quandt *et al.*, 1995) and FastM (Klingenhoff *et al.*, 1999). Searches that use weight matrices have a very high rate of false positives, but results have improved when they are used in combination with another method such as phylogenetic comparison (Guha Thakurta *et al.*, 2002).

Alternative methods for motif detection involve direct comparison of regulatory regions, either between genes thought to be co-regulated or between orthologous genes from closely related species. The Gibbs sampling method, which utilizes a modified Expectation Maximization (EM) algorithm (Lawrence *et al.*,

* Corresponding author. BRC, 1500 Partners II, 840 Main Campus Drive, NC State University, Raleigh, NC 27695-7566, USA. Tel: +1 (919) 5132512. Fax: +1 (919) 5153355. e-mail: ggibson@unity.ncsu.edu

1993), has been used in the AlignACE program to return over-represented motifs in co-regulated gene clusters and has found some success (Hughes *et al*., 2000; Manson-McGuire *et al*., 2000). Advanced application of Gibbs sampling methods in this context continues to hold promise, particularly in microorganisms (Liu *et al*., 2001). With the increasing availability of whole genome sequences of closely related species, the phylogenetic comparison of regulatory regions has increased. Comparisons between human and mouse regulatory sequence showed that phylogenetic footprinting can reduce the sequence space to be searched for transcription factor binding sites (Wasserman *et al*., 2000). Rajewsky *et al*. (2002) recovered approximately 75% of the regulatory sites compiled for *E. coli* using interspecies comparisons. Issues still remain as to how best to choose the species for comparison and how many are required to produce meaningful results. A recent study using proteobacteria takes a formal look at these issues (McCue *et al*., 2002).

A somewhat different strategy for detection of transcription factor binding motifs searches for clusters of motifs in upstream sequences (Berman *et al*., 2002; Halfon *et al*., 2002; Markstein *et al*., 2002; Rebeiz *et al*., 2002). These clustering methods require prior knowledge of characterized sites and are targeted more towards finding genes regulated by factors binding to the clusters rather than identifying the clusters themselves. Another combinatorial approach for finding synergistic motifs by Pilpel *et al*. (2001) also requires knowledge of known regulatory motifs. An underlying assumption in several of the above analyses is that binding motifs are redundant in the promoter region, as in the *Drosophila* yolk protein genes (Piano *et al*., 1999) and the *Drosophila* eve stripe 2 gene (Berman *et al*., 2002).

Capitalizing on this redundancy property, a recent study by Bussemaker *et al*. (2001) fitted a linear model of the logarithm of the expression ratio under two different experimental conditions to the counts of oligomers upstream of a set of genes. By first determining statistically significant motifs with a single-motif model of the data, a model describing the additive effects of multiple motifs can then be created. We incorporate this method by identifying the statistically significant motifs through the single-motif model, but instead of building a single additive model for a given experiment, we use the significant motifs to build regression trees. Our regression trees allow for multiple linear models to describe the data based on the prevalence of certain motifs and have the potential to uncover hierarchical or non-additive relationships between motifs.

Regression trees were originally used to generate predictive models of regression estimates. They were developed to deal with continuous-class learning problems (Quinlan, 1992; Wang & Witten, 1997), and combine a classical decision tree with linear regression estimations at the leaves of the tree. The prediction accuracy of regression trees is competitive with linear regression methods (Breiman *et al*., 1984), but the real advantage of the regression tree method lies in the model representation. The decision nodes and their position in the tree indicate which nodes together significantly affect the predicted values. We show that they can be used to identify prospective regulatory motifs bound by transcription factors, as well as combinations of motifs that aggregate to form larger motifs.

Other biological studies have also capitalized on the classificatory property of regression trees. For example, a recent investigation into the nesting habitats of smallmouth bass used regression trees to give a hierarchical view of habitat conditions that affect the smallmouth bass's choice of nesting site (Rejwan *et al*., 1999). Similarly, they have been used to identify the most predictive variables for patients who undergo angiography (Pilote *et al*., 1996). In this study, regression trees identified age as the most important variable. However, in younger patients availability of the angiography procedure was the next most predictive factor, while age was still the second most predictive factor in older patients. This illustrates the ability of regression trees to separate, or group together, cooperating factors under given circumstances.

In our model, we are using counts of binding motifs as the decision points in the tree. The decision nodes in the tree look at the counts of motifs of length k (k-mers) taken from the upstream region of a given gene. The change in estimated regression values between the leaf nodes indicates whether a combination of motifs is associated with the regulation of genes. As with the aforementioned studies, we are not using the regression tree model in its classical sense as a predictor of response, but instead to identify the predictive variables, namely regulatory motifs.

In this study, we searched for transcription factor binding motifs of genes that show sex-biased expression. Our previous study on sex, genotype and age (Jin *et al*., 2001) (subsequently referred to as the aging dataset) showed evidence for between one-third and two-thirds of the *Drosophila* transcriptome having sex-biased expression. Comparisons with *tudor* mutant animals that lack ovaries and testes have since demonstrated that most of the differences in gene expression between reproductively mature adult male and female flies is due to germline expression (Arbeitman *et al*., 2002; Parisi *et al*., 2003). To obtain a larger number of these differentially expressed genes for our analysis, we supplemented the aging dataset (Jin *et al*., 2001) with data from another experiment that tested the effects of nicotine on gene expression in flies of both sexes (G. Passador-Gurgel and G.G., in

preparation: this dataset is subsequently referred to as the nicotine dataset). Although two different clone sets were used to generate the data, a high concordance in the predicted motifs was observed, and this independent replication confirms that regression tree methods may be a valuable new approach to characterization of regulatory motifs.

2. Materials and methods

(i) *Gene selection from microarray experiments*

The genes used for analysis of sex-biased expression are from two datasets: the aging dataset (Jin *et al*., 2001) and the nicotine dataset (G. Passador-Gurgel and G.G., in preparation). The aging array experiment used a split-plot experimental design and tested for sex as a fixed effect using a mixed-models approach (Wolfinger *et al*., 2001). Array set-up and subsequent analysis for the nicotine experiment was done similarly, with 48 two-sample arrays involving three wild-type genotypes, two sexes and treatment (control versus drug) as fixed effects. The set of genes for the nicotine experiment was 4856 genes from the Drosophila Gene Collection (DGC), which were independently identified and amplified from those of the White collection used in the aging experiment. From each experiment, genes with a P value of $<0{\cdot}0001$ resulting from the test for sex effects were chosen for use in this analysis. The lists of genes from both datasets and their associated expression difference are available at http://statgen.ncsu.edu/ggibson/SupplInfo/SexSpecificList.txt

(ii) *DNA sequence motifs*

All possible motifs of length 6 were generated. Initially, we extracted counts of all possible 7-mers of the 250 differentially expressed genes from the aging dataset (Jin *et al*., 2001). Since five of the eight most significant motifs from the linear regression contained the sequence TCGATA, all subsequent analyses were conducted on 6-mer motifs. Motifs were combined with their reverse complement and the motif having the higher lexicographic order was chosen to represent the pair. No allowance for variability in the motif sequence was made. For each gene selected, the 1000 base-pair (bp) sequence upstream of the translation start site (ATG) was extracted from the Version 2 annotation of the *Drosophila* genome sequence at NCBI (March 2002, http://www.ncbi.nlm.nih.gov). This sequence includes variable lengths of 5′ untranscribed and untranslated leader sequences, which are as yet typically uncharacterized in *Drosophila*. Although enhancers in the fly genome can be several kilobases away from the translation start site, the 1000 bp upstream sequence was chosen for two reasons. First, testis-specific promoters in *Drosophila* are usually close to the start site (Arnosti, 2003). Secondly, as more sequence is added to the analysis, the signal-to-noise ratio of regulatory to non-functional motifs probably drops, and with a large number of genes we surmised that we would be most likely to find common motifs in the upstream 1 kb region. This approach is not intended to identify all the enhancer elements that regulate sex-specific gene expression in *Drosophila*, but rather to focus on those located proximal to the transcription start site.

For each gene, all motifs were counted in the upstream 1 kb sequence (allowing overlap, namely 995 motif counts per gene). All work to extract sequence, generate motifs and count motifs was done via Perl scripts.

(iii) *Single-motif linear regression*

The first stage of analysis uses a simple linear regression model to fit single-motif counts and expression data. The model is defined as:

$$Y=\beta_0+\beta_1 X$$

where Y is the base 2 logarithm of the expression difference between females and males. A positive Y indicates greater expression in females; a negative Y indicates greater expression in males. X is the count of a given motif. All genes chosen as significantly differentially expressed between the sexes (in either direction) were fitted to the model. β_1 is the relative increase or decrease in expression difference caused by each additional copy of the motif in the upstream region of the gene, and β_0 is the grand mean expression difference.

Both the nicotine and the aging datasets were run through simple linear regression. To account for the large number of motifs (2080), application of the Bonferroni correction set the experimentwise significance cutoff from regression of expression level on motif count for $\alpha=0{\cdot}05$ at $P=2{\cdot}4\times10^{-5}$. Permutation tests provided independent verification of the appropriateness of this cutoff, but for some analyses we included simply the top 20 motifs as these included a few motifs that were close to the cutoff in both datasets.

(iv) *Regression and decision trees*

Single-motif linear regression was used primarily as a data reduction technique. Motifs with a P value below the Bonferroni-corrected values were considered most likely to affect sex-biased expression and were therefore used in training and validation of the regression and decision tree models.

Regression and decision tree models were built and trained with publicly available Weka software (Witten & Frank, 1999) available at http://www.cs.waikato.ac.nz/ml/weka/. Data from the nicotine experiment

Table 1. *The most significant sex-specific motifs from single-motif regression for both the nicotine and aging datasets*

	Nicotine dataset			Aging dataset		
Rank	Motif	*P* value	sign[a]	Motif	*P* value	sign[a]
1	**TCGATA**	1.4e-19	+	**TCGATA**	1.2e-12	+
2	**CGATAG**	2.5e-11	+	**ATCGAT**	0·0000013	+
3	**ATCGAT**	3.7e-10	+	**ATATCG**	0·0000024	+
4	GGTCAC	0.000000050	+	ACGACG	0·000065	+
5	**ATATCG**	0·00000019	+	AGTCGC	0·000092	+
6	**ACACTG**	0·00000024	+	CGCAAC	0·00014	+
7	CACGTG	0·00000033	+	**CGATAG**	0·00016	+
8	TAAAAA	0·0000012	+	CCAAAG	0·00021	−
9	GGCGCA	0·0000022	+	GCAACG	0·00021	+
10	CCGTTA	0·0000030	+	**ACACTG**	0·00038	+
11	GTCACA	0·0000032	+	CACGCA	0·00058	+
12	AAGAAG	0·0000032	+	**GCACGC**	0·00063	+
13	CGCACG	0·0000057	+	CCTTTC	0·00066	−
14	AGACTC	0·0000073	−	**AGTGTG**	0·00075	+
15	CGGTAA	0·0000161	+	AGGGCC	0·00099	−
16	TTAAAA	0·000016	+	GTGTGA	0·0013	+
17	AAAATA	0·000019	+	ATCGAC	0·0015	+
18	**AGTGTG**	0·000022	+	ATTCGC	0·0015	+
19	GCGCAC	0·000022	+	AGAAGA	0·0016	+
20	**GCACGC**	0·000028	+	ACTACG	0·0020	+

Motifs in common between the two sets are indicated in bold.
[a] Positive coefficients indicate that the motif is associated with increased transcription in females. Negative coefficients indicate that the motif is associated with increased transcription in males.

were used to train the models and data from the aging experiment were used for model validation. Specifically, the regression trees were built with the M5 software using a $-O\ r$ option. The decision trees were built with the J48 software using the $-R$ option to reduce error pruning and the $-M$ option to vary the minimum number of instances per leaf.

The models were built from motifs as follows:

Model 1: Motifs that were above Bonferroni-corrected significance cutoff from single-motif regression and were seen >4% of the time within 20 bp of TCGATA/TATCGA (8 total).
Model 2: Motifs seen >5% of time within 20 bp of TCGATA/TATCGA (25 total).
Model 3: The most significant motifs from single-motif regression at or below Bonferroni-corrected cutoff (20 total).
Model 4: Combination of 20 most significant motifs from single-motif regression and 20 motifs most often seen within 20 bp of TCGATA/TATCGA.

3. Results

(i) *Identification of female-specific regulatory motifs*

The first stage of the analysis searched for motifs that may contribute to male- or female-specific gene expression in adult flies using linear regression of expression difference against motif count in the promoters of differentially expressed genes (Bussemaker *et al*., 2001). Table 1 shows the top 20 motifs after linear regression with the two different datasets. The significance threshold for regression of motif count on expression difference after Bonferroni correction is approximately $2{\cdot}4 \times 10^{-5}$. Three motifs exceed this threshold in the aging dataset, and 19 in the larger nicotine dataset. Several results stand out. Most noticeably, the two experiments converge on a similar set of motifs, with the three most significant motifs found in the aging dataset also being found within the five most significant motifs resulting from analysis of the nicotine dataset. Three other motifs are also common between each dataset's list of 20 most significant motifs. Additionally, the motif TCGATA/TATCGA is at a much higher significance level than any other motif in both datasets, with a *P* value of 10^{-19}. Lastly, almost all the motifs are associated with female-biased gene expression, and no case of a male-specific motif was replicated in both datasets. Representative linear regression profiles shown in Fig. 1 also highlight the point that none of the motifs is either necessary or sufficient for sex-specific gene expression: some genes with multiple copies of TCGATA are actually male-biased, and many female-specific genes lack the motif within 1 kb of the translation start site.

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

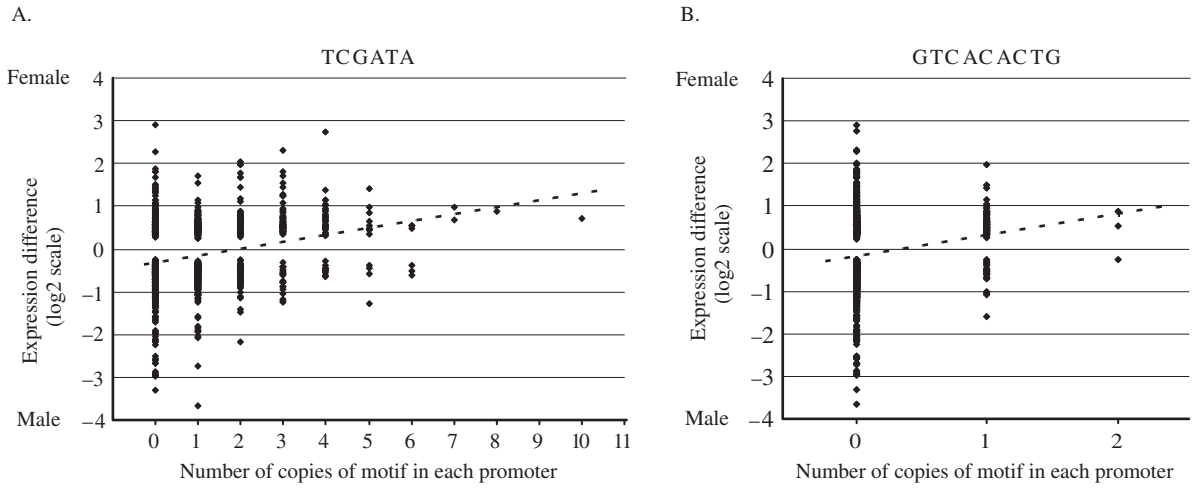


Fig. 1. Linear regression of expression difference on motif counts. Each diamond represents the normalized difference between gene expression in adult females and males on a log 2 scale, given the number of copies of the indicated motif (A: TCGATA; B: GTCACACTG) in the nicotine dataset. Only genes that are significantly different between the sexes are included. On this scale, 1 represents a two-fold difference, 2 a four-fold difference, and so on. Dashed lines shows linear regression fit. Female-biased genes are towards the top.

The most significant motif, TCGATA, is part of a known binding site for the *dGATAb* (SERPENT) protein, which enhances transcription of yolk proteins in *Drosophila* females (Lossky & Wensink, 1995). The entire binding site has been characterized as GCTATCGATAGC, which highlights the fact that TCGATA and its reverse complement TATCGA have a 4 bp overlap. The combined 8-mer is palindromic, a characteristic that is known to increase the affinity of binding sites for transcription factors but usually associated with head-to-tail dimerization of individual binding sites (Drouin *et al*., 1992). Observations of all TCGATA/TATCGA pairs in the upstream regions of the genes being analysed show that this 4 bp overlap occurs in 29% of these incidences. A chi-square contrast of the incidence of the palindrome in female-biased versus male-biased and non-sex-biased genes provides compelling evidence ($P<0{\cdot}001$) that this palindrome is strongly associated with sex-biased expression, and, specifically, that it is female-specific. A concern is that the prevalence of this overlap artificially inflates the motif counts for TCGATA and enhances its significance in the single motif regression results. However, the overlap of the motif with itself into an 8 bp palindrome creates a more likely binding site, so counting the 6-mer twice simply aids in this discovery.

The high significance of TCGATA could also be a result of its pairing with itself as a composite binding site for a transcription factor pair or for multiple fingers of a zinc-finger binding protein such as SERPENT. Since over half of the DNA-binding proteins in *Drosophila* are zinc-finger proteins (Adams *et al*., 2000), we assumed that close proximity of binding motifs would often allow for the possible binding of multiple-fingers, which prompted us to count all the non-overlapping motifs within 20 bp on either side of TCGATA/TATCGA. TCGATA was found within 20 bp of itself at a greater frequency than any other motif (Table 2), supporting the idea that it often forms a composite binding site.

(ii) *Use of regression trees to identify interacting motifs*

The most significant motifs from the single-motif regression can be used to create an additive model that accounts for the combinatorial nature of cooperative and competitive binding of transcription factors. However, in a single additive model, each included motif is assumed to affect every gene's predicted expression level. This is not always the case. Different combinations of motifs may have dramatically different effects on transcription. Consider a combination of three binding motifs that cause increased binding affinity, and thus an increase in expression levels. If one of those binding motifs is replaced by a different motif, transcriptional repression could result. Regression trees have the potential to account for these types of occurrences. Nodes at the top of the tree indicate motifs that most correlate with expression. As a path is traversed through the tree, a combination of motifs affecting expression is discerned. The values at the leaves of the tree show how the path increases or decreases the expression difference. In our case, an increase in expression difference between paths indicates that transcription tends to be enhanced in females. We are using the regression tree as a model for finding important motifs identified by nodes in the tree. A more conventional use of

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

Table 2. *Motifs within 20 base-pairs of TATCGA/TCGATA*[a]

Rank	Motif	Number	Percentage
1	TATCGA	202	18·05
2	ATCGAT	155	13·85
3	CGATAG	126	11·26
4	CGATAA	120	10·72
5	AATCGA	110	9·83
6	AAAAAT	94	8·40
7	TAAAAA	88	7·86
8	ATATCG	85	7·60
9	ATCGAA	83	7·42
10	AAAATA	82	7·33
11	AAAATT	80	7·15
12	ATTTTA	74	6·61
13	AAATAT	70	6·26
14	ATAAAA	69	6·17
15	AAAAAA	68	6·08
16	ATAAAT	68	6·08
17	AAAACA	68	6·08
18	CCGATA	68	6·08
19	GATAAC	66	5·90
20	CATCGA	65	5·81

[a] Motifs in this range may form composite binding sites with TCGATA/TATCGA, which was seen a total of 1119 times.

regression trees is as a predictive tool for estimating the values at the leaves of the tree. We instead use the predicted values simply as a test for the direction and amount of change in expression.

As inputs into the regression tree software we used the single motifs identified by simple regression, supplemented by those that occur at elevated frequency within 20 bp of TCGATA. Various combinations of these motifs and corresponding data from the nicotine dataset were used in the creation of four multiple regression model trees using Weka software (Witten & Frank, 1999; see Section 2 for details). The resulting trees were compared via their correlation coefficients, which measure the statistical correlation between the actual and predicted expression level values. These values are shown in Table 3. Models 3 and 4 show the highest correlation coefficients and were rerun with the aging dataset used as a test dataset. The test dataset correlation coefficients were 0·49 for Model 3 and 0·48 for Model 4. These values are higher than those obtained for the training dataset, and thus show strong support for the model.

Models 3 and 4 resulted in very similar regression trees and are shown in Fig. 2. Model 4 had one additional node (GATAAC), a motif found within 20 bp of TCGATA but not found to be significant by simple linear regression. We decided not to use Model 4 as our final regression tree model for two reasons: (i) the motif GATAAC was added because of its proximity to TCGATA in upstream sequences but the node containing the motif was not closely connected to TCGATA in the tree and (ii) GATAAC fell out of the model when we removed AGTGTG from the input dataset because of its 5 bp overlap with ACACTG. Since AGTGTG fits in the overlap with other genes in its path in the tree, we decided to keep that motif in the model and use the resulting tree from the set of significant motifs from single-motif regression.

Traversal of the regression tree should identify binding site combinations that may enhance or repress expression significantly in one sex or the other. On the left side of the Model 3 regression tree, we see that with 0 or 1 copy of TCGATA and 0 copies of GGTCAC, we have an estimated expression difference of $-0{\cdot}346$, indicating that genes lacking these motifs in their upstream regions are more likely differentially expressed in males. We then use $-0{\cdot}346$ as a comparison point. If we have 0 or 1 TCGATA, 1 GGTCAC and 0 copies of AGTGTG, the estimated expression difference is $-0{\cdot}320$ which is not much different from $-0{\cdot}346$. This indicates that the addition of a GGTCAC by itself does not change expression. However, if we find the combination of 0 or 1 TCGATA, 1 or more GGTCACs, 1 or more AGTGTGs and 0 ACACTGs, the expression difference changes to $-0{\cdot}176$, which is a considerable change. This motif combination may cause the gene to be less differentially expressed between the sexes. With the same combination of TCGATA, GGTCAC, AGTGTG, but addition of 1 or more copies of ACACTG, the expression difference becomes positive. This can mean either that ACACTG activates female-specific transcription, or that this motif could be a repressor-binding site for male-specific transcription. Since our analysis has not included genes that are not differentially expressed between the sexes, a change of this magnitude in comparison with our other expression differences most likely indicates up-regulation in females.

From the left traversal of the tree, a motif combination of interest is GGTCAC, AGTGTG and ACACTG. This motif trio combines to form the 10-mer GGTCACACTG that contains the palindromic sub-motif GTCACACTG. Of the 238 GGTCAC–ACACTG pairs found in the upstream regions of sex-biased genes, 84 (or 35%) were found in this overlap. Another chi-square test of motif presence associated with female-biased, male-biased or non-sex-biased genes resulted in strong evidence (P value $<0{\cdot}001$) that this larger motif is associated with female-specific expression. Detection of a larger, overlapping binding site such as this is a direct observation from regression trees. A single multiple-regression model does not provide any type of grouping of motifs that may work together. Regression trees separate independent variables that, together, change the dependent variable and create multiple groupings to explain the data.

Table 3. *Regression model tree results*

Model	Motifs in model	No. of leaf nodes in resulting tree[a]	Training set correlation coefficient
1	Most significant from SLR and seen >4% of time within 20 bp	3	0·3107
2	Seen >5% of time within 20 bp	4	0·2904
3	20 most significant from SLR	6	0·3469
4	20 most significant from SLR plus 20 seen most within 20 bp	7	0·3613

[a] The number of leaf nodes in the resulting tree gives an indication of tree complexity.

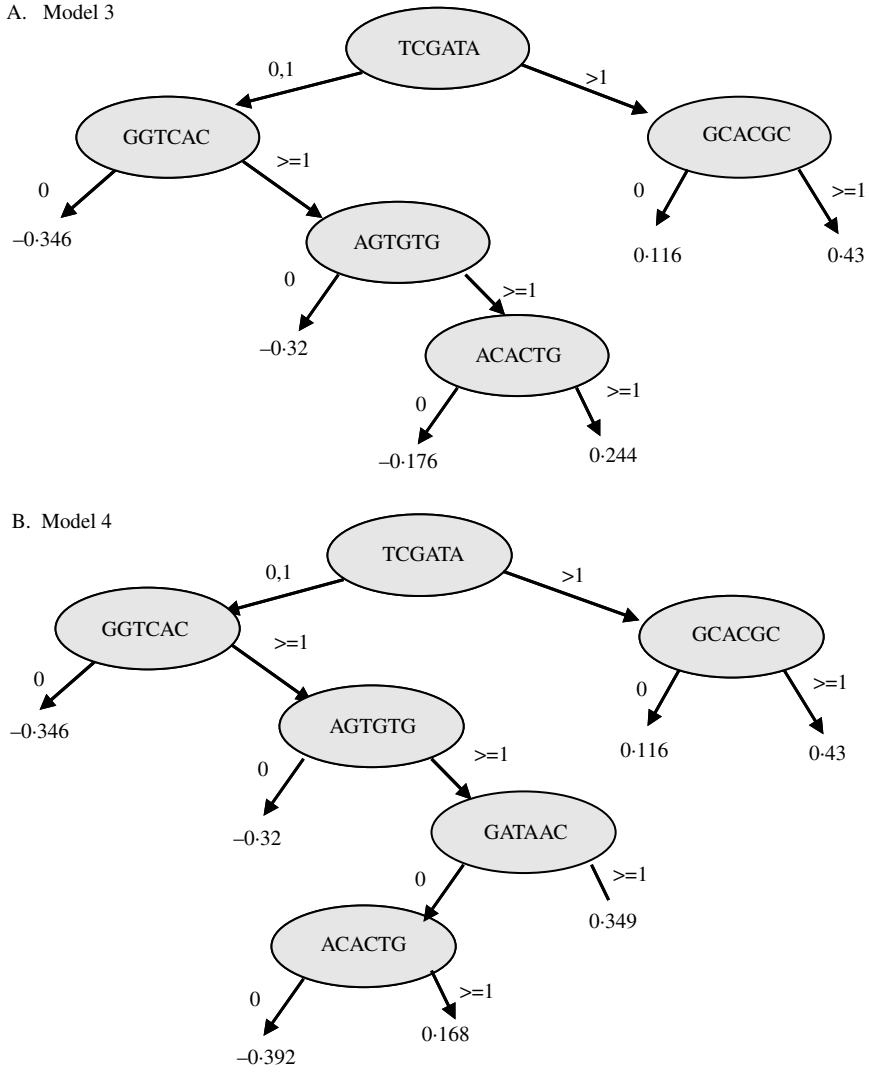


Fig. 2. Regression trees highlighting combinations of motifs that predict sex-biased gene expression in *D. melanogaster*. See text for details of Models 3 and 4.

This is a distinct advantage over multiple-regression methods.

As further verification of our method, we obtained data from a microarray experiment specifically targeting *Drosophila* ovaries and testes, since these reproductive tissues are known to contribute to much of the overall expression difference between adult male and female flies (Parisi *et al*., 2003), and ran it through our analysis. We used genes that showed a four-fold or higher difference in expression between

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

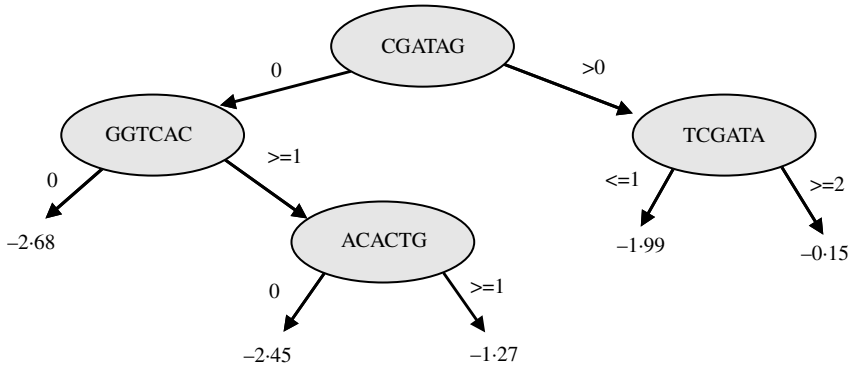


Fig. 3. Regression tree highlighting combinations of motifs that predict sex-biased gene expression in *D. melanogaster* from the ovaries/testes cDNA microarray dataset. See text for details.

the sexes in order to reduce the dataset to approximately 1600 genes. The most significant motif resulting from the single-motif linear regression was TCGATA/TATCGA, and the six most significant motifs from the ovaries/testes dataset were found in the seven most significant motifs resulting from regression on the nicotine dataset. Again, using the motifs with significance below the Bonferroni-corrected cut-off, we built a regression tree. The resulting tree (Fig. 3) was strikingly similar in structure to the regression tree built from the nicotine dataset. The top node in the ovaries/testes regression tree is the motif CGATAG, which has a 5 bp overlap with TCGATA, and TCGATA is the next node in the tree on the female-biased side. This further supports our theory of overlapping TCGATA motifs enhancing female expression. Additionally, expression becomes more female from left to right among the leaves. This tree further validates our regression tree model obtained from the nicotine dataset. The differences relative to the adult fly trees could either be due to sampling variance, or reflect the additional contribution of somatic tissues to sex-specific gene expression in whole flies.

(iii) *Use of decision trees to predict sex-specific gene expression*

With the identification of motifs affecting sex-specific expression by the regression tree, we wanted to determine whether we could use these same motifs to classify a gene as being differentially expressed in either sex from the motifs found in its upstream region. To do this, we created a decision tree, which, based on motif counts, classified a gene as significantly expressed more in males, females or neither. The structure of a decision tree is very similar to that of the regression tree except that the classification of 'male', 'female' or 'neither' is found at the leaves of the tree instead of a predicted expression difference. Again, various combinations of the significant motifs from the single-motif regression model were used as input. Data from all differentially expressed genes and a subset of genes not differentially expressed in males or females from the nicotine cDNA microarray experiment were used to construct the model tree, again using Weka software (Witten & Frank, 1999). Since the motifs used as input into the decision tree model were determined from analysis of differentially expressed genes between males and females, the expectation for the decision tree correctly classifying the differentially expressed genes from the non-differentially expressed genes was low.

Inputting only the motifs found at the regression tree nodes into the decision tree resulted in a model much more complicated than expected (49 nodes in the tree) but with a correct classification percentage of 47%. After realizing that most motifs occur closer to the promoter, we decided to narrow the upstream region of each gene to 700 bp and construct a tree using motif counts from that smaller region. The resulting tree was similar to our regression tree and highlighted certain motif pairs. It is shown in Fig. 4. This tree also had a correct classification percentage of 47% for our training set. On the entire nicotine array gene set, 67% of the observed male-biased genes and 54% of the observed female-biased genes were correctly predicted. Classification of genes not showing sex bias was low, as expected. To test our decision tree results, we created 1000 decision trees with 20 random motifs selected as input. Our model, with a 47% overall correct classification, ranked within the top 1% of all random trees created.

4. Discussion

(i) *Regression trees and sex-specific motifs in* Drosophila

Regression provides a quantitative method of combining sequence data and expression data. We describe here a two-step method for creating a multifactorial model which links the prevalence of binding motifs to changes in expression. Besides eliminating the need

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

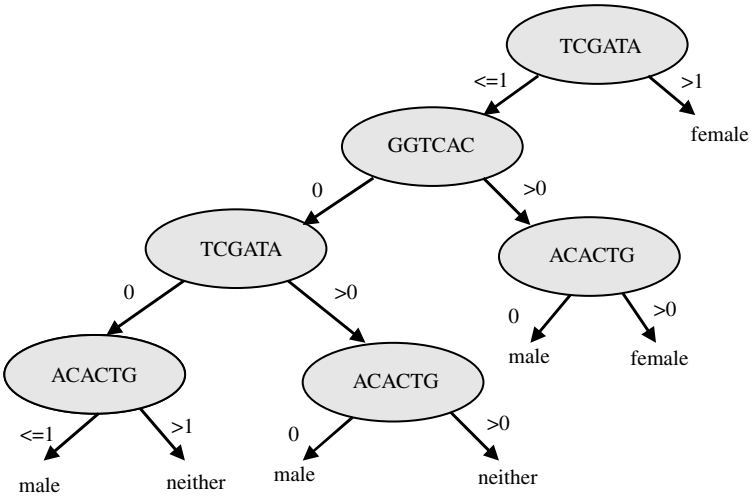


Fig. 4. Decision tree highlighting combinations of motifs that predict sex-biased gene expression in *D. melanogaster*. See text for details.

for clustering of expression data, this technique implies that the presence of multiple motifs in an upstream region is more likely to affect the level of transcription. This concept is starting to be explored in motif-clustering methods (Berman *et al*., 2002; Halfon *et al*., 2002; Markstein *et al*., 2002; Rebeiz *et al*., 2002). However, these motif-clustering methods require prior knowledge of the sequence of the binding sites which are believed to affect expression, and our approach does not. Furthermore, our method provides a straightforward procedure for focusing further analysis on a subset of the numerous significant motifs that may arise using simple linear regression.

Few binding sites for sex-specific expression have been identified in *Drosophila*. Almost all the motifs identified by our single-motif regression were associated with female-biased gene expression. Therefore, the motifs selected by the regression tree model were mostly female-specific. Verification of the function of the three major motifs that are highlighted in the regression trees was achieved by scanning TRANSFAC and the literature, which revealed that each of these motifs has previously been shown to form part of known binding sites for transcription factors during oogenesis. Most interesting is the TCGATA/TATCGA motif that forms the core of the SERPENT binding site, GCTATCGATAGC, in the promoters of the *yp1* and *yp2* genes (Lossky & Wensink, 1995). Similarly, GGTCAC/GTGACC is part of the extended TAGTGTATATAGGTCACGT binding site for chorion factor II in the chorion protein *s15* promoter during oogenesis (Shea *et al*., 1990), and ACACTG/CAGTGT is the core of the CCTACACTGTAAG binding site for DEP3 in the ovarian promoter of *Alcohol dehydrogenase* (Bayer *et al*., 1992).

Very few male-specific motifs were found by any of our single-motif models, and between the datasets, the male-specific motifs that tested with higher significance were different. Although it was surprising that our regression tree did not find any male-specific or antagonistic binding site combinations, it was encouraging that known female-specific motifs were selected and used as decision nodes in the regression tree. Since we only looked at mature adults, our motifs are actually associated with germline (ovary- and testis-specific) expression. Notwithstanding the empirical evidence discussed above that GGTCAC and ACACTG are part of female-specific enhancers, a possibility suggested by the regression trees is that the presence of these motifs is sufficient to contribute to repression of male-specific transcription. It is known, for example, that repressor binding sites in mRNA actively inhibit translation in the male germline (Crowley & Hazelrigg, 1995; Blumer *et al*., 2002). Extra power can be obtained by fitting regressions over a developmental time course, and this had led to the detection of male-specific elements as well (K. P. White & H. J. Bussemaker, personal communication).

A multiple regression model including all the significant motifs was also built on the same sets of motifs as the regression trees and resulted in a model with a correlation coefficient of 0·44. Even though this was similar to the correlation coefficient for our regression tree, the associated model does not uncover all the salient features revealed by our regression tree approach. The TCGATA motif stands out the most from all our analyses as it was always at the root of both the regression and decision trees, indicating that it is the most highly correlated motif in sex-biased expression. Additionally, the TCGATA motif was found overlapping with itself in an 8 bp palindrome 29% of the time, and this overlapping motif tested

positively for association with sex-biased expression. Because TCGATA is found in these situations so often, the motif seems to be involved somehow in regulation and deserves further investigation. Overlap of GGTCAC, AGTGTG and ACACTG into a larger motif is also highly suggested by our results.

(ii) *Advantages and drawbacks of regression trees*

There are at least two situations in which regression trees are expected to outperform direct multiple regression. As documented above, one is where the short motifs overlap and combine to perform a single binding site. Multiple linear regression does not suggest any grouping of motifs, but merely gives partial regression coefficients indicating the contribution of the motif to the change in expression. In fact, overlapping motifs will tend not to add significance to the overall model fit once the most strongly associated motif has been accounted for. The second situation where regression trees should provide an advantage is where multiple different combinations of motifs give rise to similar expression patterns. Though not strongly indicated here, most likely because only a short section of each promoter was examined, in theory combinations of motifs that act together should generate their own arms of the regression tree. It should even be possible for the same motif to appear on different arms at different frequencies, as for example TCGATAT in our decision tree, and for repressor and activator functions to be distinguished.

The utility of regression trees is thus more likely to lie in the perspective they provide concerning the relationship among motifs, rather than superior performance in identifying single motifs. The major factors restricting the application of regression trees relate to the enormous range of possible ways of combining and formulating motifs. While 8-mer and longer motifs may often be functional, perfect matches will often be rare in promoters of co-regulated genes so statistical power is reduced, particularly given that the increased number of possible longer motifs requires more stringent significance thresholds. Similarly, formulation of trees that combine motifs of different lengths, or link motifs in two different regions of a gene (for example, putative promoter and distal enhancer elements), creates so many possible combinations that it will be difficult to assess *a priori* which trees are more or less probable. If the number of co-regulated genes for which a regulatory motif is sought is less than 20 or so, it may never be possible to use regression-based approaches since P values of the order of 10^{-5} would require an unreasonably tight relationship between motif count and transcript abundance. Nevertheless, systematic simulation studies and statistical modelling, including use of other evidence to define candidate regulatory regions within which motifs may lie (Wasserman *et al*., 2000), should improve the performance of regression trees in the context of regulatory motif detection.

(iii) *Do computational approaches identify enhancer elements?*

The standard approach to confirmation that a motif actually regulates gene expression is to demonstrate that it is sufficient to drive expression of a reporter gene in the predicted pattern in a transgenic organism. In our case, the expression data themselves demonstrate, however, that the identified motifs are insufficient to drive female-specific expression, since a large number of genes with each motif combination are expressed more strongly in males than females. Several other recent studies have failed to confirm that sequences identified using bioinformatic approaches are functional. For example, Halfon *et al*. (2002) extracted 34 potential dorsal mesodermal enhancers consisting of multiple binding sites for known transcription factors, but only 8 of the 18 of these for which data are available appear to drive transcription in embryonic *Drosophila* mesoderm. They concluded that there can be a high false-positive identification rate associated with computational strategies.

Given the extremely high significance associated with particular test statistics, it should also be considered that some potential regulatory motifs are not classical enhancers, but rather define a class of 'modulator' elements that act in a more probabilistic manner. Either the effects of individual elements are too subtle to detect in transgenic assays, or the elements act in a context-dependent manner. Promoter-proximal elements such as those characterized in this study are likely to require distal true enhancer sequences, as regulatory regions in flies typically extend over tens of kilobases. The corollary may also be true, that enhancers require the context of modulator elements, such as those identified here, more commonly than generally recognized.

The problem remains as to how to confirm the biological function of statistically significant motifs. One approach is to ask whether the motifs are polymorphic in the promoters of genes that show variable expression within and among species. We sequenced the promoters of 10 wild-type strains of *D. melanogaster* for eight genes that differed between genotypes in the level of sex-specific transcription in our microarray studies. Nine of the 72 SNPs and indel polymorphisms were located within the top 10 motifs described here, but this fraction is not greater than expected given the motif frequencies in the sequenced regions. Nevertheless, polymorphism in modulator elements is an intuitively appealing mechanism for quantitative variation in gene expression that could contribute to gradual evolution of gene expression.

Phylogenetic shadowing (Boffelli *et al*., 2003; Kellis *et al*., 2003), namely extensive genomic comparison of promoter sequences in multiple sibling species among which tissue-specific gene expression diverges, is likely to aid in the functional footprinting of subtle regulatory motifs.

We thank John Doyle for encouraging us in the application of regression and decision trees to motif detection, Brian Oliver for discussions concerning sex-biased gene expression, and Kevin White for communicating unpublished data. Rebecca Riley-Berger and Jennifer King sequenced the promoters of the eight genes. E.H. is the recipient of an IGERT training fellowship in Genome Sciences, and microarray research in G.G.'s laboratory has been supported by the David and Lucille Packard Foundation and NIH award P01 GM45344.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, P. G., *et al*. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W. & White, K. P. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**, 2270–2275.

Arnosti, D. N. (2003). Analysis and function of transcriptional regulatory elements: insights from *Drosophila*. *Annual Reviews in Entomology* **48**, 579–602.

Bayer, C. A., Curtiss, S. W., Weaver, J. A. & Sullivan, D. T. (1992). Delineation of *cis*-acting sequences required for expression of *Drosophila mojavensis Adh-1*. *Genetics* **131**, 143–153.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M. & Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proceedings of the National Academy of Sciences of the USA* **99**, 757–762.

Blumer, N., Schreiter, K., Hempel, L., Santel, A., Hollmann, M., Schafer, M. A. & Renkawitz-Pohl, R. (2002). A new translational repression element and unusual transcriptional control regulate expression of don juan during *Drosophila* spermatogenesis. *Mechanisms of Development* **110**, 97–112.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K., Ovcharenko, I., Pachter, L. & Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. New York: Chapman & Hall/CRC.

Bussemaker, H. J., Li, H. & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics* **27**, 167–171.

Crowley, T. E. & Hazelrigg, T. (1995). A male-specific 3′-UTR regulates the steady-state level of the exuperantia mRNA during spermatogenesis in *Drosophila*. *Molecular and General Genetics* **248**, 370–374.

Drouin, J., Sun, Y. L., Tremblay, S., Lavender, P., Schmidt, T. J., de Lean, A. & Nemer, M. (1992). Homodimer formation is rate-limiting for high affinity DNA binding by glucocorticoid receptor. *Molecular Endocrinology* **6**, 1299–1309.

Guha Thakurta, D., Palomar, L., Stormo, G. D., Tedesco, P., Johnson, T. E., Walker, D. W., Lithgow, G., Kim, S. & Link, C. D. (2002). Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Research* **12**, 701–712.

Halfon, M. S., Grad, Y., Church, G. M. & Michelson, A. M. (2002). Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated computational model. *Genome Research* **12**, 1019–1028.

Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000). Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296**, 1205–1214.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G. & Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* **29**, 389–395.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254.

Klingenhoff, A., Kornelie, F., Quandt, K. & Werner, T. (1999). Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15**, 180–186.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.

Liu, X., Brutlag, D. L. & Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, 127–138.

Lossky, M. & Wensink, P. C. (1995). Regulation of *Drosophila* yolk protein genes by an ovary-specific GATA factor. *Molecular and Cellular Biology* **15**, 6943–6952.

Manson-McGuire, A., Hughes, J. D. & Church, G. M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Research* **10**, 744–757.

Markstein, M., Markstein, P., Markstein, V. & Levine, M. S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences of the USA* **99**, 763–768.

McCue, L. A., Thompson, W., Carmack, C. S. & Lawrence, C. E. (2002). Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Research* **12**, 1523–1532.

Parisi, M., Nuttall, R., Naiman, D., Bouffard, G., Malley, J., Andrews, J., Eastman, S. & Oliver, B. (2003). Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science* **299**, 697–700.

Piano, F., Parisi, M. J., Karess, R. & Kambysellis, M. P. (1999). Evidence for redundancy but not *trans* factor–*cis* element coevolution in the regulation of *Drosophila Yp* genes. *Genetics* **152**, 605–616.

Pilote, L., Miller, D. P., Califf, R. M., Rao, J. S., Weaver, W. D. & Topol, E. J. (1996). Determinants of the use of coronary angiography and revascularization after thrombolysis for acute myocardial infarction. *New England Journal of Medicine* **335**, 1198–1205.

Pilpel, Y., Sudarsanam, P. & Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **29**, 153–159.

Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research* **23**, 4878–4884.

Quinlan, J. R. (1992). Learning with continuous classes. *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, World Scientific, pp. 343–348.

Rajewsky, N., Socci, N. D., Zapotocky, M. & Siggia, E. D. (2002). The evolution of DNA regulatory regions for Proteo-gamma bacteria by interspecies comparisons. *Genome Research* **12**, 298–308.

Rebeiz, M., Reeves, N. L. & Posakony, J. W. (2002). SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. *Proceedings of the National Academy of Sciences of the USA* **99**, 9888–9893.

Rejwan, C., Collins, N. C., Brunner, L. J., Shuter, B. J. & Ridgway, M. S. (1999). Tree regression analysis on the nesting habitat of smallmouth bass. *Ecology* **80**, 341–348.

Shea, M. J., King, D. L., Conboy, M. J., Mariani, B. D. & Kafatos, F. C. (1990). Proteins that bind to *Drosophila* chorion *cis*-regulatory elements: a new C2H2 zinc finger protein and a C2C2 steroid receptor-like component. *Genes and Development* **4**, 1128–1140.

Wang, Y. & Witten, I. H. (1997) Inducing model trees for continuous classes. *Proceedings of the European Conference on Machine Learning*.

Wasserman, W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. (2000). Human–mouse genome comparisons to locate regulatory sites. *Nature Genetics* **26**, 225–228.

Werner, T. (2002). Finding and decrypting of promoters contributes to the elucidation of gene function. *In Silico Biology* **2**, 249–255.

Witten, I. H. & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Diego, CA: Morgan Kaufmann.

Wolfinger, R. D., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.

Genet. Res., Camb. (2004), **83**, *pp.* 189–196. © 2004 Cambridge University Press
DOI: 10.1017/S0016672304006792 Printed in the United Kingdom

Insecticide resistance genes confer a predation cost on mosquitoes, *Culex pipiens*

C. BERTICAT, O. DURON, D. HEYSE AND M. RAYMOND*
Institut des Sciences de l'Evolution (*UMR CNRS 5554*), *C.C. 065, Université de Montpellier II, 34095 Montpellier cedex 05, France*

(*Received 5 January 2004 and in revised form 19 February 2004*)

Summary

Newly occurring adaptive genes, such as those providing insecticide resistance, display a fitness cost which is poorly understood. In order to detect subtle behavioural changes induced by the presence of resistance genes, we used natural predators and compared their differential predation on susceptible and resistant *Culex pipiens* mosquitoes, using strains with a similar genetic background. Resistance genes were either coding an overproduced detoxifying esterase (locus *Ester*), or an insensitive target (locus *ace-1*). Differential predation was measured between susceptible and resistant individuals, as well as among resistant mosquitoes. A backswimmer, a water measurer, a water boatman and a predaceous diving beetle were used as larval predators, and a pholcid spider as adult predator. Overall, the presence of a resistance gene increased the probability of predation: all resistance genes displayed predation costs relative to susceptible ones, at either the larval or adult stage, or both. Interestingly, predation preferences among the susceptible and the resistance genes were not ranked uniformly. Possible explanations for these results are given, and we suggest that predators, which are designed by natural selection to detect specific behavioural phenotypes, are useful tools to explore non-obvious differences between two classes of individuals, for example when they differ by the presence or absence of one recent gene, such as insecticide resistance genes.

1. Introduction

Genes responsible for an adaptation to a new environment are usually assumed to have a fitness cost, i.e. to be at a disadvantage in the previous environment (e.g. Fisher, 1958; Lande, 1983; Orr & Coyne, 1992; Carrière *et al*., 1994). This assumption is based on the general view that resource reallocation occurs or that metabolic or developmental processes are affected, thus decreasing other fitness-enhancing characters (Davies *et al*., 1996). Cost can be important in the evolution of adaptation since it can lead to allelic replacement (an allele is replaced by a less costly one) or to selection of modifier genes (Lenski, 1988*a*, *b*; Cohan *et al*., 1994). Few situations exist where both the environmental changes and the adaptive genes are clearly identified. Resistance to pesticides, and in particular resistance to organophosphorus insecticides (OP) in *Culex pipiens* L. mosquitoes, is one of them.

* Corresponding author. Fax: +33 4 67144615. e-mail: raymond@isem.univ-montp2.fr

Two loci are involved in OP resistance in *C. pipiens*, the super-locus *Ester* and the locus *ace-1*. Several resistance alleles have been described at both loci (for a review see Raymond *et al*., 2001). The resistance conferred by *Ester* is due to an esterase over-production which is the result of two non-exclusive mechanisms (Raymond *et al*., 1998): gene amplification (for instance, $Ester^4$, $Ester^2$ and $Ester^5$ alleles), or change in gene regulation ($Ester^1$ allele). The *ace-1* locus codes for the OP target, acetylcholinesterase (AChE). Resistance alleles $ace\text{-}1^R$ code an AChE with a reduced sensitivity towards OP, associated with modified catalytic properties (Bourguet *et al*., 1997).

Resistance genes have been studied in the Montpellier area for more than 30 years. Resistance first appeared in 1972 with the occurrence of $Ester^1$, followed by $ace\text{-}1^R$ in 1978, $Ester^4$ in 1984 and $Ester^2$ in 1990 (Guillemaud *et al*., 1998). Estimations of overall fitness costs from population surveys have shown that *ace-1* is associated with higher deleterious effects than *Ester* (Lenormand *et al*., 1999; Lenormand & Raymond, 2000). This difference is also observed for a

specific life history trait, survival during the overwintering period (Chevillon *et al*., 1997; Gazave *et al*., 2001). The functional differences between the two loci could explain this phenomenon (Chevillon *et al*., 1997). The over-production of esterases by the *Ester* locus could be at the expense of producing something else, with the resulting alteration of some fitness-related traits. The modified AChE could lead to changes in some behavioural fitness-related traits, since it alters the optimal functioning of cholinergic synapses of the central nervous system. It has been observed that, during the 1990s, $Ester^4$ has replaced $Ester^1$ (Guillemaud *et al*., 1998). As $Ester^4$ is known to confer a slightly lower OP resistance level, its advantage over $Ester^1$ could possibly come from a lower cost (Guillemaud *et al*., 1998). The proximal causes of such variability in the fitness cost between resistance alleles are still unknown.

In order to better understand this fitness cost and its variability, the effects of these resistance genes on several fitness-related traits are being studied, using strains sharing the same genetic background. In a recent study, a mating competition cost associated with $Ester^1$, $Ester^4$ and $ace\text{-}1^R$ resistance alleles was demonstrated, but no cost difference between them was detected (Berticat *et al*., 2002*a*). Here, we investigate how these three resistance alleles affect the probability of predation at larval and adult stages, relative to susceptible alleles. We also attempt to compare the resistance alleles with one another. Avoiding predation is an important fitness component of *C. pipiens* (Sih, 1986), and confrontation with a predator could constitute a risky situation, liable to amplify the physiological differences between the resistance genotypes, thus potentially allowing us to detect cost difference between the resistance alleles.

2. Materials and methods

(i) *Mosquito strains*

Four strains sharing the same genetic background and only differing by their genotype at *Ester* and/or *ace-1* locus were used: the insecticide-susceptible strain S-LAB, homozygous for $ace\text{-}1^S$ and $Ester^0$ (Georghiou *et al*., 1966); the resistant strains SA1 and SA4, homozygous for $ace\text{-}1^S$ and for the resistance alleles $Ester^1$ and $Ester^4$, respectively; and finally, the resistant strain SR, homozygous for $Ester^0$ and for the resistance allele $ace\text{-}1^R$ (Berticat *et al*., 2002*a*). Before all experiments, all strains were reared under the same standardized conditions for a minimum of 5 generations, preventing possible maternal effects.

(ii) *Predation on adult mosquitoes*

The adult predator used in this experiment was a spider, *Holocnemus pluchei* (Scopoli) (Araneae, Pholcidae), a common inhabitant of homes, which is known to feed on flying insects, including *C. pipiens* (Déom, 1990). *H. pluchei*, through vibrations of its web, locates its prey, which is eventually immobilized and rapidly packed with silk threads. Then *H. pluchei* injects its digestive saliva into a captured insect, and ingests the content. The external skeleton of an empty individual remains, tightly packed like a mummy, allowing easy detection of eaten adults. *H. pluchei* used here were locally collected in one University building.

Differential predation between two strains was assessed by introducing, into the same cage ($20 \times 20 \times 20\ cm^3$), 20 one-day-old male mosquitoes from each of the two strains considered, together with one *H. pluchei*. Predators were starved for 10 days before each experiment. Every day, predated adults ('mummies') were collected, and the spider was replaced by a new starved one. This procedure ensured that the predation rate did not decrease due to satiation. The experiment was ended when approximately 50% of all adults were eaten. In order to recognize the strain of origin of each mummy, adults of each strain were marked just before the start of an experiment, using fluorescent powders of different colour (yellow or orange). For each experiment, at least two replicates were performed by switching the colour of each strain. Additionally, experiments with adults marked with orange or yellow from the same strain were conducted for all strains. The different experiments performed and their number of replicates are indicated in Table 1.

(iii) *Predation on mosquito larvae*

The larval predator used in this experiment was the pigmy backswimmer, *Plea minutissima* Leach (Hemiptera, Pleidae), which is about 2 mm in size. This insect is a common inhabitant of ponds of the Palearctic, and feeds on small aquatic prey such as other small insects or crustaceans. *P. minutissima* is a potential predator of *C. pipiens*, as both often co-occur in the same breeding sites (Laird, 1988), and *P. minutissima* readily feeds on young (L1 or L2) *C. pipiens* larvae in the laboratory. *P. minutissima* injects its digestive saliva into a captured larvae, and ingests the contents. The external skeleton of an empty larva remains, allowing easy detection of captured larvae. *P. minutissima* used here were collected locally (around the Montpellier area) and reared in the laboratory.

Differential predation between two strains was assessed by introducing, into the same container, an equal number of L2 larvae from the two strains considered, together with two or three *P. minutissima*. The experiment was ended when approximately 50% of all larvae had been preyed upon, and eaten larvae of each strain were recorded. Predators were starved for 10 days before each experiment. In order to

Table 1. *Adult predation. (A) Effect of powder coloration on each strain, (B) effect of resistance genes compared with a susceptible one, and (C) effect of different resistance genes between them*

Effect tested	Confronted strains		No. of replicates	*P* values	$\hat{\beta}$ of the strain mentioned
	Orange	Yellow			
(A) Effect of coloration	S-LAB	S-LAB	4	0·1345	–
	SA1	SA1	3	1	–
	SA4	SA4	3	0·74	–
	SR	SR	2	0·88	–
	All		–	0·69	–
(B) Effect of resistance vs susceptible genes	SA1	S-LAB	2	0·018	–
	S-LAB	SA1	2	0·03	–
	All		–	0·001	SA1 **0·67** (0·048)
	SA4	S-LAB	2	0·48	–
	S-LAB	SA4	2	0·001	–
	All		–	0·02	SA4 **0·64** (0·076)
	SR	S-LAB	2	0·25	–
	S-LAB	SR	2	0·89	–
	All		–	0·59	SR 0·50 (0·075)
(C) Effect of different resistance genes	SA1	SA4	2	0·56	–
	SA4	SA1	2	0·22	–
	All		–	0·36	SA4 0·41 (0·044)
	SR	SA1	2	0·93	–
	SA1	SR	2	0·13	–
	All		–	0·35	SR 0·44 (0·060)
	SR	SA4	2	0·32	–
	SA4	SR	2	0·76	–
	All		–	0·57	SR 0·57 (0·033)

The *P* value refers to a two-sided (A and C) or a one-sided test (B), when the alternative hypothesis is a higher predation rate for resistant mosquitoes. For all cases, the *P* value refers to a global exact test across replicates. Estimates of average predation coefficients ($\hat{\beta}$) refer to the strain mentioned and bold characters indicate $\hat{\beta}$ values significantly ($P<0·05$) higher than 0·5. SE is given in parentheses. See text for explanations.

recognize the strain of origin of each larva, two protocols were used. For the first protocol, each experiment was conducted in 100 ml of tap water (water depth 1·5 cm), with a total number of 40 larvae. No refugium was available for the mosquito larvae. Larvae of one of the strains considered were stained just before the start of an experiment, using diluted methylene blue. For each experiment, two replicates were performed by switching the stained strain. Additionally, experiments with stained and unstained larvae from the same strain were conducted for all the strains. The number of replicates of the different experiments are indicated in Table 2. For the second protocol, when larvae from the SR strain were involved, a propoxur (a carbamate insecticide) concentration of 5 mg/l was applied during 24 h to the non-eaten larvae. In this case, each experiment was conducted in 500 ml of tap water (water depth 1 cm), with a total number of 200 larvae and no refugium was available for the mosquito larvae. This dose kills in a few hours only those larvae without the *ace-1*^{R} resistance gene (i.e. all individuals except those from the SR strain), as the propoxur concentration required to kill SR larvae after 24 h exposure is more than 100-fold higher (Bourguet *et al.*, 1997). This procedure allowed the identification of SR individuals among non-eaten larvae. As a control, the same propoxur dose was simultaneously applied only to susceptible *ace-1*^{S} (S-LAB, or SA1 or SA4) and only to *ace-1*^{R} resistant (SR) larvae. The number of replicates of the different experiments is indicated in Table 3. The same procedure could not be used for the other resistant strains, as their relatively low OP resistance level does not allow the use of a discriminative dose.

The larval predation cost of SR relative to S-LAB was further evaluated using three additional predators: a water boatman *Sigara lateralis* (Leach) (Hemiptera, Corixidae), a predaceous diving beetle *Guignotus pusillus* Fabricius, 1781 (Coleoptera, Dytiscidae) and the water measurer *Hydrometra stagnorum* (Linnaeus, 1758) (Hemiptera, Hydrometridae). Their size is approximately 5–6, 2 and 10 mm, respectively. All these predators are commonly found in mosquito breeding sites around the Montpellier area, and also at a larger scale (Laird, 1988). They can feed only on young (L1 or L2) *C. pipiens* larvae in laboratory conditions, and inject their digestive saliva into a captured larva in order to ingest its content. Water boatmen appear to be very effective predators, and seem to hunt like *P. minutissima*. In comparison with

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

Table 2. *Larval predation by* Plea minutissima. (*A*) *Effect of dye on each strain, and* (*B*) *effect of resistance genes compared with a susceptible one*

Tested effect	Confronted strains		No. of replicates	*P* value	$\hat{\beta}$
	Not stained	Stained			
(A) Effect of coloration	S-LAB	S-LAB	16	$<10^{-5}$	–
	SA1	SA1	4	0·14	–
	SA4	SA4	5	0·24	–
	SR	SR	3	1	–
	All		–	$<10^{-4}$	–
(B) Effect of resistance vs susceptible genes	SA1	S-LAB	8	0·68	0·49 (0·038)
	SA4	S-LAB	12	$<10^{-5}$	**0·63** (0·050)
	SR	S-LAB	12	$<10^{-8}$	**0·71** (0·050)

The *P* value refers to a two-sided (A) or a one-sided test (B), when the alternative hypothesis is a higher predation rate for resistant mosquitoes. For all cases, the *P* value refers to a global exact test across replicates. Estimates of average predation coefficients ($\hat{\beta}$) refer to the resistant strain and bold characters indicate $\hat{\beta}$ values significantly ($P<0·05$) higher than 0·5. SE is given in parentheses. See text for explanations.

Table 3. *Larval predation by* Plea minutissima*, using an insecticide for genotype identification*

Confronted strains	No. of replicates	*P* value	$\hat{\beta}$
SR S-LAB	5	$<10^{-8}$	**0·65** (0·014)
SR SA1	5	0·22	0·52 (0·024)
SR SA4	5	$<10^{-8}$	**0·86** (0·036)

The *P* value refers to a two-sided (lines 2 and 3) or a one-sided test (line 1), when the alternative hypothesis is a higher predation rate for SR mosquitoes. For all cases, the *P* value refers to a global exact test across replicates. Estimates of average predation coefficients ($\hat{\beta}$) refer to the SR strain and bold characters indicate $\hat{\beta}$ values significantly ($P<0·05$) higher than 0·5. SE is given in parentheses. See text for explanations.

Table 4. *Estimates of average predation coefficients* ($\hat{\beta}$) *for resistant larvae* (*SR strain*) *compared with susceptible ones* (*S-LAB strain*)*, in the presence of various predators* (SE *in parentheses*)

Predator	No. of replicates	*P* value	$\hat{\beta}$
Sigara lateralis	9	0·22	0·56 (0·020)
Guignotus pusillus	9	$<10^{-2}$	**0·69** (0·043)
Hydrometra stagnorum	11	$<10^{-8}$	**0·68** (0·033)

The *P* value refers to a one-sided test, when the alternative hypothesis is a higher predation rate for SR larvae. For all cases, the *P* value refers to a global exact test across replicates. Estimates of average predation coefficients ($\hat{\beta}$) refer to the SR strain and bold characters indicate $\hat{\beta}$ values significantly ($P<0·05$) higher than 0·5. SE is given in parentheses. See text for explanations.

other Dytiscidae, adults of *Guignotus pusillus* are very small, and feed only on tiny prey. The water measurer walks slowly onto the water surface, usually among vegetation, and spears small prey under the water surface with its long rostrum. Differential predation between S-LAB and SR was assessed with the same protocol described above with *P. minutissima*, although only L1 larvae were used, and only one predator per replicate. Experiments were conducted in 250, 50 and 50 ml of tap water, with a total number of larvae of 200, 100 and 40 for the water boatman, water beetle and water measurer, respectively. Non-eaten larvae were assigned to each strain by treating them with a discriminating dose of propoxur (5 mg/l), as described above. The numbers of replicates of the different experiments are indicated in Table 4.

(iv) *Statistics*

A predation experiment corresponds to sampling without replacement. The null hypothesis (H_0) is that both morphs (here strains) are equally preyed upon. At the end of the experiment, the number of eaten individuals of each morph follows a hypergeometric distribution, and the probability of the observed data, under H_0, is: $P_{obs}=(C_{A_1}^{r_1}\,C_{A_2}^{r_2})/C_{A_1+A_2}^{r_1+r_2}$, where Aj denotes the total number of morph j at the beginning of the experiment, rj is the number of morph j remaining after predation, and $C_i^j=i!/j!(i-j)!$. To test H_0, a hypergeometric exact test was constructed. The *P* value is defined as: $P=\sum_{P_i\leq P_{obs}}P_i$, where P_i is the probability (under H_0) of all i cases describing all possible ways of distributing the observed number of eaten individuals among both morphs, with the total number of individuals of both morphs kept constant. When an alternative hypothesis was present (e.g. resistant individuals were more preyed upon than susceptible ones), a one-sided test was performed. When no alternative hypothesis was obvious (e.g. when differently coloured adults of the same strain were together), a two-sided

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

test was done. A quick-basic program was written to perform these tests, and was checked by comparison with hand calculations. A global test across replicates was performed by generating the joint distribution, and computing the P value as $P=\sum_{P_j \leqslant Pg_{obs}} P_j$, where P_j is the probability of element j of the joint distribution, and Pg_{obs} is the joint probability of the observed data. When a specified alternative hypothesis was present (e.g. type 1 individuals were more preyed upon than type 2), the P value was $P=\sum_{N_j \geqslant N_{obs}} P_j$, where N_j is the total number of type 1 preyed upon individuals in element j of the joint distribution, and N_{obs} is the total number of observed type 1 preyed upon individuals across replicates. A quick-basic program was written to perform the global exact test for up to five replicates, using the complete enumeration method. A PowerBasic program was written to perform the global exact test for an unspecified number of replicates, using the resampling method to estimate the P value. Program checking was done by comparing the P values generated by the two programs (which use very different algorithms) when used on the same data, for 2–5 replicates. When the number of resamplings was 500 000, the estimated P values diverged by less than 0·4% from the computed exact values. The exact P value was computed for cases with 2–4 replicates, and also for 5 replicates when the number of assayed individuals was lower than 40. In all other cases, the exact P value was estimated using 500 000 resamplings.

Preference was measured using the index proposed by Manly (1974, 1985):

$$\hat{\beta}_i = \frac{\log_e(r_i/A_i)}{\sum_{j=1}^{K} \log_e(r_j/A_j)},$$

where K is the number of morphs (here $K=2$). This measure is appropriate for experiments in which the prey are not replaced during the experiment. This index varies between 0 and 1, and $\sum_i \hat{\beta}_i = 1$. The absence of preference between two morphs corresponds here to $\hat{\beta}=1/2$.

3. Results

(i) *Adult predation*

Each predation experiment lasted about 3 days (range 1–4 days). In order to recognize susceptible and resistant mosquitoes in the experimental cage, adults were marked with a fluorescent powder, either yellow or orange. The colour of the powder had no significant effect ($P>0·69$) on the predation frequency, for all the strains used (Table 1). When susceptible and resistant adults were in the same cage, the latter were significantly more preyed upon than the former (SA1: $P<0·001$, $\hat{\beta}=0·67\pm0·048$; SA4: $P=0·02$, $\hat{\beta}=0·64\pm0·076$). However, no difference in predation rate relative to susceptible individuals was apparent for the SR strain (Table 1). When the resistant strains were confronted pairwise within the same cage, predation was not different ($P>0·3$) according to the resistance genes present.

(ii) *Larval predation*

Each predation experiment lasted about 2 days (range 1–3 days). In order to recognize susceptible and resistant mosquitoes in the experimental container, larvae were stained with a blue dye. This dye slightly increased the risk of predation by *P. minitissima* for the susceptible strain (Table 2). As the hypothesis considered is a higher predation for resistant larvae compared with susceptible ones, only assays where the susceptible strain is stained are presented, in order to be conservative (assays where the resistant strain is stained are all supportive of the hypothesis tested, but they are not conclusive due to the dye bias). Despite this disadvantage, stained susceptible larvae were significantly less predated than resistant ones ($P<10^{-5}$), with the exception of SA1 larvae (Table 2).

When SR individuals were used, they could be recognized within the non-eaten larvae as they survive a high concentration of propoxur. Thus no dye was required in these experiments. SR larvae were significantly more preyed upon than susceptible individuals ($P<10^{-8}$, $\hat{\beta}=0·65\pm0·014$). SR larvae were also significantly more eaten than SA4 ($P<10^{-8}$, $\hat{\beta}=0·86\pm0·036$), although no difference ($P=0·22$) in predation rate was apparent when SR and SA1 were together (Table 3).

To evaluate whether the differences detected by *Plea minutissima* were also detected by other larval predators, SR were confronted with S-LAB larvae in the presence of the three other aquatic predators. For these predators, SR larvae were significantly more preyed upon than susceptible ones (diving beetle: $P<10^{-2}$, $\hat{\beta}=0·69\pm0·043$; water measurer: $P<10^{-8}$, $\hat{\beta}=0·68\pm0·033$; Table 4), with the exception of the water boatman ($P=0·22$, $\hat{\beta}=0·56\pm0·020$; Table 4).

4. Discussion

Overall, the presence of a resistance gene increased the probability of predation, at both the larval and the adult stage: there is thus a 'predation cost' associated with these genes.

(i) *Origin of the predation cost*

Hunting techniques of backswimmers and water boatmen (families Notonectidae, Corixidae and Pleidae) rely essentially upon prey motion (Murphey & Mendenhall, 1973; Sih, 1979). Behaviour underlying

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

backswimmers' preferences seems to be stereotyped and inflexible (Scott & Murdoch, 1983). Many mosquito larvae, including those of *C. pipiens*, are natural prey items for several backswimmer species, and thus share an evolutionary history with them (e.g. Sunish & Reuben, 2002; Chesson, 1984; Blaustein, 1998; Mogi *et al*., 1999). It is thus not surprising that upon a backswimmer attack, mosquitoes try most of the time to escape by becoming motionless, although other strategies are also occasionally observed (such as wriggling away) (Scott & Murdoch, 1983; Sih, 1979). *C. pipiens* larvae are apparently able to detect chemicals released by conspecifics which have been preyed upon by backswimmers, and adjust their behaviour to reduce the predation risk by choosing a less risky microhabitat (a vegetation refugium, the edge of the breeding site, etc.) and moving less (Sih, 1986). Similarly, prey motion is reduced following the introduction of a dytiscid (Kruuk & Gilchrist, 1997). This behavioural change is probably an adaptation to escape predators using motion and/or vibration to detect and locate their prey.

The higher predation cost inflicted by three larval predators could be explained if resistant larvae are more active, and thus are detected more frequently by the predator. Another possibility is that resistant larvae are not changing their microhabitat and/or their moving frequency after conspecifics have started to be preyed upon, unlike susceptible individuals. SR larvae display a distinct feeding behaviour, as they replace their gut contents at a faster rate than the other strains (Agnew *et al*., 2004). This is consistent with the former hypothesis (resistant larvae are more active), although a direct measurement is required to confirm this. The absence of predation cost in the presence of the water boatman is surprising, and suggests that its hunting technique is different. The identification of this difference could potentially shed some light on the modified behaviour of resistant larvae.

The pholcid spider's principal means of capturing prey is to throw silk with the aid of its hind legs. This method is used to immobilize mosquitoes which are entangled in the standing web, or to catch flying mosquitoes directly (Strickman *et al*., 1997; Déom, 1990). Once a mosquito has been in contact with the web, it could escape a spider attack. Apparently, mosquitoes possessing $Ester^1$ or $Ester^4$ have a higher predation probability (Table 1), suggesting that they are either more active (thus with a higher probability of flying near the web or the spider), or have fewer chances to escape an attack by *H. pluchei*. However, possessing $ace\text{-}1^R$ does not seem to affect predation probability. There are several physiological differences between susceptible and resistant mosquitoes. For example, susceptible adults live longer (Agnew *et al*., 2004), and have a lower density of endocellular *Wolbachia* (Berticat *et al*., 2002*b*). *Wolbachia* affect locomotive performance, at least in a parasitic wasp (Fleury *et al*., 2000), and thus may represent a causal link between the effect of a resistant gene and the predation cost. Further experiments, using aposymbiotic strains, could settle this issue.

(ii) *Variability of the predation cost*

All the resistance genes studied present a predation cost relative to susceptible ones, at either the larval or adult stage, or both.

For the *ace-1* locus, the predation cost of the resistance alleles seems to be restricted to the larval stage: spiders seem to capture susceptible and resistant adult mosquitoes equally. This indicates that the high survival cost associated with the $ace\text{-}1^R$ gene during the overwintering period (Chevillon *et al*., 1997; Gazave *et al*., 2001), could not be attributed to pholcid predation. However, it is still possible that other spider species use distinct cues or use different catching techniques which are more discriminatory towards the behavioural changes between mosquitoes resistant and susceptible at the *ace-1* locus. It is also possible that the predation cost is only apparent in female mosquitoes (which were not used in the experiments), as only females overwinter in caves. Only empirical data using the most common spider predators in local caves (*Meta bourneti* (Simon, 1922), *Tegenaria parietina* (Fourcroy, 1785), *Pholcus phalangioides* (Fuesslin, 1775)) could settle this point. The first two species have already been observed catching hibernating *C. pipiens* (M. Michaud, personal communication), although no quantitative data are yet available.

As regards the *Ester* locus, the allele $Ester^4$ displays a predation cost in both larvae and adults, although $Ester^1$ induces a cost only in adults. This absence of predation cost in larvae must be considered with caution, as the procedure used was very conservative: it could be safely concluded only that the predation cost of $Ester^1$ in larvae is not significantly higher than that induced by the staining procedure in susceptible individuals.

There is one example of transitivity for predation preferences (e.g. if the preference is ranked as $A<B$ and $B<C$, then $A<C$): adults with $Ester^1$ or $Ester^4$ are equally more preyed upon than susceptible mosquitoes ($\hat{\beta}=0{\cdot}67$ and $0{\cdot}64$, respectively), and thus adults with $Ester^1$ or $Ester^4$ are equally preferred when they are presented together to the predator ($\hat{\beta}$ values not different from $0{\cdot}5$). However, this transitivity is not always observed: for example, larvae with $Ester^4$ or $ace\text{-}1^R$ are approximately equally preferred to susceptible mosquitoes ($\hat{\beta}=0{\cdot}63$ and $0{\cdot}65$–$0{\cdot}71$, respectively), although larvae with $ace\text{-}1^R$ are strongly preferred when the alternative is larvae with $Ester^4$ ($\hat{\beta}=0{\cdot}86$). The other possible example of non-transitivity

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

in larval predation, involving individuals with *Ester*^{1}, *ace-1*^{R} and susceptible, is not conclusive because $\hat{\beta}$ for the pair SA1/S-LAB is probably underestimated (see Section 3). The non-transitivity observed for both larval and adult predation suggests that several phenotypic traits of the prey are affected by the resistance genes, and that the predator uses these cues differently according to environmental conditions.

In conclusion, predators seem to be useful tools to detect behavioural changes that are caused by these genes of recent origin. There is a large variety of potential predators for any given insect species, each with its own detection method, stimulus type and capture strategy (Lima & Dill, 1990). It is likely that any phenotypic variation will result in differential predation for at least one type of predator. We suggest that predators, which are designed by natural selection to detect specific behavioural phenotypes, are useful tools to explore non-obvious differences between two classes of individuals, for example when they differ by the presence or absence of a gene such as insecticide resistance.

We are very grateful to N. Pasteur, F. Rousset and M. Weill for helpful comments on the manuscript, to C. Bernard, M. Marquine, and G. Pistre for technical assistance, to R. Connes for the identification of *P. minutissima*, to M. Martinez for the identification of the other aquatic predators and to V. Durand for help in the literature search. This work was financed in part by MATE (PE00/122000/024). Contribution 2004.019 of the Institut des Sciences de l'Evolution de Montpellier (UMR CNRS 5554). The experiments performed here comply with the current French laws.

References

Agnew, P., Berticat, C., Bedhomme, S., Sidobre, C. & Michalakis, Y. (2004). Parasitism increasing and decreasing the costs of insecticide resistance in mosquitoes. *Evolution*, in press.

Berticat, C., Boquien, G., Raymond, M. & Chevillon, C. (2002*a*). Insecticide resistance genes induce a mating competition cost in *Culex pipiens* mosquitoes. *Genetical Research* **79**, 41–47.

Berticat, C., Rousset, F., Raymond, M., Berthomieu, A. & Weill, M. (2002*b*). High *Wolbachia* density in insecticide resistant mosquitoes. *Proceedings of the Royal Society of London, Series B* **269**, 1413–1416.

Blaustein, L. (1998). Influence of the predatory backswimmer, *Notonecta maculata*, on invertebrate community structure. *Ecological Entomology* **23**, 246–252.

Bourguet, D., Lenormand, T., Guillemaud, T., Marcel, V. & Raymond, M. (1997). Variation of dominance of newly arisen adaptive genes. *Genetics* **147**, 1225–1234.

Carrière, Y., Deland, J.-P., Roff, D. A. & Vincent, C. (1994). Life-history cost associated with the evolution of insecticide resistance. *Proceedings of the Royal Society of London, Series B* **258**, 35–40.

Chesson, J. (1984). Effect of notonectids (Hemiptera: Notonectidae) on mosquitoes (Diptera: Culicidae): predation or selective oviposition? *Environmental Entomology* **13**, 531–538.

Chevillon, C., Bourguet, D., Rousset, F., Pasteur, N. & Raymond, M. (1997). Pleiotropy of adaptive changes in populations: comparisons among insecticide resistance genes in *Culex pipiens*. *Genetical Research* **68**, 195–203.

Cohan, F. M., King, E. C. & Zawadzki, P. (1994). Amelioration of the deleterious pleiotropic effects of an adaptive mutation in *Bacillus subtilis*. *Evolution* **48**, 81–95.

Davies, A. G., Game, A. Y., Chen, Z., Williams, T. J., Goodall, S., Yen, J. L., McKenzie, J. A. & Batterham, P. (1996). *Scalloped wings* is the *Lucilia cuprima Notch* homologue and a candidate for the *Modifier* of fitness and asymmetry of diazinon resistance. *Genetics* **143**, 1321–1337.

Déom, P. (1990). L'araignée de Pluche. *La Hulotte* (*4th Edition*) **55**, 36–41.

Fisher, R. A. (1958). *The Genetical Theory of Natural Selection*. New York: Dover.

Fleury, F., Vavre, F., Ris, N., Fouillet, P. & Boulétreau, M. (2000). Physiological cost induced by the maternally-transmitted endosymbiont *Wolbachia* in the *Drosophila* parasitoid *Leptopilina heterotoma*. *Parasitology* **121**, 493–500.

Gazave, E., Chevillon, C., Lenormand, T., Marquine, M. & Raymond, M. (2001). Dissecting the cost of insecticide resistance genes during the overwintering period of the mosquito *Culex pipiens*. *Heredity* **87**, 441–448.

Georghiou, G. P., Metcalf, R. L. & Gidden, F. E. (1966). Carbamate-resistance in mosquitoes: selection of *Culex pipiens fatigans* Wied (= *Culex quinquefasciatus*) for resistance to Baygon. *Bulletin of the World Health Organization* **35**, 691–708.

Guillemaud, T., Lenormand, T., Bourguet, D., Chevillon, C., Pasteur, N. & Raymond, M. (1998). Evolution of resistance in *Culex pipiens*: allele replacement and changing environment. *Evolution* **52**, 430–440.

Kruuk, L. E. B. & Gilchrist, J. S. (1997). Mechanism maintaining species differentiation: predator-mediated selection in a *Bombina* hybrid zone. *Proceedings of the Royal Society of London, Series B* **264**, 105–110.

Laird, M. (1988). *The Natural History of Larval Mosquito Habitats*. London: Academic Press.

Lande, R. (1983). The response to selection on major and minor mutations affecting a metrical trait. *Heredity* **50**, 47–65.

Lenormand, T. & Raymond, M. (2000). Clines with variable selection and variable migration: model and field studies. *American Naturalist* **155**, 70–82.

Lenormand, T., Bourguet, D., Guillemaud, T. & Raymond, M. (1999). Tracking the evolution of insecticide resistance in the mosquito *Culex pipiens*. *Nature* **400**, 861–864.

Lenski, R. E. (1988*a*). Experimental studies of pleiotropy and epistasis in *Escherichia coli*. I. Variation in competitive fitness among mutants resistant to virus T4. *Evolution* **42**, 425–432.

Lenski, R. E. (1988*b*). Experimental studies of pleiotropy and epistasis in *Escherichia coli*. II. Compensation for maladaptive effects associated with resistance to virus T4. *Evolution* **42**, 433–440.

Lima, S. L. & Dill, L. M. (1990). Behavioral decisions made under the risk of predation: a review and prospectus. *Canadian Journal of Zoology* **68**, 619–640.

Manly, B. F. J. (1974). A model for certain types of selection experiments. *Biometrics* **30**, 281–294.

Manly, B. J. F. (1985). *The Statistics of Natural Selection on Animal Populations*. London: Chapman and Hall.

Mogi, M., Sunahara, T. & Selomo, M. (1999). Mosquito and aquatic predator communities in ground pools

on lands deforested for rice field development in central Sulawesi, Indonesia. *Journal of the American Mosquito Control Association* **15**, 92–97.

Murphey, R. K. & Mendenhall, B. (1973). Localization of receptors controlling orientation to prey by the backswimmer *Notonecta undulata*. *Journal of Comparative Physiology* **84A**, 19–30.

Orr, H. A. & Coyne, J. A. (1992). The genetics of adaptation: a reassessment. *American Naturalist* **140**, 725–742.

Raymond, M., Chevillon, C., Guillemaud, T., Lenormand, T. & Pasteur, N. (1998). An overview of the evolution of overproduced esterases in the mosquito *Culex pipiens*. *Philosophical Transactions of the Royal Society of London, Series B* **353**, 1–5.

Raymond, M., Berticat, C., Weill, M., Pasteur, N. & Chevillon, C. (2001). Insecticide resistance in the mosquito *Culex pipiens*: What have we learned about adaptation? *Genetica* **112/113**, 287–296.

Scott, M. A. & Murdoch, W. M. (1983). Selective predation by the backswimmer, *Notonecta*. *Limnology and Oceanography* **28**, 352–366.

Sih, A. (1979). Stability and prey behavioural responses to predator density. *Journal of Animal Ecology* **48**, 79–89.

Sih, A. (1986). Antipredator responses and the perception of danger by mosquito larvae. *Ecology* **67**, 434–441.

Strickman, D., Sithiprasasna, R. & Southard, D. (1997). Bionomics of the spider, *Crossopriza lyoni* (Araneae, Pholcidae), a predator of dengue vectors in Thailand. *Journal of Arachnology* **25**, 195–201.

Sunish, I. P. & Reuben, R. (2002). Factors influencing the abundance of Japanese encephalitis vectors in ricefields in India. II. Biotic. *Medical and Veterinary Entomology* **16**, 1–9.

Genet. Res., Camb. (2004), **83**, *pp.* 197–209. *With 4 figures.* © 2004 Cambridge University Press
DOI: 10.1017/S0016672304006779 Printed in the United Kingdom

Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth

ÖRJAN CARLBORG*, PAUL M. HOCKING, DAVE W. BURT AND CHRIS S. HALEY
Roslin Institute, Roslin, Midlothian, EH25 9PS, UK

(*Received 21 July 2003 and in revised form 19 January 2004*)

Summary

We used simultaneous mapping of interacting quantitative trait locus (QTL) pairs to study various growth traits in a chicken F_2 intercross. The method was shown to increase the number of detected QTLs by 30% compared with a traditional method detecting QTLs by their marginal genetic effects. Epistasis was shown to be an important contributor to the genetic variance of growth, with the largest impact on early growth (before 6 weeks of age). There is also evidence for a discrete set of interacting loci involved in early growth, supporting the previous findings of different genetic regulation of early and late growth in chicken. The genotype–phenotype relationship was evaluated for all interacting QTL pairs and 17 of the 21 evaluated QTL pairs could be assigned to one of four clusters in which the pairs in a cluster have very similar genetic effects on growth. The genetic effects of the pairs indicate commonly occurring dominance-by-dominance, heterosis and multiplicative interactions. The results from this study clearly illustrate the increase in power obtained by using this novel method for simultaneous detection of epistatic QTL, and also how visualization of genotype–phenotype relationships for epistatic QTL pairs provides new insights to biological mechanisms underlying complex traits.

1. Introduction

The desire to dissect the underlying mechanisms of complex traits has led to detection of major genes and quantitative trait loci (QTLs) for many traits in various species. The traditional way to detect major genes and QTLs is by looking for marginal (additive and dominance) effects of the individual loci. Larger sample sizes in QTL mapping studies have increased the opportunity to study the importance of more complex genetic mechanisms such as epistasis. Epistasis has been sought by estimation of the epistatic effects of combinations of QTLs detected by their marginal effects (e.g. Chase *et al*., 1997) or by using one-dimensional searches with an epistatic model, while including markers to control background genetic effects (e.g. Fijneman *et al*., 1996). Some attempts have also been made to develop methods that assess the physiological importance of epistasis (Cheverud & Routman, 1995). More recently, several new methods and technologies have been proposed to increase the power to map epistatic QTLs by performing genome-wide mapping of epistatic QTLs (e.g. Boer *et al*., 2002; Carlborg *et al*., 2000; Carlborg & Andersson, 2002; Kao *et al*., 1999; Jannink & Jansen, 2001; Sen & Churchill, 2001). Several of the methods have also been evaluated by simulation and several have also been applied to map interacting QTLs in various experimental populations (e.g. Carlborg *et al*., 2003; Leamy *et al*., 2002; Peripato *et al*., 2002; Shimomura *et al*., 2001; Zeng *et al*., 2000). The application of newly developed methods to experimental datasets is an important part of the process of developing improved method, because it gives new insights into various properties of the analytical method. It also gives an indication of the potential of the new method for revealing previously unnoticed phenomena in experimental data.

Conventional genetic selection has resulted in lines of laying fowl that are small and lean, and produce many eggs in the course of a laying year. Selection of fowl for high growth rates, high muscle yields and

* Corresponding author. Tel: +44 (0)131 527 4258. Fax: +44 (0)131 440 0434. e-mail: Orjan.Carlborg@bbsrc.ac.uk

improved feed efficiency has led to the creation of very large, heavily muscled broiler lines with relatively poor reproductive fitness. Several recent studies have reported associations between genetic markers and quantitative traits of economic importance in chickens (e.g. Dunnington *et al*., 1992; van Kaam *et al*., 1998, 1999; Ikeobi *et al*., 2002; Sewalem *et al*., 2002). The current study is based on a cross between a layer line with a small body size and a sire line of broiler parent stock with a very large body size that were crossed to produce an F_2 in which many traits were characterized. This cross has previously been analysed using a variety of traditional QTL mapping techniques (Ikeobi *et al*., 2002; Sewalem *et al*., 2002). Here, we use the method described by Carlborg *et al*. (2003) to map epistatic QTLs and to evaluate the relative contribution of epistasis to live weight at 3, 6 and 9 weeks of age and for the growth in the age intervals 3–6 weeks and 6–9 weeks of age.

2. Animal material

The mapping population consisted of a three generation F_2 cross between a White Leghorn line and a commercial broiler sire line. The layer was derived from a commercial pure line and the broiler sire line had been selected for high growth rates and breast muscle yields as part of a commercial breeding program. Three females and three males from both lines were used to generate six F_1 families. Subsequently, four of these families (two each of broiler male × layer female and layer male × broiler female) were used to create the F_1 population. Each F_1 family contained 10–16 birds. Eight male and 32 female F_1 were selected to produce an F_2 generation of 546 chickens. The recorded traits were body weight at 3, 6 and 9 weeks of age, and, from these, growth rates at 3–6 and 6–9 weeks of age were calculated. For the total genome scan, 134 microsatellite markers covering 30 autosomal linkage groups and the sex chromosomes were typed on eight F_0 grandparents, 40 F_1 parents and 510 F_2 chickens. After parentage checking and genotyping edits in the F_2, data from 466 F_2 chicks in 30 full-sib families with genotypes on 101 microsatellites covering 27 linkage groups were available for analysis. The total map length was 2499 cM. The average marker spacing was 40 cM and the average polymorphic information content was 0·61 (ranging from 0·19 to 0·98). The sex chromosomes were excluded in the search for epistatic QTLs. A more thorough description of the mapping population can be found in Sewalem *et al*. (2002).

3. QTL mapping methods

This report uses two QTL mapping methods based on two genetic models (without and with epistasis) and two genomic search strategies, forward selection (FS) and simultaneous search (SIM) to map QTLs in an outbred F_2 chicken cross. The methods are compared based on the differences in the number of significant QTLs detected and the amount of genetic variance explained by the detected QTLs. Method I (FS) is a traditional strategy for QTL mapping based on a linear model with marginal (additive and dominance) effects for multiple QTLs. The final genetic model is built by forward selection of significant marginal effects of individual QTLs. Method II (SIM) is a method for simultaneous mapping of epistatic QTLs (Carlborg & Andersson, 2002; Carlborg *et al*., 2003), which is based on a linear model with marginal effects for a pair of QTLs and their four possible pairwise interactions. The locations for the two QTLs in the model are selected simultaneously using either an exhaustive search (in the real data) or a genetic algorithm (during randomization testing). The contribution of epistasis to the genetic variance explained by the pair was evaluated for all the pairs. The procedures outlined here are described in more detail in the following sections.

(i) *Linear models for single and multiple QTLs*

In the marginal effects genetic model, used for forward selection of non-epistatic QTL (method I, FS), each QTL is modelled by its marginal (additive and dominance) effects

$$y = \beta_0 + FZ + \beta_{1j} a_j + \beta_{2j} d_j + \epsilon_j \quad (1)$$

where y_i is a vector of phenotypes, β_0 is the mean, F is a vector of regression coefficients for full-sib family, sex, rearing pen and earlier detected QTLs, Z is a matrix of regression variables for full-sib family, sex, rearing pen and earlier detected QTLs, β_{1j}, β_{2j} are regression coefficients for additive and dominance effects at genomic location j, and a_j and d_j are regression indicator variables for additive and dominance effects at genomic location j.

For simultaneous mapping of QTL pairs (method II, SIM), the linear model is a non-orthogonal expansion of model I to include also the marginal genetic effects of a second QTL and the four pairwise interaction terms for a QTL pair

$$y = \beta_0 + FZ + \beta_{1jk} a_j + \beta_{2jk} d_j + \beta_{3jk} a_k + \beta_{4jk} d_k + \beta_{5jk} aa_{jk} + \beta_{6jk} ad_{jk} + \beta_{7jk} da_{jk} + \beta_{8jk} dd_{jk} + \epsilon_{jk} \quad (2)$$

where y, β_0, F and Z are the same as in model I, β_{1jk}, β_{2jk}, β_{3jk} and β_{4jk} are regression coefficients for additive and dominance effects for QTLs at locations j and k cM, β_{5jk}, β_{6jk}, β_{7jk} and β_{8jk} are regression coefficients for epistatic effects between QTLs at locations j and k cM, a_j, d_j, a_k and d_k are regression indicator variables for additive and dominance effects for QTLs

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

at locations j and k cM, and aa_{jk}, ad_{jk}, da_{jk} and dd_{jk} are regression indicator variables for epistatic effects for QTLs at locations j and k cM.

(ii) *Parameter estimation*

Estimation of the genetic effects for QTLs was performed using variations of the commonly used least squares framework for QTL mapping in inbred and outbred crosses (Haley & Knott, 1992; Haley *et al*., 1994). This framework involves two independent tasks. First, QTL genotype probabilities are estimated throughout the genome conditional on the measured marker genotypes. Second, the QTL genotype probabilities are used to calculate regression indicator variables for the genetic effects of QTL, which are then used to estimate the genetic effects using least squares. In this F_2 population, the marker genotypes were used to estimate the probability of an F_2 offspring being each of the four QTL genotypes (QQ, Qq, qQ and qq) at 1 cM intervals throughout the genome. The marginal QTL effects considered are additive (allele substitution) and dominance (heterozygote deviation) effects (model I above). Haley & Knott (1992) describe how to form additive (a_i) and dominance (d_i) indicator regression variables as

$$a_i = P(QQ)_i - P(qq)_i$$
$$d_i = P(Qq)_i + P(qQ)_i,$$

where i is the genome location of QTL$\subset$[1$\cdots$ genome size cM], and $P(\mathrm{XX})_i$ is the conditional probability of the individual having QTL genotype XX at location i given the flanking marker genotypes. We did not consider parental origin effects because there was no evidence of imprinting in this population (Sewalem *et al*., 2002).

Method II involves a search for pairwise interactions between QTLs, and the genetic model to evaluate these effects includes four interaction effects in addition to the marginal effects of the two QTLs in the pair. To estimate these effects (additive by additive, additive by dominance, dominance by additive and dominance by dominance interactions), a new set of indicator regression variables needs to be calculated. Haley & Knott (1992) indicated that the indicator regression variables could be calculated by multiplying the respective additive and dominance regression indicator variables for the QTL in the pair

$$aa_{ij\mathbf{12}} = a_{i\mathbf{1}} \times a_{j\mathbf{2}},$$
$$ad_{ij\mathbf{12}} = a_{i\mathbf{1}} \times d_{j\mathbf{2}},$$
$$da_{ij\mathbf{21}} = d_{i\mathbf{2}} \times a_{j\mathbf{1}},$$
$$dd_{ij\mathbf{22}} = d_{i\mathbf{1}} \times d_{j\mathbf{2}},$$

where i and j are the genome locations in cM of QTLs 1 and 2$\subset$(1$\cdots$ genome size cM).

Using these indicator regression variables, the genetic parameters for single QTL (model I) and epistatic QTL pairs (model II) can be estimated using ordinary least squares.

(iii) *Forward selection interval mapping*

A simple way to map multiple QTLs is by forward selection of non-interacting QTLs. This was the first analysis we performed to detect significant marginal (additive and dominance) effects of QTLs (Fig. 1, step I). QTL genotype probabilities were calculated at 1 cM intervals and QTLs were fitted using model I at 1 cM intervals using ordinary least squares (Haley *et al*., 1994). The additive and dominance regression indicator variables for the most significant single QTL in this scan were added as cofactors to model I and a new genome scan was performed using the updated model. This procedure was repeated until no additional significant QTLs were detected. Statistical significance was assessed by randomization testing (Churchill & Doerge, 1994) in each step of the procedure using a 5% genome-wide threshold for significant and a 20% genome-wide significance threshold for nearly significant QTLs. All randomization tests are based on analyses of 1000 permuted datasets.

(iv) *Simultaneous interval mapping*

Simultaneous mapping of epistatic QTLs increases the power to detect interacting QTL. The principle of the SIM method performed here is as follows (Fig. 1, step II). First, QTL genotype probabilities were calculated at 1 cM intervals according to Haley *et al*. (1994). An exhaustive simultaneous search for interacting QTL pairs in the real data was performed using model II. For all fitted pairs, the parameters of the model were estimated using least squares and the model fit (residual sum of squares) was retained. Significance of fitted QTL pairs was assessed in three ways depending on the number of QTLs in the pair that had significant marginal effects (for further detail on the randomization procedures see Carlborg & Andersson, 2002). (i) When both QTLs in the pair had significant marginal effects in the FS procedure described above, the QTL pair was declared significant without further significance testing. (ii) Where one of the QTLs in the pair had significant marginal effects, a randomization test was used to test for the combined effects of the marginal effects of the second QTL and the interaction parameters for the pair, conditional on the significant marginal effects of the first QTL. (iii) Where neither of the QTLs had significant marginal effects, the significance of the pair is assessed using a randomization test for a QTL pair without significant marginal effects. For all these tests, a 5% genome-wide threshold was used to declare significant

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

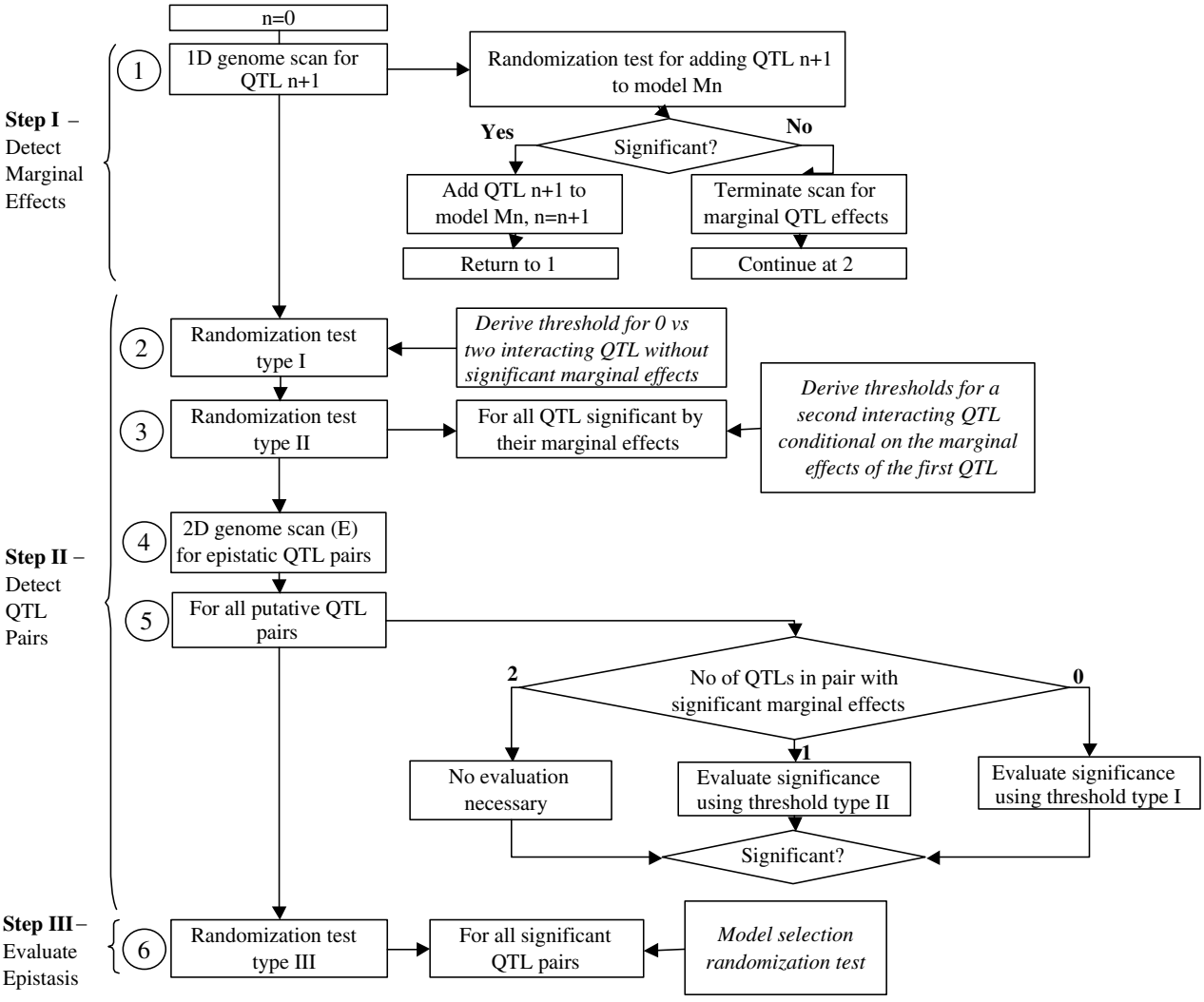


Fig. 1. The analysis procedure used for detection of QTL. Further explanation of the figure can be found in the text. Abbreviations: 1D, one-dimensional; 2D, two-dimensional; E, exhaustive search.

and a 20% genome-wide significance threshold to declare nearly significant QTL pairs.

(v) *Model selection for significant QTL pairs*

To evaluate whether epistasis contributed significantly to the genetic variance explained by significant and nearly significant QTL pairs, a randomization test was used to test whether a model including both marginal (additive and dominance) and epistatic QTL parameters significantly improved the fit over a model including only marginal QTL effects (Carlborg & Andersson, 2002). A nominal 5% significance threshold was used for each of these tests.

(vi) *Multiple regression modelling*

To compare the explanatory power of the QTL detected by the SIM and FS procedures, we used multiple regression modelling to fit simultaneously all QTL detected using the FS and SIM procedures. For all traits in turn, we fitted model I with (i) all QTLs detected by FS and (ii) all QTLs detected by SIM. The fits of the models were compared by the reduction of the residual sums of squares of these two models by including the genetic effects of the QTLs. The relative importance of epistasis for the analysed traits was assessed by comparing the fit of model I, including the marginal effects for all QTL detected by the SIM procedure, and model II, including the same marginal effects together with interaction effects for pairs where an epistatic QTL model was significantly better than a marginal effects model. The variances contributed by the marginal effects (model I) and by the marginal and epistatic effects (model II) were compared using the reduction of the residual sum of squares by the respective models by including the genetic effects of the QTLs.

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

(vii) *Interpretation of epistasis*

If the genetic mechanism behind the observed pairwise QTL interactions could be understood, the information would be valuable for identifying candidate genes for detected QTLs. We have therefore plotted the nine genotype class means for all the significant and nearly significant QTL pairs to identify similarities among the interaction patterns and those of classic mendelian patterns of epistasis. The genotype class means were estimated using the SAS software, by regressing phenotypes on fixed effects and the two-locus genotype probabilities (calculated by multiplying the single locus genotype probabilities described above) of the QTL pair.

4. Computational methods

In QTL mapping, the genome is modelled as a grid based on genetic markers (where each marker is a node in the grid) or on genetic map locations (where each node in the grid is a genomic location in cM). The grid is one-dimensional during a search for a single QTL and multidimensional when multiple QTLs are sought. A genome scan involves fitting a statistical model at multiple locations in the genomic grid with the objective of finding the location(s) in the genome with significant statistical support for a QTL or multiple QTLs. We use a genetic-map-based grid with a genetic distance of 1 cM (Kosambi) between the nodes.

We have used three different algorithms to select the QTLs to be evaluated among all the possible combinations of QTLs that exist in the grid. Below, we give a short introduction to these methods but, for a more thorough discussion of methods to search for QTLs in genetic grids, we refer to Carlborg (2002).

(i) *Exhaustive search*

An exhaustive search involves fitting the statistical model at all nodes in the one- or multidimensional grid. The method guarantees that the best location in the grid, at the given resolution, is found, but at the price of a high computational demand. The computational demand for using an exhaustive search in a one-dimensional grid (i.e. a search for a single QTL), randomization testing in one-dimensional grids or isolated scans in two-dimensional grids (i.e. fitting two QTLs simultaneously in real data) is not prohibitively high, especially when parallel computers are used for the analysis (Carlborg, 2002). However, randomization testing based on two-dimensional grids and scans in grids of dimensions higher than two is computationally intractable using an exhaustive search and, for this, alternative search methods are needed.

(ii) *Forward selection*

Forward selection is a method to reduce a scan of a multidimensional grid to a series of one-dimensional scans. In QTL mapping, the method has been used to map multiple non-interacting QTLs, where the most significant QTL from a series of successive exhaustive one-dimensional genome scans are sequentially added to a multiple-QTL model. The method is expected to perform well when the QTLs are independent, which is the case for non-interacting and non-linked QTLs, and has been widely used for this purpose. We have selected this method to represent a traditional method to search for multiple non-interacting QTLs.

(iii) *Genetic algorithm*

Genetic algorithms are search algorithms based on the mechanisms of genetics and natural selection, and can be used to perform a multidimensional search in a more computationally efficient way than using an exhaustive search. The advantage of using a true multidimensional search instead of a search based on repetitive one-dimensional searches is expected to be greater for interacting QTLs than for non-interacting QTLs, because pairs of QTLs with non-significant marginal effects will not be found in a series of one-dimensional searches. The importance of using true multidimensional searches when mapping interacting QTLs was first shown by Carlborg *et al*. (2000), where a genetic algorithm was shown to be more efficient in detecting interacting QTL than an FS-based method. Here, a genetic algorithm has been used to reduce the computational demand during randomization testing for interacting pairs of QTLs without significant marginal genetic effects. We used a genetic algorithm (GA) from a library named PGAPack (Levine, 1996). Ten independent GA populations of 20 individuals with 1000 iterations per population were used for two-dimensional genome scan. For each independent GA population, a local exhaustive search of ± 5 cM was performed around the found optimum after the GA had converged. More information on specific parameters settings for PGAPack can be found in Carlborg *et al*. (2000).

5. Results

(i) *Detection of non-interacting and interacting QTLs*

For the five analysed bodyweight and growth traits, a total of nine QTL regions were detected as significant using a 5% genome-wide significance threshold (Table 1). Three of the regions (chromosome 1, 150 cM; chromosome 1, 470 cM; chromosome 27, 0 cM) were only detected by their marginal effects and one region was only detected using simultaneous mapping of epistatic QTL pairs (chromosome 2,

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

Table 1. *Genomic regions with a significant or nearly significant QTL affecting at least one growth trait. The information content at the location of the QTL and the markers flanking the QTL peak are also given*

QTL		LM		RM		BW3		BW6		BW9		GR36		GR69		Sum			IC	Int	
GGA	Pos	Name	Pos	Name	Pos	FS	SIM	FS	SIM	FS	SIM	FS	SIM	FS	SIM	FS	SIM	T		Y/N	Pairs
1	70	MCW010	48	ADL188	109	c	c	a	a	a	–					a	a	a	L	Y	1
1	150	LEI146	145	MCW007	178	–	c	a	c			–	c			a	c	a	H	Y	5(4)
1	390	MCW036	362	LEI106	394	–	c	–	b	a	a	a	c	c	–	a	a	a	M	Y	4(3)
1	470	LEI079	422	ROS025	503	a	c	a	b							a	b	a	L	Y	1
2	240	ADL196	225	LEI127	270			–	a			–	b			–	a	a	M	Y	3(2)
2	290	LEI127	270	ROS074	302			–	b	c	–	c	–			c	b	b	H	N	
3	50	MCW083	51	MCW083	51			–	b					–	c	–	b	b	M	Y	1
4	165	ADL266	126	LEI073	231			a	a	a	c	a	b	a		a	a	a	L	N	
5	127	ROS084	57	ADL298	166									–	c	–	c	c	L	Y	1
6	35	ROS003	33	ADL142	51			a	a			a	b			a	a	a	H	Y	4
7	105	ROS019	101	ADL180	109							–	c			–	c	c	H	Y	1
8	15	ADL179	11	ROS075	80					c	c	c	c			c	c	c	L	Y	1
13	55	ADL147	32	ADL255	70	a	c	a	a	c	a	a	b			a	a	a	M	Y	5(3)
18	15	ROS022	0	ROS027	23			–	c			–	c			–	c	c	H	Y	2(1)
27	0	ROS071	0	ROS071	0			c	–			a	c			a	c	a	H	Y	2

Abbreviations: GGA = *Gallus gallus* chromosome, Pos = Estimated chromosomal position (cM) based on the results from all traits, No = QTL ID number, LM/RM = Left/Right Marker flanking QTL interval (LM = RM if the QTL peak is located at a marker), BW3/6/9 = Body weight at 3/6/9 weeks of age, GR36/69 = Growth from 3 to 6 and 6 to 9 weeks of age, Sum = Summary of QTL mapping results, FS = Significance of QTL mapped using forward selection, SIM = Significance of QTL mapped using Simultaneous mapping, T = Significance of QTL mapped by the entire SIM procedure, a/b/c = QTL significant at 5/10/20% genome-wide significance threshold, Int = QTL involved in interactions, Y/N = Yes/No, Pairs = No. of epistatic pairs (No. of *unique* epistatic pairs) in which QTL is involved, IC = combined information content for QTL location classified as 0 < Low (L) < 0·30, 0·31 < Medium (M) < 0·60, 0·61 < High (H) < 1·00.

240 cM). When a 20% genome-wide significance threshold was used, 15 QTL regions were detected, and five of these (chromosome 2, 240 cM; chromosome 3, 50 cM; chromosome 5, 127 cM; chromosome 7, 105 cM; chromosome 18, 15 cM) were only detected using simultaneous mapping of epistatic QTL pairs. A summary of all QTL pairs that were detected for the five analysed traits and the model that was selected for each of the pairs are given as supplementary information on the publisher's website. Two QTL regions (chromosome 2, 290 cM; chromosome 4, 165 cM) were never included in a significant epistatic QTL pair, and seven regions were significant on more than one occasion and two of these were only detected using SIM (chromosome 2, 240 cM; chromosome 18, 15 cM). The least squares estimates for the two-locus genotypes for all detected QTL pairs are given in Table 2.

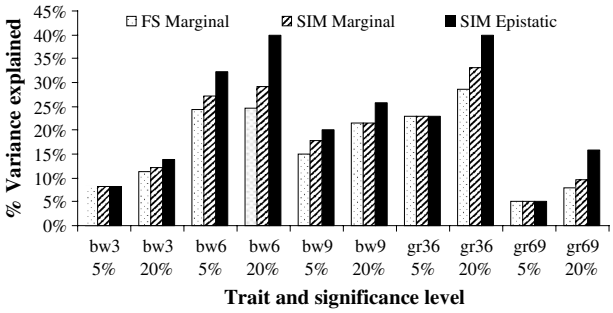


Fig. 2. The variance explained by the QTLs detected by their marginal effects using forward selection and by the SIM procedure for the five analysed traits, calculated as the reduction of the residual sums of squares by adding marginal and epistatic genetic effects to the model. Abbreviations: FS, QTL mapped using forward selection; SIM, QTL mapped using simultaneous mapping; BWX, bodyweight at X weeks of age; GRXY, growth from X to Y weeks of age; Marginal, model used included additive and dominance effects; Epistatic, model used included additive, dominance and epistatic effects.

(ii) *Variation explained by epistasis*

The additional residual phenotypic variance explained by adding significant epistatic parameters to the genetic model varied from 0% to 34% using a 5% genome-wide threshold and from 20% to 103% using a 20% genome-wide threshold. The largest contribution of epistasis when using the 5% threshold was for bodyweight at 6 weeks and 9 weeks. For the QTLs detected using the 20% threshold, the largest contribution of epistasis was found for bodyweight at 6 weeks and the growth rates at 3–6 and 6–9 weeks. Fig. 2 shows the amount of residual phenotypic variation explained by the QTLs detected by their marginal effects using FS and by the SIM procedure for the five analysed traits.

(iii) *Interpretation of epistasis*

In total, 30 QTL pairs were detected for the five analysed traits (Table 3) and, for 16 of these, an epistatic model was selected. Among the 30 pairs, there were 21 unique combinations of loci that had significant genetic effects on at least one trait. The patterns among the genotypic effects for the locus pairs that had significant effects on multiple traits were very similar and so only the unique combinations were evaluated further. Four clusters of QTL pairs with similar genetic effect patterns were identified by visual inspection (representative pairs are given in Fig. 3).

The first cluster consists of five pairs, in which several of the homozygote–heterozygote genotypes have lower phenotypes than expected under a two-locus additive model. An epistatic model was significant for four of the pairs in this group. An example of a pair from this cluster is shown in Fig. 3A.

The second cluster contains six pairs, in which the broiler double homozygote has a lower phenotype than expected given the other genotypic-effects for the pair. The QTL pair in Fig. 3B is an example from this cluster. An epistatic model was selected for the three most significant of the pairs in this group. Three of the six pairs include a QTL on chromosome 1, closely linked to marker LEI106 at 393 cM, and two more pairs include a close, but unlinked, QTL located at 455 cM.

The third cluster includes four pairs that show a continuous increase in the phenotype from the low phenotype Leghorn double homozygote to the high phenotype broiler double homozygote. The transition of the phenotype between the genotypes varies from near linear ('additive') to non-linear. Fig. 3C shows a pair from this cluster with a non-linear phenotype transition between the genotype classes. An epistatic model was selected for the two most non-linear of the pairs in this group.

The last identified cluster includes two pairs that have their genotypic effects divided into three distinct classes, in which the high-effect group contain broiler homozygotes or the double heterozygote, the intermediary-effect group only contains the Leghorn double homozygote and the low-effect group contain the rest of the genotype classes. Both of the pairs are significantly epistatic. Fig. 3D shows one of these pairs.

There are no striking similarities with a mendelian pattern of digenic epistasis or other similarities among the remaining five pairs and they have not been classified further. The plots of the genotypic effects for all the 22 genotype combinations are given as supplementary information on the publishers website.

QTLs with significant pairwise interactions or non-significant interactions for QTL clustering into groups 1 and 2 above were joined by connecting arrows to

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

Table 2. *Estimates of the genotypic effects* (*as deviations from the LLLL genotype*) *and the respective standard errors for the QTL pairs detected in the study*

QTL information			Genotypes																	
			BBBB		BBBL		BBLL		BLBB		BLBL		BLLL		LLBB		LLBL		LLLL	
Pair no	Trait	Location	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	bw3	1-147|1-373	72·0	19·6	82·4	18·8	39·2	20·1	35·2	19·1	47·2	17·7	50·4	19·4	50·9	19·6	60·3	21·8	0·0	–
2	bw3	1-147|1-471	18·1	34·4	36·7	32·7	−53·8	34·8	2·3	31·9	−7·5	29·5	−37·6	31·2	35·4	27·9	−43·7	43·6	0·0	–
3	bw3	1-471|13-59	141·1	51·3	167·8	46·3	70·3	38·0	116·5	50·0	93·7	35·7	121·9	56·3	114·5	50·2	106·6	51·7	0·0	–
4	bw6	1-71|6-34	123·6	100·9	84·3	82·3	96·2	82·1	126·9	85·3	118·2	75·4	−47·5	116·6	−51·9	92·0	−72·9	83·8	0·0	–
2	bw6	1-150|1-455	126·0	107·4	96·8	101·0	−109·0	106·5	35·9	100·4	4·8	88·7	−73·5	102·8	112·9	87·6	−127·4	134·1	0·0	–
5	bw6	1-150|4-168	381·6	144·8	160·2	134·1	−12·5	149·6	165·7	135·1	136·7	114·0	3·5	143·8	339·8	113·1	4·3	191·3	0·0	–
6	bw6	1-150|18-13	61·7	64·3	73·2	56·5	−83·9	62·9	39·2	59·1	−54·3	53·8	−50·7	61·1	−65·3	64·8	−67·4	65·7	0·0	–
7	bw6	1-383|6-34	87·3	59·7	198·2	49·5	125·7	52·6	210·5	48·7	141·4	44·3	112·1	58·8	131·4	55·6	91·7	55·1	0·0	–
8	bw6	1-383|13-56	173·4	86·9	391·0	75·0	120·5	74·9	398·5	77·9	233·9	63·7	243·0	79·7	218·1	83·4	265·2	91·4	0·0	–
9	bw6	1-455|4-168	81·6	324·8	566·4	352·7	−105·7	259·5	495·3	328·5	74·6	191·1	152·2	390·3	194·0	216·6	140·6	391·8	0·0	–
3	bw6	1-455|13-56	387·8	147·1	512·9	132·5	219·6	108·9	365·3	143·2	325·0	102·3	346·8	161·2	366·0	143·9	296·9	148·2	0·0	–
10	bw6	2-239|6-34	178·8	59·7	160·8	53·5	136·2	56·2	159·7	53·5	155·8	49·1	85·6	63·2	123·0	64·4	86·6	54·6	0·0	–
11	bw6	2-239|13-56	360·5	93·8	208·9	69·6	77·9	70·2	88·2	72·3	246·1	58·1	117·7	75·9	344·1	79·7	62·6	85·2	0·0	–
12	bw6	3-34|6-34	148·5	52·3	55·6	47·4	−61·2	51·3	1·0	43·4	59·2	40·4	−30·5	47·9	67·2	53·6	−52·4	48·4	0·0	–
13	bw6	4-168|6-34	476·6	144·1	317·8	124·0	323·5	103·2	210·9	129·2	259·6	115·8	165·3	180·8	118·6	144·1	89·7	127·5	0·0	–
14	bw6	4-168|13-56	481·1	186·9	286·5	171·7	250·1	132·6	59·5	181·7	316·4	122·2	112·0	209·2	364·0	188·0	−13·6	180·5	0·0	–
15	bw9	1-397|4-168	495·3	308·8	456·8	263·0	−60·8	271·4	607·4	256·0	303·5	219·5	72·5	266·7	363·9	202·2	184·1	352·3	0·0	–
8	bw9	1-397|13-55	301·9	129·6	508·8	110·1	113·4	113·4	447·4	112·3	307·6	97·7	318·9	108·6	352·6	127·1	239·8	131·7	0·0	–
16	bw9	8-13|13-55	911·4	270·9	−83·1	216·6	99·0	188·7	−328·7	230·3	334·0	147·4	−131·0	259·9	347·0	241·9	−270·2	230·5	0·0	–
6	gr36	1-145|18-16	46·9	43·8	44·7	38·1	−67·5	42·6	34·8	39·7	−39·7	36·6	−28·7	38·8	−22·7	44·5	−54·6	43·8	0·0	–
17	gr36	1-145|27-0	182·3	41·9	100·4	35·5	34·5	39·4	102·6	37·4	57·0	34·6	47·1	35·9	60·6	44·8	54·1	37·8	0·0	–
15	gr36	1-400|4-170	134·3	149·1	255·1	127·4	−53·3	129·9	306·1	122·7	107·2	104·6	49·2	130·5	170·5	95·6	73·4	168·1	0·0	–
7	gr36	1-400|6-33	92·5	37·8	145·3	32·2	90·1	36·4	129·4	31·9	116·5	28·8	79·1	36·1	117·0	36·1	48·8	34·7	0·0	–
11	gr36	2-245|13-52	269·4	79·3	152·5	60·6	94·5	60·5	65·3	65·2	205·8	50·0	90·4	67·4	275·7	68·8	42·3	75·7	0·0	–
13	gr36	4-170|6-33	294·9	109·2	243·4	93·2	258·3	77·6	180·0	97·5	167·5	88·3	75·8	136·9	22·3	109·3	50·9	94·9	0·0	–
14	gr36	4-170|13-52	375·6	153·7	208·2	141·4	254·1	106·3	60·7	151·4	247·5	96·6	71·7	168·2	251·2	153·6	2·1	148·0	0·0	–
18	gr36	6-33|27-0	146·4	38·9	104·8	30·9	72·5	35·1	154·1	32·5	90·7	28·9	65·2	32·2	64·7	39·0	58·6	32·4	0·0	–
19	gr36	7-105|27-0	77·8	40·9	121·0	33·3	52·5	40·9	142·3	35·1	44·6	31·6	44·2	35·4	92·5	40·9	53·3	34·9	0·0	–
20	gr36	8-25|27-0	333·5	135·3	181·1	86·8	239·1	78·6	120·5	93·7	206·5	73·6	141·1	117·5	271·2	108·2	69·6	83·8	0·0	–
21	gr69	3-63|5-127	720·3	205·0	−386·1	202·2	260·2	193·1	−331·9	203·2	213·6	160·5	−98·4	219·5	240·2	152·3	−162·2	270·9	0·0	–

Abbreviations: Pair no, number of unique QTL pairs if the same pair has significant effects for more traits the pair has the same number; L, Leghorn allele; B, broiler allele; XXYY, genotype XX at locus 1 and genotype YY at locus 2.
Locations: a-b|c-d, first QTL at chromosome a in location b and second QTL at chromosome c in location d.

Table 3. *Number of QTL pairs identified by a simultaneous mapping strategy for epistatic QTL pairs (SIM) and the number of pairs detected with a marginal effects model including additive and dominance effects (A + D) and an epistatic QTL model (E) were selected. Also, the number of times two, one or none of the QTLs in the detected pair were also detected using forward selection (FS) and a marginal effects model*

	5% genome-wide significance						20% genome-wide significance					
		Selected model		Detected by FS				Selected model		Detected by FS		
	No of pairs by SIM	A + D	E	2	1	0	No of pairs by SIM	A + D	E	2	1	0
BW3	0	0	0	0	0	0	3	1	2	2	1	0
BW6	4	1	3	2	2	0	13	6	7	8	5	0
BW9	1	0	1	0	1	0	3	1	2	3	0	0
GR36	0	0	0	0	0	0	10	6	4	6	3	1
GR69	0	0	0	0	0	0	1	0	1	0	0	1

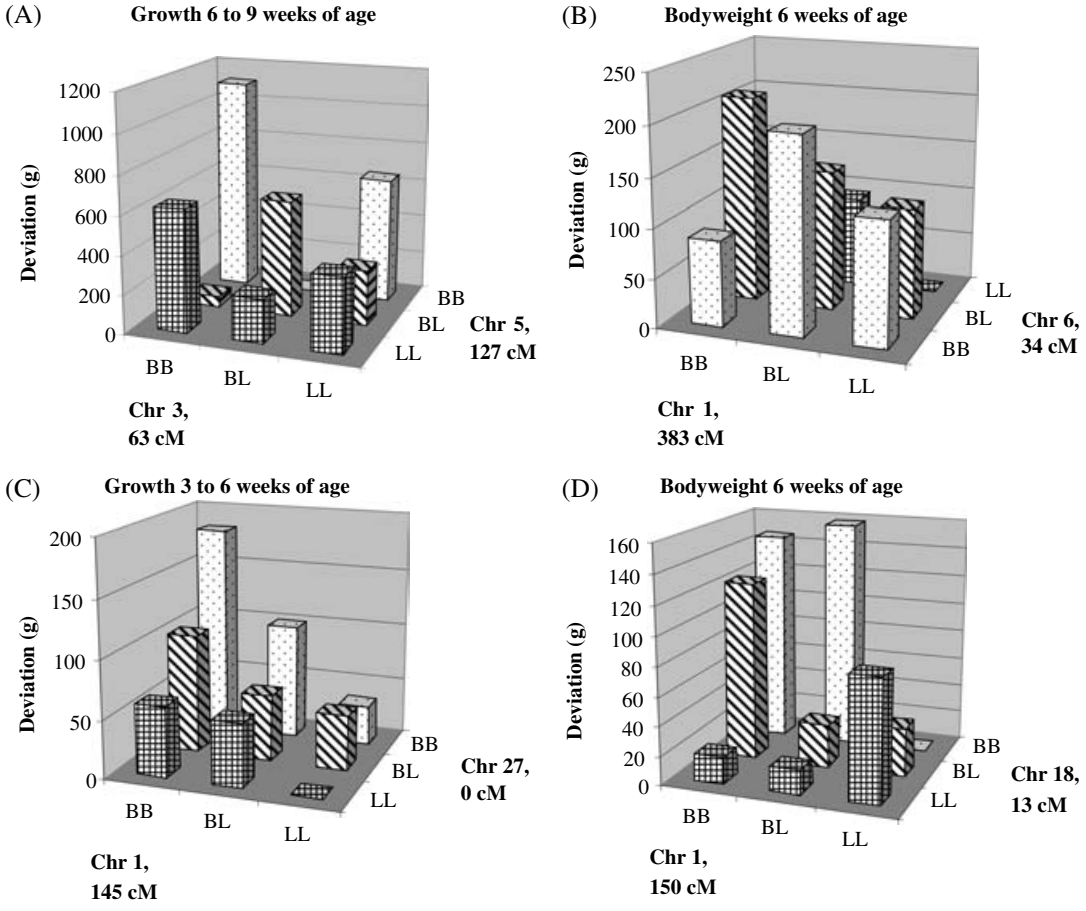


Fig. 3. Phenotypic expression in the nine genotype classes for representative epistatic QTL pairs from the four clusters of QTL pairs with similar genetic effects on growth in a broiler layer cross. Abbreviations: Deviation, phenotype expressed as the deviation (in grams) of the phenotype from the genotype class with the lowest mean; Chr, chromosome; Pos, position; BB, genotype is homozygote broiler; BL, genotype is heterozygote; LL, genotype is homozygote layer.

understand further the genetic architecture of the traits (Fig. 4). The figure shows a chain of eight QTLs linked by pairwise interactions, two branches with a single QTL and three loops on the chain (two of which are created by non-linear type interactions). Two QTLs (chromosome 2, 290 cM, and chromosome 4, 165 cM) are not included in the figure because they were not involved in any significant pairwise interactions.

6. Discussion

The use of efficient computational algorithms in QTL mapping allows researchers to move from approximate

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

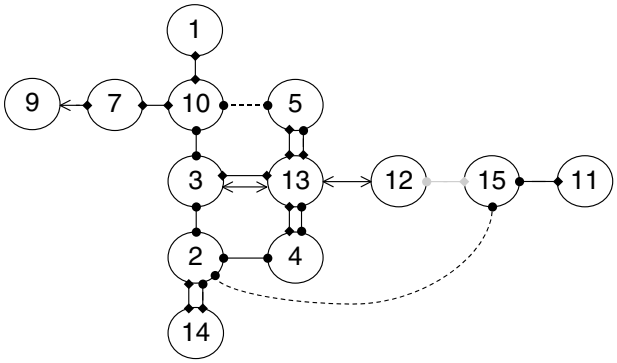


Fig. 4. Graphical representation of the interactions detected between the QTLs affecting growth in a broiler × layer cross. Circles indicate QTLs and the number in the circle is the relevant QTL number given in Table 1. The connectors between the circles indicate two-locus interactions, where black connectors indicate significant epistatic interactions and grey connectors indicate interactions where the significance threshold is not reached but where inspection of the genotype means indicate that the QTL pair belongs to one of the four interaction clusters identified. Dashed connectors indicate additive-like interactions. The style of the ends of the connectors indicate the trait for which the interaction was significant: circular, body weight at 3 weeks; square, body weight at 6 weeks; arrow, body weight at 9 weeks; circular + square, growth between 3 weeks and 6 weeks; square + arrow, growth between 6 weeks and 9 weeks.

methods to screen for epistasis to true multi-dimensional searches (Carlborg *et al*., 2001; Carlborg, 2002; Ljungberg *et al*., 2002). We have previously shown by simulations that simultaneous mapping of multiple epistatic QTLs has the potential to increase the power to map interacting QTLs (Carlborg *et al*., 2000; Carlborg & Andersson, 2002). The benefits of using this method are, however, not universal. Because the true genetic architecture of complex traits, and therefore the impact of epistasis, is unknown, the potential benefit of using these methods can only be assessed when they are applied in analyses of experimental data. High power to detect epistasis can only be expected in reasonably large datasets with high-quality phenotypic measurements and highly informative markers. Owing to limited practical experience from applying these methods to experimental data, the practical usefulness and limitations of the method in experimental datasets are still largely unknown. The analysis of this dataset serves as an exploration of potential use of the method in a reasonably sized experimental dataset, which was not initially designed for detection of epistasis. The model selection procedures used in this study were based on stringent, population-based genome-wide thresholds in order to control the rate of false-positive epistatic QTL pairs. We felt that this was justified, because this study is one of the first aiming to detect genome-wide epistatic QTLs and it is important to avoid making inferences and recommendations for future studies based on false-positive QTLs. This does limit the power of the study but, when the method has been more thoroughly evaluated, other thresholds can be used to obtain a balance between type I and type II errors that is suitable for each individual experiment.

The first application of the QTL mapping method used in this study was for analyses of growth traits from an exotic cross between the Red Jungle Fowl and a White Leghorn layer (Carlborg *et al*., 2003). The use of the method increased the number of QTLs detected dramatically, and epistasis was shown to be a large contributor to the genetic variance of early growth in that cross. The study described in this report serves two purposes. First, we analysed a set of similar growth traits in a different chicken cross between broiler and layer chickens. By doing this, we hoped to evaluate further the importance of epistasis in chicken growth and to identify genetic mechanisms underlying detected epistasis. Second, this dataset is significantly smaller (466 rather than 752 F_2 individuals) and the results from this study will indicate the potential of the method in the more moderately sized experiments that are common used.

The previous analyses of the growth traits in this dataset (Sewalem *et al*., 2002) were based on an older version of the dataset and different combinations of background QTLs and fixed effects in the models used for analyses. Despite this, a brief comparison of the results from these studies shows that the two studies together report 16 QTLs as significant using a 20% genome-wide significance threshold. Sewalem *et al*. (2003) detected 12 QTLs, one of which was unique to that study, and we here report 15 QTLs, four of which could only be detected using SIM and were unique to this study. Two of these unique QTLs were, however, detected for growth rate at 3–6 and 6–9 weeks of age, which have not previously been analysed.

Both the number and the significance of epistatic QTLs were lower in this cross than in the exotic cross between the Red Jungle Fowl and a White Leghorn. This is expected because this cross has about 300 fewer F_2 individuals and a sparser genetic map (average marker spacing is more than 15 cM greater). There are also fewer unique QTLs detected by the SIM procedure and this could be due to the decrease in power caused by the above reasons, but also to the considerably shorter time since the divergence of the broiler and layer than of the domesticated chicken and its wild ancestor (a few hundred compared with several thousand years). This could influence the opportunities for co-adaptation of genes within the lines that might be one cause of the large amount of epistasis detected in the more exotic cross. However, a significant amount of epistasis was still detected in this study, which implies that epistasis is a rather important mechanism for generation of poultry lines in

general, and that certain favourable allelic combinations occur at high frequencies within the lines. The creation of an experimental cross creates new allelic combinations, which in turn increases the power to detect epistasis. Further studies are needed to evaluate how much epistasis that is segregating within natural chicken populations.

In the Red Jungle Fowl × White Leghorn cross, epistasis was found to be very influential on early growth (8–46 days of age), whereas the importance of epistasis on later growth was low. In the broiler × layer cross, the largest total genetic and epistatic contribution to growth is to the bodyweight at 6 weeks of age (42 days of age) and to growth between 3 weeks and 6 weeks of age (21–42 days). There also seems to be a discrete set of epistatic QTLs involved in earlier growth. This study is therefore consistent with the previous finding that there could be different genetic regulation of early and late growth in chickens, and that epistasis is more important for early than for late growth.

There were 101 unique QTL pairs detected in the Red Jungle Fowl × White Leghorn cross, and of the 21 unique pairs detected in the broiler × layer cross, ten mapped to the same chromosome pairs and six mapped to closely linked marker intervals in the two crosses. When the genotype–phenotype relationships were compared between the crosses for these pairs, the Leghorn alleles for one pair (chromosome 1, 417 cM, and chromosome 13, 7 cM) appeared to have a very similar phenotypic effect in both a broiler and a Red Jungle Fowl background.

The marker spacing in this cross is on average 40 cM and, owing to this, there are relatively large proportions of the genome where the genetic information for detecting a QTL is low. Several QTLs have been located in low information regions both as single QTLs and as part of epistatic QTL pairs. The method used for mapping epistatic QTL pairs is designed to detect significant additional variation explained by an epistatic QTL model. There is no evidence here that would suggest that the additional variation explained for the pairs is due to low information content. On the contrary, there is an indication that the additional QTLs found are in most cases located in more informative regions in the genome than the QTLs detected by their marginal effects. There is furthermore no evidence that segregation distortion is more common in the regions detected using the simultaneous mapping procedure. There is also no evidence that there was a deviation from normality within the QTL genotype classes of the detected epistatic QTL pairs. This observation strengthens the evidence that the method is robust when applied to real data.

Close linkage causes some genotype combinations to be rare, which could cause problems in estimating genetic interactions. Several QTLs were detected on chromosome 1 but only in one pair did the epistatic model fit significantly better than the marginal effects model. In that specific case, the QTLs were located 226 cM apart (chromosome 1, 147 cM, and chromosome 1, 373 cM), which means that they are virtually unlinked. For the QTLs that were located closer than that, no interactions were detected, which could be because either there are no interactions or there is a lack of recombination and hence limited information means that the power is too low to detect interactions.

The genotypic patterns for means of the detected QTL pairs suggested four clusters of pairs with similar patterns of genotype–phenotype expression. The first group contain pairs where several (and in some instances all) homozygote–heterozygote genotype combinations have inferior phenotypes. In the estimates for the two-locus interaction model, this type of genotype pattern becomes apparent by large estimates of the two single-locus dominance and the dominance-by-dominance interaction terms. For example, almost all of the variation for the single epistatic QTL pair for growth at 6–9 weeks of age (chromosome 3, 63 cM, and chromosome 5, 127 cM) is due to the dominance and dominance-by-dominance components. The underlying genetic mechanism for this is unclear but the relatively frequent occurrence of the pattern indicates that there could be some general mechanism that causes this phenomenon.

A second commonly occurring interaction pattern is where the hybrid genotypes (i.e. genotypes that contain at least one heterozygote genotype) have higher phenotypes than both double homozygotes. The broiler genotype has a higher phenotypic effect on growth than the layer genotype in all genotypic combinations. This pattern indicates a pair of QTLs with a heterosis-type interaction. One possible explanation for this could be that the broiler line is fixed for an allele with deleterious effect on growth in homozygous form and that the layer allele is able to complement this allele in the hybrid individuals. Five of the six pairs that exhibit this pattern contain QTLs located on the distal end of chromosome 1 (three pairs with one of the QTLs located around 400 cM and two pairs with one of the QTLs located around 455 cM). This similarity could indicate that these QTLs are the same, even though their estimated locations are more than 50 cM apart.

A third group reflect QTL pairs with a genotype–phenotype pattern with a smooth transition from low phenotypes for Leghorn double homozygotes to high phenotypes for broiler homozygotes. The deviation from additivity and dominance for these pairs is generally due to a non-linear ('multiplicative') rather than a linear ('additive') increase in the phenotypic values with genotype. This pattern indicates a co-adaptation between the alleles at the two loci, where the broiler double homozygote was associated with the highest phenotypic values.

The fourth identified group involves two pairs where there are three levels of phenotypes – high, medium and low. One of the pairs (chromosome 1, 150 cM, and chromosome 18, 13 cM) showed a pattern where a high phenotype is expressed when either or both of the loci contain the broiler homozygote and the other loci contain one broiler allele (BBB– or B–BB). The intermediary phenotype was expressed only by the layer double homozygotes (LLLL), and the rest of the genotypes express a low phenotype. Biologically, this could indicate an inhibitory action on growth by the layer alleles at these loci, unless they are present in the double homozygote (where the inhibition is lower) or it is overridden by homozygote broiler alleles from either locus. The second pair in this group has a similar appearance but is more difficult to interpret genetically.

For some of the QTL pairs it is, however, not possible to cluster or find immediate biological explanations for the patterns of the genotypic means. This could be due to our limited knowledge about the relationships between gene interactions and phenotype. It could also be due to violations of assumptions made in the QTL mapping procedure (e.g. segregation of multiple QTL alleles within the original lines, existence of multiple linked genes in the QTL region or simply poor estimates of the genotypic effects owing to chance or low information content at the genomic location of interest). The results for most QTL pairs will therefore only be an estimate of the importance of epistasis for the combined effects of the two genomic regions and aid in the selection of genotypes for further genetic characterisation of these regions.

By creating a figure joining pairwise interacting QTLs, we obtained a visual representation of the complexity of the genetic network behind the analysed traits. The pairs that were detected or assigned to be epistatic in this study are connected as shown in Fig. 4. The interpretation of this figure is speculation until the true genetic components of each QTL have been identified, but it is possible to suggest alternative interpretations of the figure. It could be viewed as an enzymatic chain (the eight connected horizontal QTLs) in which each step is affected by the result of the enzymatic processes that precede and proceed from that step. The branches represent modulators of the enzymatic chain or provide alternative substrates for the chain. The loops (and especially the loops involving non-linear additive type of epistasis) indicate feedback inhibitors or accelerators of the enzymatic activity. An alternative interpretation is that the QTLs in the centre of the chain involved in most interactions are central to the process of growth (e.g. for deposition of protein or fat). There are then several branches (enzymatic chains) leading to these QTLs and providing substrates for the growth process. The loops could also in this scenario indicate feedback regulation.

A QTL study can be used to find the chromosomal locations that contribute to the variation of the F_2 individuals. It can also be used to predict the genetic effects of individual QTL genotypes. The latter is more difficult because many individuals are needed to draw strong conclusions about the magnitude of the effects. In the Jungle Fowl × Leghorn cross described by Carlborg *et al.* (2003), about half of the detected QTL pairs had epistasis patterns that conformed to previously described mendelian patterns of epistasis (Ö. Carlborg, unpublished results). The evaluations of the effects of the genotypes of the individual QTL pairs in this cross shows that 17 of the 21 unique QTL pairs can be classified into four clusters of similar types of interactions. From this, we conclude that, even though this study was based on a population with a rather low-resolution genetic map and relatively few individuals in each genotype class, the extra effort to map epistatic QTL pairs and inspection of the genotype class made a valuable contribution to interpreting the results.

The method used for mapping of interacting QTLs is based on detection and estimation of epistatic QTL pairs one at the time. Owing to a high computational demand and the small number of individuals in the cross relative to the number of parameters that would need to be estimated, it is not possible simultaneously to fit all QTLs and to estimate their joint effects. Therefore, some of the QTL pairs that are proposed in this article might not be significant if all parameters were fitted jointly. The major aim of this study is, however, not to describe an optimal method for detection of epistatic QTL but rather to highlight genetically interesting findings that deserve to be further evaluated in future generations in this pedigree (e.g. in an advanced intercross line) as well as to indicate the potential benefits of considering epistasis in genome scans for QTLs. If an experiment was designed with the aim of exploring further the importance of epistasis, we recommend that many individuals and a more informative genetic map should be used. Nonetheless, the results from this study show that this method for mapping epistatic QTLs can be valuable for experimental datasets of limited size that are initially not designed for detection of epistasis.

Ö.C. was funded by a fellowship from the Knut and Alice Wallenberg foundation. The research was funded by grants from the Biotechnology and Biological Sciences Research Council, the Department for Environment, Food and Rural Affairs, and the European Union.

References

Boer, M. P., ter Braak, C. J. F. & Jansen, R. C. (2002). A penalized likelihood method for mapping epistatic

quantitative trait loci with one-dimensional genome searches. *Genetics* **162**, 951–960.
Carlborg, Ö. (2002). New methods for mapping quantitative trait loci. PhD thesis, Acta Universitatis Agriculturae Sueciae. Veterinaria 121. Swedish University of Agricultural Sciences.
Carlborg, Ö., Andersson, L. & Kinghorn, B. (2000). The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics* **155**, 2003–2010.
Carlborg, Ö., Andersson-Eklund, L. & Andersson, L. (2001). Parallel computing in interval mapping of quantitative trait loci. *Journal of Heredity* **92**, 449–451.
Carlborg, Ö. & Andersson, L. (2002). The use of randomization testing for detection of multiple epistatic QTL. *Genetical Research* **79**, 175–184.
Carlborg, Ö., Kerje, S., Schutz, K., Jacobsson, L., Jensen, P. & Andersson, L. (2003). A global search reveals epistatic interaction between QTLs for early growth in the chicken. *Genome Research* **13**, 413–421.
Chase, K., Adler, F. R. & Lark, K. G. (1997). Epistat: a computer program for identifying and testing interactions between pairs of quantitative trait loci. *Theoretical and Applied Genetics* **94**, 724–730.
Cheverud, J. M. & Routman, E. J. (1995). Epistasis and its contribution to genetic variance components. *Genetics* **139**, 1455–1461.
Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.
Dunnington, E. A., Haberfeld, A., Stallard, L. C., Siegel, P. B. & Hillel, J. (1992). Deoxyribonucleic acid fingerprint bands linked to loci coding for quantitative traits in chickens. *Poultry Science* **71**, 1251–1258.
Fijneman, R. J., De Vries, S. S., Jansen, R. C. & Dermant, P. (1996). Complex interactions of new quantitative trait loci, *Sluc1*, *Sluc2*, *Sluc3*, and *Sluc4*, that influence the susceptibility to lung cancer in the mouse. *Nature Genetics* **14**, 465–467.
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.
Haley, C. S., Knott, S. A. & Elsen, J. M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195–1207.
Ikeobi, C. O., Woolliams, J. A., Morrice, D. R., Law, A., Windsor, D., Burt, D. W. & Hocking, P. M. (2002). Quantitative trait loci affecting fatness in the chicken. *Animal Genetics* **33**, 428–435.
Jannink, J. L. & Jansen, R. C. (2001). Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* **157**, 445–454.
Kao, C.-H., Zeng, Z.-B. & Teasdale, R. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.
Leamy, L. J., Routman, E. J. & Cheverud, J. M. (2002). An epistatic genetic basis for fluctuating asymmetry of mandible size in mice. *Evolution; International Journal of Organic Evolution* **56**, 642–653.
Levine, D. (1996). *Users Guide to the PGAPack Parallel Genetic Algorithm Library*. Mathematics and Computer Science Division, Argonne National Laboratory, IL, USA.
Ljungberg, K., Holmgren, S. & Carlborg, Ö. (2002). Efficient algorithms for quantitative trait loci mapping problems. *Journal of Computational Biology* **9**, 793–804.
Peripato, A. C., De Brito, R. A., Vaughn, T. T., Pletscher, L. S., Matioli, S. R. & Cheverud, J. M. (2002). Quantitative trait loci for maternal performance for offspring survival in mice. *Genetics* **162**, 1341–1353.
Sen, S. & Churchill, G. A. (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**, 371–387.
Sewalem, A., Morrice, D. M., Law, A., Windson, D., Haley, C. S., Ikeobi, O. N., Burt, D. W. & Hocking, P. M. (2002). Mapping of quantitative trait loci for body weight at three, six and nine weeks of age in a broiler layer cross. *Poultry Science* **81**, 1775–1781.
Shimomura, K., Low-Zeddues, S. S., King, D. P., Steeves, T. D. L., Whiteley, A., Kushla, J., Zemenides, P. D., Lin, A., Vitaterna, M. H., Churchill, G. A. & Takahashi, J. S. (2001). Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Research* **11**, 959–980.
van Kaam, J. B., van Arendonk, J. A., Groenen, M. A., Bovenhuis, H., Vereijken, A. L., Croijmans, J. J., van der Poel, J. J. & Veenendaal, A. (1998). Whole genome scan for quantitative trait loci affecting body weight in chickens using a three generation design. *Livestock Production Science* **54**, 133–150.
van Kaam, J. B., Groenen, M. A., Bovenhuis, H., Veenendaal, A., Vereijken, A. L. & van Arendonk, J. A. (1999). Whole genome scan in chickens for quantitative trait loci affecting growth and feed efficiency. *Poultry Science* **78**, 15–23.
Zeng, Z.-B., Liu, J., Stam, L. F., Kao, Z.-H., Mercer, J. M. & Laurie, C. C. (2000). Genetic architecture of a morphological shape difference between two *Drosophila* species. *Genetics* **154**, 299–310.

Genet. Res., Camb. (2004), **83**, *pp.* 211–220. *With 2 figures.* © 2004 Cambridge University Press
DOI: 10.1017/S0016672304006846 Printed in the United Kingdom

Segregation of QTL for production traits in commercial meat-type chickens

D. J. DE KONING[1]*, C. S. HALEY[1], D. WINDSOR[1], P. M. HOCKING[1], H. GRIFFIN[1], A. MORRIS[2]†, J. VINCENT[2] AND D. W. BURT[1]
[1] *Roslin Institute, Roslin, Midlothian EH25 9PS, UK*
[2] *The Cobb Breeding Company Ltd, Chelmsford, Essex CM3 8BY, UK*

(*Received 18 August 2003 and in revised form 30 January 2004*)

Summary

This study investigated whether quantitative trait loci (QTL) identified in experimental crosses of chickens provide a short cut to the identification of QTL in commercial populations. A commercial population of broilers was targeted for chromosomal regions in which QTL for traits associated with meat production have previously been detected in extreme crosses. A three-generation design, consisting of 15 grandsires, 608 half-sib hens and over 15 000 third-generation offspring, was implemented within the existing breeding scheme of a broiler breeding company. The first two generations were typed for 52 microsatellite markers spanning regions of nine chicken chromosomes and covering a total of 730 cM, approximately one-fifth of the chicken genome. Using half-sib analyses with a multiple QTL model, linkage was studied between these regions and 17 growth and carcass traits. Out of 153 trait × region comparisons, 53 QTL exceeded the threshold for genome-wide significance while an additional 23 QTL were significant at the nominal 1% level. Many of the QTL affect the carcass proportions and feed intake, for which there are few published studies. Given intensive selection for efficient growth in broilers for more than 50 generations it is surprising that many QTL affecting these traits are still segregating. Future fine-mapping efforts could elucidate whether ancestral mutations are still segregating as a result of pleiotropic effects on fitness traits or whether this variation is due to new mutations.

1. Introduction

In chicken, as in other species, crosses between extreme lines have been used to detect quantitative trait loci (QTL) that explain phenotypic differences between the lines. These experimental populations include crosses between native jungle fowl and White Leghorn (Carlborg *et al*., 2003), broiler and White Leghorn (Sewalem *et al*., 2002) and two extreme broiler lines (Van Kaam *et al*., 1998). This approach has proved very successful in identifying QTL that explain differences between these lines, but they provide no insight as to whether these QTL are segregating within current commercial lines that have been selected for at least 50 generations. Indeed, following more than 50 generations of selection for efficient growth, it is expected that loci with major effects on growth will be fixed for the high-growth alleles within the broiler lines, unless there are other mechanisms that maintain variation at these loci. Hence, most of the extreme crosses have been analysed under the assumption that the founder breeds are completely fixed for alternative QTL alleles (Haley *et al*., 1994). However, for successful implementation of marker-assisted selection within a population, segregation of QTL needs to be verified within the commercial lines. Confirmation of QTL within a commercial line is only realistic using the existing family structure and data recording of the breeding population and requires different study designs and statistical analyses compared with line-cross experiments. Following the preliminary results

* Corresponding author. Tel: +44 131 5274258. Fax: +44 131 4400434. e-mail: DJ.deKoning@BBSRC.AC.UK
† Present address: British United Turkeys, Hockenhull Hall, Tarvin, Chester, CH3 8LE.

Table 1. *Trait means and genetic parameters for 13 traits in a commercial broiler breeding population*

Trait	Mean[a]	SD	$h^2 \pm$ SE	Maternal effect[b]	Average reliability[c]
Body weight 40 days, g	2415	276	0·11 ± 0·01	0·02/0·01	0·30
Feed conversion during test	1·82	0·31	0·07 ± 0·01	–	0·10
Residual feed intake during test, g	1042	223	0·11 ± 0·02	0·02	0·16
Conformation score	3·35	0·88	0·23 ± 0·02	0·01	0·43
Dissection weight at 41 days, g	2291	268	0·10 ± 0·03	0·04	0·18
Abdominal fat weight, g	28	10	0·00 ± 0·01	–	–
Breast muscle weight, g	450	67	0·43 ± 0·04	–	0·33
Thighbone weight, g	20	4·5	0·06 ± 0·02	–	0·10
Thigh muscle weight, g	92	13	0·10 ± 0·03	0·02	0·16
Thigh meat to bone ratio	4·8	1·1	0·10 ± 0·02	–	0·10
Drumbone weight, g	33	7·3	0·07 ± 0·02	–	0·12
Drum muscle weight, g	76	13	0·16 ± 0·03	–	0·21
Drum meat to bone ratio	2·4	0·7	0·04 ± 0·02	–	0·08

[a] Raw phenotypic means.
[b] Proportion of total variance explained by the direct maternal effect. Second value is for the maternal genetic effect.
[c] Expected fraction of additive genetic variance explained by breeding values (EBV).
– Indicates that the direct maternal effect was not significant.

for a region on chicken chromosome 4 (De Koning *et al*., 2003), we have tested eight additional candidate regions for which QTL have been reported in extreme crosses on a commercial broiler line.

2. Material and methods

(i) *Experimental population and phenotypic traits*

Following power calculations (De Koning *et al*., 2003), 15 males of a broiler dam line (The Cobb Breeding Company Ltd, Chelmsford, UK) were selected as grandsires in a three-generation half-sib design. Blood samples were collected on the grandsires (G1), their mates (104) and 608 second-generation (G2) hens. For 80 hens only their own observations for body weight, conformation and test data were available, leaving 524 G2 hens with phenotypic data on at least one offspring with an average family size of 35. On the offspring of these hens, the third generation (G3), only phenotypic information was gathered. Traits that are routinely measured on all birds included body weight at 40 days and conformation score. Prior to selection, a proportion of the birds were randomly selected for carcass dissection to allow sufficient numbers for QTL analysis. Following truncation selection on body weight, a proportion (~20%) of birds was tested for 2 weeks for feed consumption and growth, while the remaining birds were culled at 40 days of age. Thirteen traits were derived from the observations (Table 1). For body weight and conformation score observations were available on >50 000 birds (15 000 G3 offspring and their contemporaries) with an average of 28 offspring for every G2 hen. For nine carcass-related traits, an average of nine offspring phenotypes were available for each of 477 G2 hens. For the feed intake and growth data during a 2 week test, an average of five offspring was available for 440 G2 hens. Following exploratory analyses using GENSTAT (Lawes Agricultural Trust, Harpenden, UK), variance components were estimated using ASREML (Gilmour *et al*., 2000). The initial model included the fixed effects (sex, hatch within flock, and age of dam for all traits), covariates (body weight for all carcass proportions, mid-weight and growth during test for feed efficiency traits) as well as a random polygenic component. The initial model included all the fixed effects and covariates as well as a random polygenic component:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \quad (1)$$

where **y** is a vector of phenotypes, **b** is a vector of fixed effects and covariates, **u** is a vector of random direct polygenic effects (estimated breeding values: EBV) and **e** is a vector of residuals. **X** is an incidence matrix relating fixed effects and covariates to observations and **Z** is an incidence matrix relating observations to random direct polygenic effects. Subsequently a direct maternal effect was added to the model and tested against a polygenic model with a likelihood ratio test.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Vc} + \mathbf{e}. \quad (2)$$

Variables are as in (1) with the addition of **c**, a vector of random direct maternal effects and **V**, an incidence matrix relating direct maternal effects to observations. When the direct maternal effect was significant the model was extended with a genetic maternal

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

Table 2. *Candidate regions from four experimental crosses*

Chromosome	Marker interval (positions)[a]	QTL in experimental crosses[b]
1	MCW0011–MCW0112 (98–205)	Body weight[1,3,4], feed intake[1], thigh yield[3]
3	ADL0237–MCW0037 (275–317)	Body weight[4], fatness[3]
4	ADL0241–LEI0076 (80–182)	Body weight[2,3,4], feed intake[1,2]
5	MCW0090–MCW0032 (57–128)	Body weight[4], fatness[3], lean-to-bone ratio[3]
7	LEI0064–MCW0236 (0–109)	Body weight[3,4], leg yield[3], fatness[3], lean-to-bone ratio[3]
8	ROS0026–MCW0100 (14–46)	Body weight[3,4], breast yield[3]
9	ROS0078–MCW0135 (0–57)	Body weight[4], fatness[3], lean-to-bone ratio[3]
11	LEI0110–ROS0112 (18–88)	Body weight[4]
13	MCW0213–ADL0214 (22–74)	Body weight[3,4], fatness[3], leg yield[3], lean-to-bone ratio[3]

[a] Positions on consensus linkage map in Schmid *et al*. (2000).
[b] Restricted to traits that resemble those in the present study. Superscripts indicate in which cross the QTL was detected: [1] Wageningen University extreme broiler cross (Van Kaam *et al*., 1998, 1999*a*, *b*); [2] Agrifood Research Finland extreme layer cross (Tuiskula-Haavisto *et al*., 2002); [3] Roslin Institute broiler × layer cross (Ikeobi *et al*., 2002; Sewalem *et al*., 2002; Ikeobi *et al*., 2004); [4] Uppsala Red Jungle Fowl × White Leghorn cross (Carlborg *et al*., 2003).

component and its significance evaluated with a likelihood ratio test against model (2):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wd} + \mathbf{Vc} + \mathbf{e}, \quad (3)$$

where **d** is a vector of random maternal genetic effects and **W** is an incidence matrix relating maternal genetic effects to observations. For the QTL analyses, trait scores for the G2 dams were derived from the EBV of the G2 hens, adjusted for information coming from other relatives besides their offspring by deducting the mean of the parental EBV of each hen. An alternative to using adjusted EBV is to calculate offspring yield deviations (OYD) as was done by Van Kaam *et al*. (1998) and in our previous work (De Koning *et al*., 2003). However, the adjustment of the EBV is more straightforward than obtaining the OYD, especially because the EBV of the G2 sires (mated to our G2 hens) may be biased when they were mated only to a single or few hens. Furthermore, Dolezal *et al*. (2003) showed that adjusted EBV and OYD are very closely correlated. To account for different numbers of offspring between G2 hens, the reliability of the EBV was used as a weighting factor in the QTL analyses (Table 1). The estimation of direct maternal (MD) and the maternal genetic (MG) effects used additional information compared with the EBVs for traits that were also measured on the G2 hens. Therefore, the estimated maternal effects from ASREML for body weight (direct and genetic), conformation score and residual feed intake were included as four additional traits in the QTL analyses. For a more detailed description of the phenotypes and the derivation of QTL trait scores see De Koning *et al*. (2003).

The adjusted EBVs for all traits were analysed jointly by GENSTAT to obtain estimates of the correlations between trait scores. A principal component analysis was also performed to assess the number of independent traits among the 17 traits that were analysed.

(ii) *Genotyping and map construction*

Nine regions on chicken chromosomes 1, 3, 4, 5, 7, 8, 9, 11 and 13 were selected because they showed evidence for body-weight-related QTL in one or more genome scans. Marker coverage of these chromosomal regions and the QTL identified in these regions in four experimental populations are summarized in Table 2. Microsatellite markers in the candidate regions were selected from the consensus linkage map (Schmid *et al*., 2000) and tested for heterozygosity in the 15 grandsires. Genotypes were obtained on the G1 and G2 animals for 52 microsatellite markers with three to ten markers per candidate region. Details on PCR amplification and gel electrophoresis are given by Sewalem *et al*. (2002). Marker distances were estimated using the 'build' option of Crimap (Green *et al*., 1990), subsequently using the 'flips' option to evaluate alternative marker orders compared to the marker order of the consensus map.

(iii) *QTL analysis*

The methodology is based on the half-sib analyses proposed by Knott *et al*. (1996). Exploratory QTL analyses were performed using the QTL Express

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

software at http://qtl.cap.ed.ac.uk/ (Seaton *et al.*, 2002), followed by analyses under a multiple QTL model using a modification of the methodology proposed by De Koning *et al.* (2001). In the first step of the multiple QTL analyses, the candidate regions are analysed individually fitting a single QTL within every family:

$$Y_{ij} = \mu_i + b_i X_{ij} + e_{ij}, \quad (4)$$

where Y_{ij} is the phenotype of j, offspring of sire i, μ_i is the mean of sire family i, b_i the allele substitution effect of the QTL within family i, X_{ij} the probability that animal j inherited the (arbitrarily assigned) first haplotype of sire i, and e_{ij} is the residual effect. In the second step, the best positions on every chromosome that exceeded a point-wise 5% threshold were identified and all the regions were re-analysed with the QTL that were on all other chromosomes as cofactors:

$$Y_{ij} = \mu_i + b_i X_{ij} + \sum_{k=1}^{n} b_{ik} X_{ijk} + e_{ij}, \quad (5)$$

where variables are identical to (1), except for the term $\sum_{k=1}^{n} b_{ik} X_{ijk}$, which describes the multiple regression of the n cofactors that are on chromosomes other than the one under study. If this analysis revealed additional putative QTL, or the best positions of the QTL change, the selection of cofactors was modified and the regions were re-analysed. This step was repeated until no new QTL were identified or dropped from the model, and the positions of the QTL were stable. The difference between this analysis and that of De Koning *et al.* (2001) is that in the present study the cofactors were maintained in the model continuously, while De Koning *et al.* (2001) adjusted the trait scores for cofactor effects prior to re-analysing the chromosomes. The proportion of within-family variance explained by each QTL (h^2_{QTL}) was approximated following Knott *et al.* (1996):

$$h^2_{QTL} = 4*[1-(MSE_{full}/MSE_{reduced})], \quad (6)$$

where MSE_{full} is the mean squared error of the model including the QTL (4) and $MSE_{reduced}$ is the mean squared error of the model fitting only a family mean. For comparison, we also estimated the proportion of variance explained (r^2) by the joint QTL and cofactors. Empirical thresholds were obtained using permutation tests (Churchill & Doerge, 1994). Marker genotypes for the region under study were permuted within half-sib families, while the phenotypes and the genotype scores for the cofactors were maintained. Note that this provides an empirical test for the region under study, not for the cofactors, but the significance of every cofactor was tested when its region was re-analysed. For significance testing we imposed two thresholds: (1) Following the recommendations of Lander & Kruglyak (1995) we used an empirical point-wise threshold (not accounting for multiple testing) of $P<0{\cdot}01$ to claim confirmed linkage when a QTL for a given trait had already been reported for a certain region. (2) Because each region represented on average 1/50 of the chicken genome, we imposed an empirical 'region-wise' threshold (accounting for multiple tests on part of a linkage group) of $P<0{\cdot}001$ to claim genome-wide significant linkage (Lander & Kruglyak, 1995).

It is not trivial to determine which QTL are confirming published QTL and which QTL are 'new'. Trait definitions vary between studies and some traits are measured in only a single study. Published studies use different molecular markers, further compromising any comparison of QTL positions. This is no problem for genome-wide significant QTL because they do not rely on published results for interpretation of their significance. Accounting for the imprecision of QTL detection, we used a maximum distance of 30 cM from a published QTL to infer whether that QTL had been confirmed in the present study.

3. Results

(i) *Trait heritabilities and correlations*

Heritabilities were low to moderate (Table 1) and significant direct maternal effects were detected for body weight, residual feed intake, conformation, dissection weight and thigh muscle weight. For body weight, the maternal genetic effect was also significant.

Many of the traits were closely correlated and a principal component analysis on all adjusted breeding values showed that five independent vectors explained 99% of all the variation in the 17 traits. The principal component vector loadings and the partial correlations between the EBVs show that conformation, bodyweight and dissection weight grouped together with correlations ranging from 0·30 to 0·96. Residual feed intake and feed conversion ratio were a separate group with a correlation of 0·64. Correlations of the feed intake traits with the other traits were all within −0·10 to 0·10, with the exception of feed conversion ratio and dissection weight (0·17). The thigh and drum proportion traits were at least moderately correlated with the absolute correlation varying between 0·20 and 0·81.

(ii) *QTL analyses*

The multiple QTL analyses found 53 genome-wide significant QTL, varying from a single genome-wide significant QTL for body weight and conformation score up to six genome-wide significant QTL for

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

residual feed intake and thighbone weight. Twenty-three additional putative QTL exceeded the threshold for confirmed linkage. An overview of all these QTL and a comparison with published QTL is given in Fig. 1. Seventeen genome-wide significant QTL map to regions where similar QTL have been published (Fig. 1). From the 23 QTL exceeding the threshold for confirmed linkage, 10 map within 30 cM of published QTL for a similar or identical trait, while the remaining 13 putative QTL have to be classified as suggestive new QTL because they do not map to a published QTL. Fig. 1 also shows that the QTL appear clustered rather than uniformly distributed across the candidate regions. Many of these QTL clusters may reflect pleiotropic action of a single QTL, with the actual number of genome-wide significant QTL between 9 and 53. Although Schrooten & Bovenhuis (2002) propose a method to identify pleiotropic effects of QTL in a half-sib design there is at present no multi-trait software available to distinguish between linked and pleiotropic QTL in half-sib designs.

(iii) *QTL effects*

The approximate proportions of within-family variance explained by the QTL (h^2_{QTL}) are summarized in Table 3 and range between 0·04 and 0·26 for the genome-wide significant QTL. Summed together, the QTL and QTL used as cofactors have r^2 (Table 3) between 0·16 (body weight) and 0·52 (direct maternal effect for residual feed intake). Multiplying the r^2 by 4 to approximate the within-family variance explained by the joint QTL would give very unrealistic values, thus illustrating that the variances explained by the joint QTL are overestimated. Hayes & Goddard (2001) quantified the level of upward bias using empirical pig and dairy cattle data and the present results agree with their trend. The total overestimation of the QTL variances increases with the number of QTL that are detected.

To evaluate the proportion of the additive genetic variance that is explained by the joint QTL it is important to note that the trait scores are EBV that would explain all additive genetic variance (i.e. have a 'heritability' of 1·0) if there were an infinite number of offspring. The reliability of the EBV, also defined as the squared correlation between the estimated and true EBV, is an indicator of the proportion of additive genetic variance explained by the EBV. The average reliabilities vary between ~0·1 for the thigh and drum traits and 0·4 for conformation score, clearly reflecting the effect of the estimated heritability (Table 1) on the reliability. Although the QTL explain up to half of the variance in adjusted EBV (Table 3), this only accounts for a small part of the additive genetic variance because of the low to modest reliabilities of the EBV (Table 1).

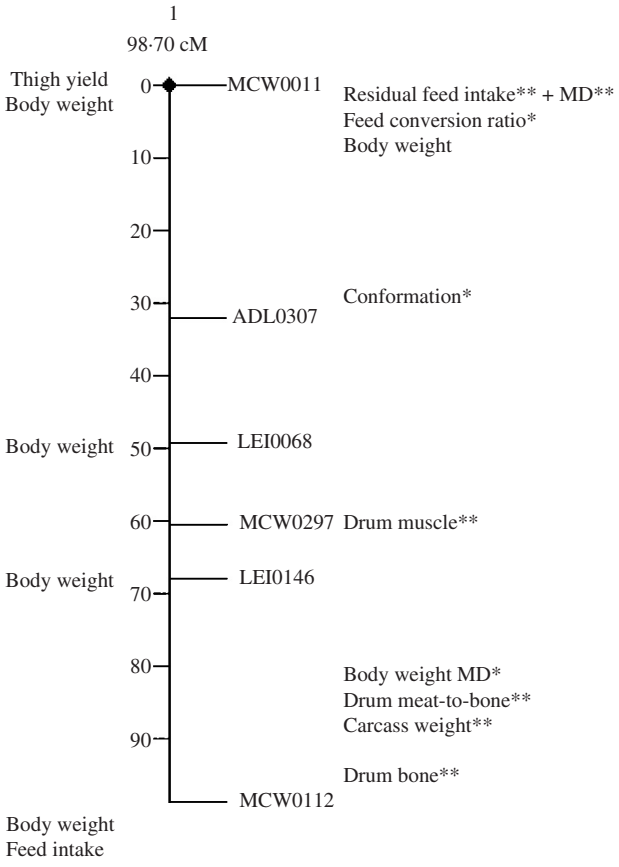
4. Discussion

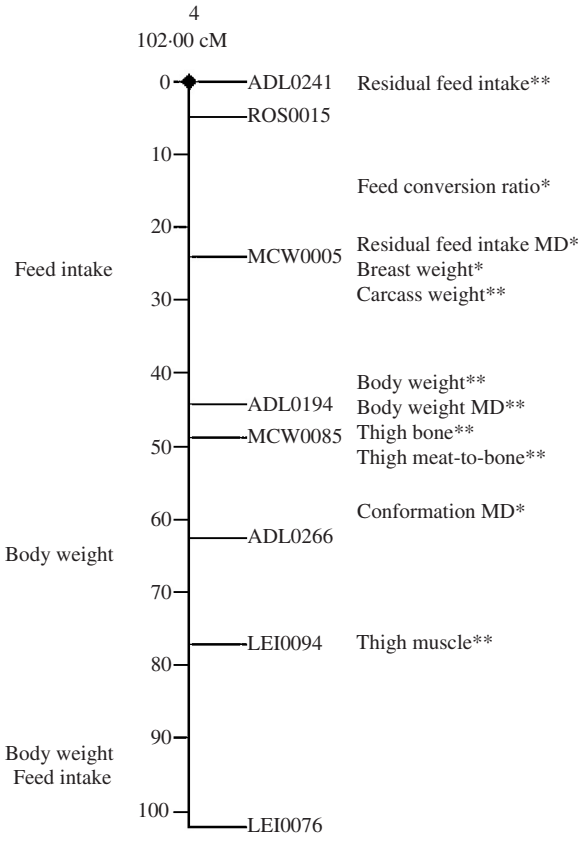
(i) *Multiple QTL analysis*

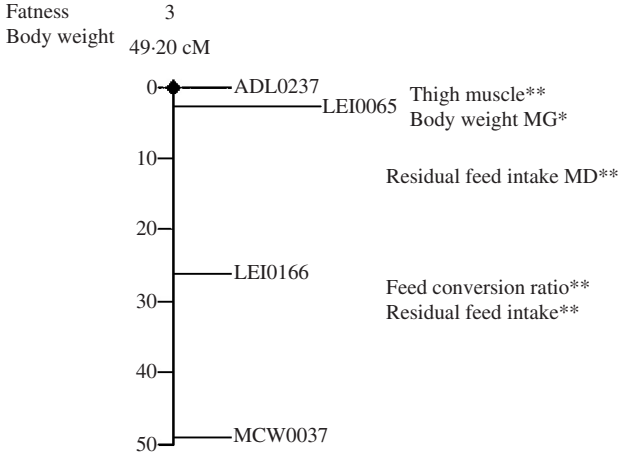
The number of genome-wide significant QTL (53) is very high compared with published studies of poultry QTL, even accounting for the fact that many of the QTL are counted more than once because they affect multiple traits (Fig. 1, Table 3). The only comparison with a family based experimental design is offered by the studies of Van Kaam *et al.* (1998, 1999*a*, *b*) who identified only four genome-wide QTL. They used a cross between two different broiler strains that is expected to be segregating for more QTL than the present study of a single closed population. One possible explanation for this discrepancy could be the use of cofactors to account for unlinked QTL in the present analyses. Using only single QTL analyses we detected 24 instead of 53 genome-wide significant QTL. De Koning *et al.* (2001) first introduced the use of cofactors for the analyses of half-sib designs in dairy cattle. Despite the apparent effectiveness of this approach and its relatively straightforward implementation, it has not been widely used in the analyses of experimental data except for the population where it was first implemented (Viitala *et al.*, 2003). The present results use a refined approach of the cofactor analysis where cofactors are continuously included in the analyses rather than adjusting the phenotypic data for the cofactor effects (De Koning *et al.*, 2001). Fig. 2 shows the effect of multiple QTL analyses on the test statistic along chromosome 1 for two traits. The main effect of using cofactors is the reduction in the residual variance leading to a higher test statistic.

(ii) *Segregation of QTL within a selected line*

The selection line for this experiment is a broiler-dam line with about 50 000 contemporaries at any given time in overlapping generations. Although all birds are potential selection candidates the effective population size is much smaller, which is exemplified by the present experiment where 15 grandsires give rise to approximately one-third of the animals in the G3. The initial selection of candidate parents is based on body weight at 6 weeks of age and conformation score. The selected birds are then entered into a 2 week feed efficiency trial, while a proportion of unselected relatives is dissected to provide carcass measurements. From the 53 genome-wide QTL, 21 are for traits for which selection is applied directly on the selection candidates (body weight, feed intake, conformation) and 32 for carcass-related traits that have been measured on relatives of the selection candidates. For many decades selection has been mainly been on juvenile growth and conformation. This may be reflected in the present results because we find the least number







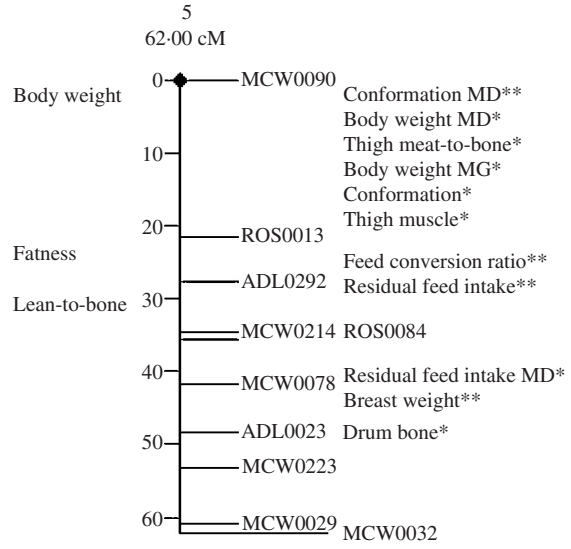


Fig. 1. (*Cont.*)

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

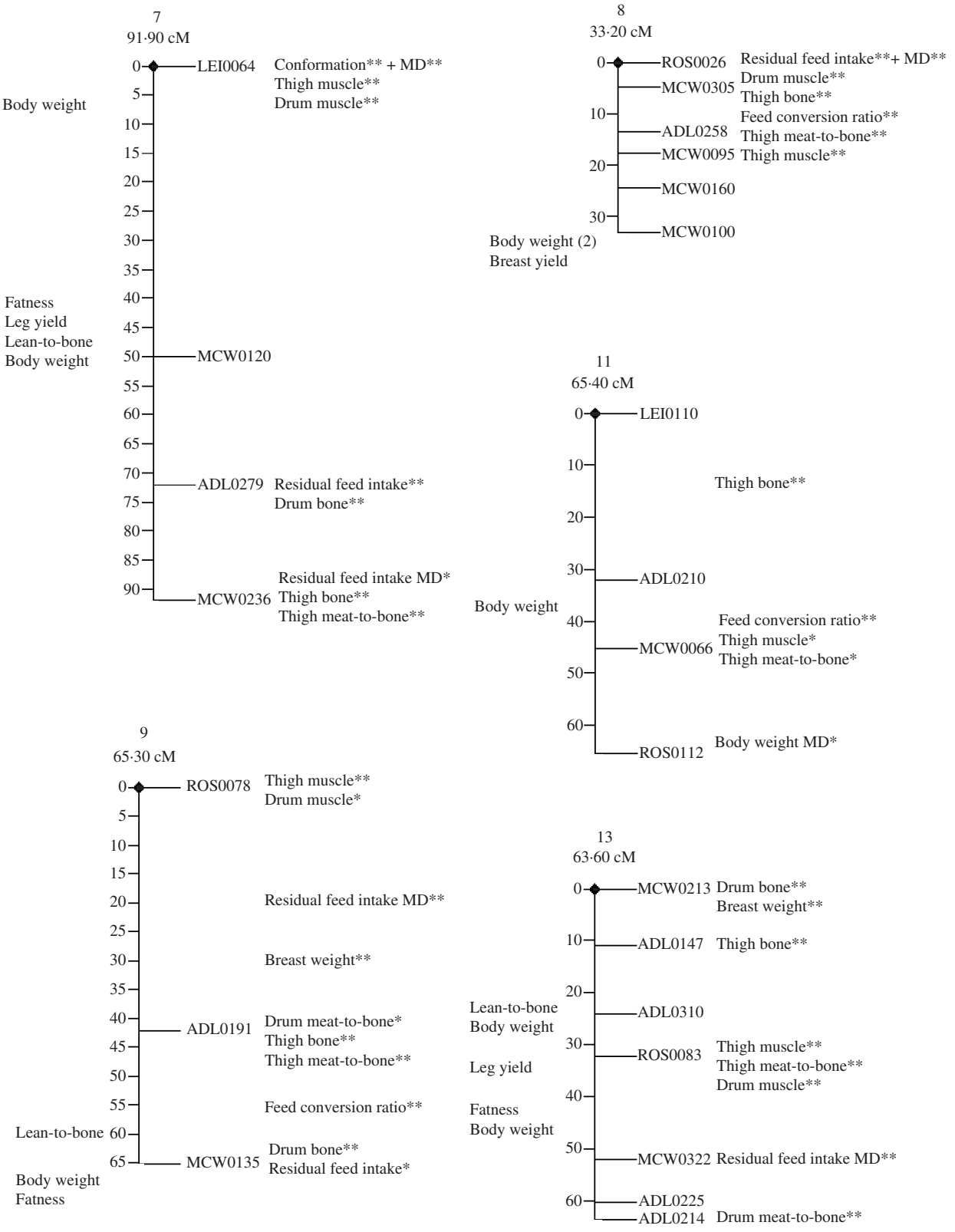


Fig. 1. Overview of poultry QTL in nine candidate regions in four experimental crosses and the present study. The marker maps are in Kosambi cM. Trait names on the left of the maps indicate approximate locations of QTL in the four experimental crosses while trait names on the right indicate approximate locations of QTL in the present study. Significance exceeding the threshold for confirmed linkage (*); genome-wide significant linkage (**). MD and MG denote, respectively, the direct maternal and the maternal genetic effects of the trait.

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

Table 3. *Approximate proportion of within-family variance* (h^2_{QTL}) *explained by QTL in nine candidate regions and the proportion of EBV variance* (r^2) *explained by joint QTL and cofactors*

Trait[a]	GGA1	GGA3	GGA4	GGA5	GGA7	GGA8	GGA9	GGA11	GGA13	r^2 joint cofactors and QTL
Body weight	0·07*		0·24**					C		0·16
MD	0·06*		0·16**	0·08*				0·08*		0·21
MG	C[b]	0·09*		0·12*			C			0·24
Feed conversion	0·09*	0·18**	0·09*	0·14**		0·14**	0·18**	0·18**		0·40
Residual feed intake	0·14**	0·10**	0·16**	0·15**	0·10*	0·11**	0·06*	C		0·41
MD	0·05**	0·04**	0·03*	0·05*	0·02*	0·08**	0·18**	C	0·06**	0·52
Conformation score	C[b]			0·11*	0·20**					0·15
MD	0·10*		0·10*	0·19**	0·20**					0·23
Dissection weight	0·24**		0·23**							0·13
Breast yield			0·04*	0·11**			0·20**		0·13**	0·26
Thighbone			0·12**		0·17**	0·10**	0·10**	0·10**	0·10**	0·38
Thigh muscle		0·15**	0·16**	0·03*	0·10**	0·10**	0·05**	0·05*	0·08**	0·47
Thigh meat to bone ratio			0·05**	0·04*	0·07**	0·09**	0·16**	0·04*	0·09**	0·39
Drumbone	0·30**			0·05*	0·14**		0·10**		0·26**	0·32
Drum muscle	0·18**				0·13**	0·15**	0·07*		0·21**	0·29
Drum meat to bone ratio	0·20**						0·08*		0·20**	0·20

* Denotes significance at the empirical $P<0·01$ (confirmed linkage) and ** denotes significance at the empirical region-wide $P<0·001$ ($\sim$genome-wide significant).
[a] MD and MG denote respectively, the direct maternal and the maternal genetic effect of the preceding trait.
[b] C indicates that the best position was included as a cofactor although this region was not significant.

of QTL for body weight and conformation. Breeding objectives have changed over time to include feed efficiency and breast yield, for which we find large numbers of QTL. Selection on carcass proportions is expected to be less effective because it is based on information coming from relatives. As more traits are combined in the selection index, the total efficiency of selection for any given trait will decrease. Although the development of broiler breeding over time may offer some explanations, it is nevertheless surprising that so many QTL with moderate to large effect are still segregating within this line. It is even more surprising that many of these QTL map to regions that explain phenotypic differences between broilers, layers and their wild progenitor. Furthermore the number of detected QTL suggests that the present design is at least as powerful as a moderately sized F2 design for the detection of QTL. This raises questions as to whether there is just as much variation within chicken lines as there is between lines, and whether the same loci or even the same alleles might be involved.

The large population size would certainly contribute to maintain considerable genetic variation by preventing fixation of alleles by drift and/or inbreeding. However, for QTL with moderate to large effects to be present it could be hypothesized that considerable mutation variance should have contributed (Keightley & Hill, 1987). If there were new mutations giving rise to many of the detected QTL it is not obvious why they would map to the same regions as QTL explaining differences between broilers and layers. However, no firm conclusions can be drawn because we do not know how many QTL are segregating outside the candidate regions nor whether the QTL that map to similar regions as published studies represent the same functional mutation. Furthermore, we do not know whether the QTL represent single Mendelian loci or complexes of multiple linked effects. Fine mapping efforts in both the commercial line and the experimental crosses would reveal conserved haplotypes around the mutation(s) that give rise to the QTL in each population. If these haplotypes are identical in the crosses and the commercial lines this points to the same mutation while different haplotypes and/or QTL locations point to independent mutations in different populations.

(iii) *Conclusions*

The use of nine candidate regions from experimental crosses to target a commercial line has proved very powerful. By typing only approximately 20% of the chicken genome we detected QTL explaining between 14% and 50% of the variation in the analysed traits, although this is most likely an inflated estimate. The detection of many QTL within a selection line is the first step to the implementation of marker-assisted selection within this line. With the present knowledge this would require large amounts of genotyping and analyses because the QTL effects have to be estimated within every family. However, if these QTL can be fine-mapped to the level of a functional haplotype by

https://doi.org/10.1017/S0016672304006834 Published online by Cambridge University Press

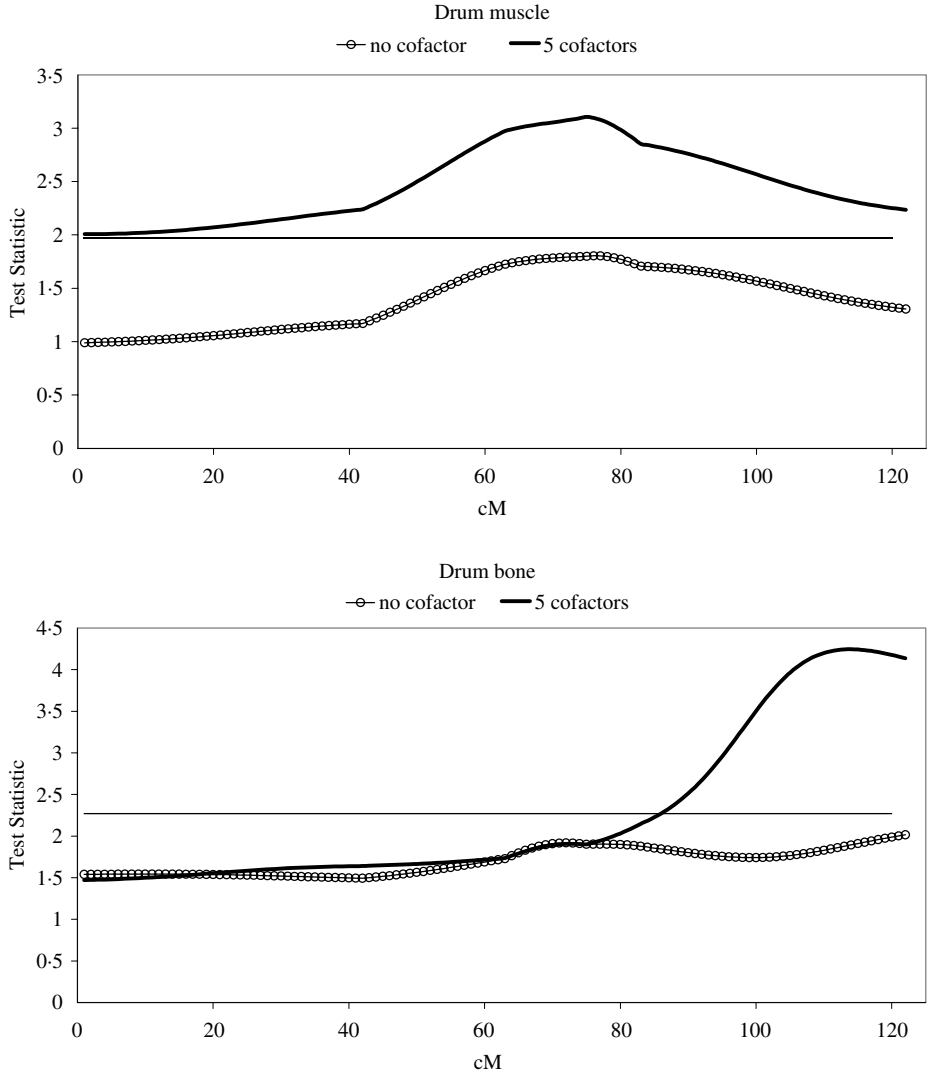


Fig. 2. Effect of cofactors on the test statistic along chromosome 1 for two traits. The horizontal line denotes the approximate threshold for genome-wide significance under the multiple QTL model.

using across-family haplotype comparison (Riquet *et al*., 1999), they could be used for direct association and selection at the population level.

Our results inspire some interesting hypotheses about variation within versus between lines and whether the same loci could be involved. The present results lack precision of QTL positions and information about QTL on the remaining chromosomes that would be required to draw any firm conclusions, but clearly point to commercial populations as a valuable addition to experimental crosses for the location of QTL that affect performance traits.

This project was financially supported by the Department of Environment, Food, and Rural Affairs of the UK government through LINK project LK0625. We gratefully acknowledge The Cobb Breeding Company Ltd for additional financial support and data collection. We acknowledge referees' comments, which increased the clarity of the manuscript.

References

Carlborg, Ö., Kerje, S., Schütz, K., Jacobsson, L., Jensen, P. & Andersson, L. (2003). A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Research* **13**, 413–421.

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

De Koning, D. J., Schulman, N. F., Elo, K., Moisio, S., Kinos, R., Vilkki, J. & Mäki-Tanila, A. (2001). Mapping of multiple quantitative trait loci by simple regression in half-sib designs. *Journal of Animal Science* **79**, 616–622.

De Koning, D. J., Windsor, D., Hocking, P. M., Burt, D. W., Law, A., Haley, C. S., Morris, A., Vincent, J. & Griffin, H. (2003). Quantitative locus detection in commercial broiler lines using candidate regions. *Journal of Animal Science* **81**, 1158–1165.

Dolezal, M., Schwarzenbacker, H., Soelkner, J. & Fürst, C. (2003). Analysis of different selection criteria for selective DNA pooling. Book of abstracts of the 54th meeting of the EAAP, Rome, p. 3.

Gilmour, A. R., Cullis, B. R., Welham, S. J. & Thompson, R. (2000). *ASREML Reference Manual*. University of New South Wales, Orange, Australia.

Green, P., Falls, K. & Crooks, S. (1990). *Documentation for CRI-MAP*, version 2.4. St Louis, MO: Washington School of Medicine.

Haley, C. S., Knott, S. A. & Elsen, J. M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195–1207.

Hayes, B. & Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics, Selection, Evolution* **33**, 209–229.

Ikeobi, C. O. N., Woolliams, J. A., Morrice, D. R., Law, A., Windsor, D., Burt, D. W. & Hocking, P. M. (2002). Quantitative trait loci affecting fatness in the chicken. *Animal Genetics* **33**, 428–435.

Ikeobi, C. O. N., Woolliams, J. A., Morrice, D. R., Law, A. S., Windsor, D., Burt, D. W. & Hocking, P. M. (2004). Quantitative trait loci for muscling in a broiler layer cross. *Livestock Production Science*, in press. Accessible online (doi:10.1016/j.livprodsci.2003.09.020)

Keightley, P. D. & Hill, W. G. (1987). Directional selection and variation in finite population. *Genetics* **117**, 573–582.

Knott, S. A., Elsen, J. M. & Haley, C. S. (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93**, 71–80.

Lander, E. S. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**, 241–247.

Riquet, J., Coppieters, W., Cambisano, N., Arranz, J. J., Berzi, P., Davis, S. K., Grisart, B., Farnir, F., Mni, M., Simon, P., Taylor, J. F., Vanamanshoven, P., Wagenaar, D., Womack, J. E. & Georges, M. (1999). Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. *Proceedings of the National Academy of Sciences of the USA* **96**, 9252–9257.

Schmid, M., Nanda, I., Guttenbach, M., Steinlein, C., Hoehn, H., Schartl, M., Haaf, T., Weigend, S., Fries, R., Buerstedde, J. M., Wimmers, K., Burt, D. W., Smith, J., A'Hara, S., Law, A., Griffin, D. K., Bumstead, N., Kaufman, J., Thomson, P. A., Burke, T. A., Groenen, M. A. M., Crooijmans, R. P. M. A., Vignal, A., Fillon, V., Morisson, M., Pitel, F., Tixier-Boichard, M., Ladjali-Mohammedi, K., Hillel, J., Mäki-Tanila, A., Cheng, H. H., Delany, M. E., Burnside, J. & Mizuno S. (2001). First report on chicken genes and chromosomes 2000. *Cytogenetics and Cell Genetics* **94**, 169–218.

Schrooten, C. & Bovenhuis, H. (2002). Detection of pleiotropic effects of quantitative trait loci in outbred populations using regression analysis. *Journal of Dairy Science* **85**, 3503–3513.

Seaton, G., Haley, C. S., Knott, S. A., Kearsey, M. & Visscher, P. M. (2002). QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics* **18**, 339–340.

Sewalem, A., Morrice, D. M., Windsor, D., Haley, C. S., Ikeobi, C. O. N., Burt, D. W. & Hocking, P. M. (2002). Mapping of quantitative trait loci (QTL) for body weight at 3, 6, and 9 weeks of age in a broiler layer cross. *Poultry Science* **81**, 1775–1781.

Tuiskula-Haavisto, M., Honkatukia, M., Vilkki, J., De Koning, D.-J., Schulman, N. & Mäki-Tanila, A. (2002). Mapping of quantitative trait loci affecting quality and production traits in egg layers. *Poultry Science* **81**, 919–927.

Van Kaam, J. B. C. M. H., van Arendonk, J. A. M., Groenen, M. A. M., Bovenhuis, H., Vereijken, A. L. J., Crooijmans, R., van der Poel, J. J. & Veenendaal, A. (1998). Whole genome scan for quantitative trait loci affecting body weight in chickens using a three generation design. *Livestock Production Science* **54**, 133–150.

Van Kaam, J. B. C. H. M., Groenen, M. A. M., Bovenhuis, H., Veenendaal, A., Vereijken, A. L. J. & van Arendonk, J. A. M. (1999*a*). Whole genome scan in chickens for quantitative trait loci affecting growth and feed efficiency. *Poultry Science* **78**, 15–23.

Van Kaam, J. B. C. H. M., Groenen, M. A. M., Bovenhuis, H., Veenendaal, A., Vereijken, A. L. J. & van Arendonk, J. A. M. (1999*b*). Whole genome scan in chickens for quantitative trait loci affecting carcass traits. *Poultry Science* **78**, 1091–1099.

Viitala, S. M., Schulman, N. F., De Koning, D. J., Elo, K., Kinos, R., Virta, A., Virta, J., Mäki-Tanila, A. & Vilkki, J. H. (2003). Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle. *Journal of Dairy Science* **86**, 1828–1836.

Genet. Res., Camb. (2004), **83**, *pp.* 221–223.
Printed in the United Kingdom

Book Reviews

DOI: 10.1017/S0016672304216913

Introduction to Conservation Genetics. R. Frankham, J. D. Ballou and D. A. Briscoe. Cambridge University Press.

Published in 2002, this book is much, much more than an 'introduction' to conservation genetics, it is an in-depth treatment of the subject suitable (as the preface tells us) as a university textbook and for professionals in the field. After a preface and a couple of introductory chapters, it has two large sections on 'Evolutionary Genetics of Natural Populations' (182 pages; 7 chapters, including one specifically on small populations) and 'Effects of Population Size Reduction' (135 pages; 5 chapters), which together amount to a textbook in evolutionary population genetics, liberally illustrated with examples from endangered animals and plants or laboratory models, often the Frankham-Briscoe lab's own highly illustrative experiments with *Drosophila*. The final section, 'From Theory to Practice' (170 pages; 6 chapters) covers the practical application of the tools and principles discovered in the first two sections and ranges widely, from species definitions through uses of molecular markers to management advice in captive breeding. As in any good textbook, each chapter ends with further reading and problems, and at the back there is a list of take home messages, a set of revision problems, an extensive glossary, the reference list and the index. Truly this is a massive enterprise.

Rightly, the book places quantitative genetic variation, especially for reproductive fitness, firmly at centre stage in conservation genetics, and its main business is to consider the consequences of population size (and change) for this kind of variation. Molecular markers, especially non-expressed DNA markers, have many useful roles in conservation genetics documented here, but we should not (and here do not) lose sight of the central concern, which is expressed variation. The clarity of presentation of many population genetics issues is excellent. Although the treatment becomes reasonably mathematical at times, to a greater extent than most textbooks, this one delivers worked examples of even the simplest formulae, and line-by-line algebra, making it very accessible to the mathematically challenged. The authors are also appropriately candid about uncertainties, for example in chapter 13, when writing about the mutational meltdown hypothesis and the effective population size required to avoid inbreeding depression; and they are appropriately dismissive, for example in chapter 12 when writing about fluctuating asymmetry as a way of detecting inbreeding depression. In fact sections 1 and 2 struck me as an excellent teaching resources for university-level evolutionary genetics, with the added benefit that undergraduates will be attracted to a textbook that is so overtly about conservation.

I found few serious issues to argue about in the text, which, considering the scale of the enterprise, is remarkably free of errors. The errors that have been spotted so far can be found on the book's web page at http://consgen.mq.edu.au/.

The main criticisms one can make of this book concern its organisation and presentation. In their preface, the authors say that the organisation of material has been arrived at from teaching experience, and that some repetition results from trying to make each chapter the basis for a free-standing lecture. However, as a reader I found the extensive revisiting of topics and examples trying. Part of the problem comes from the separation of the two first sections, because events in small populations are described in a section 1 chapter, but then revisited in section 2 as events in *declining* populations. In consequence, by the time we reach the main treatment of inbreeding and inbreeding depression in section 2, we have already read quite a lot about them, both in an introductory chapter and in section 1. Furthermore, the main discussion of selfing in plants, a normal behaviour for some 20% of plants, ends up in section 2 about declining populations, which seems strange. Similarly, other topics such as effective population size and genetic rescue of inbred populations by supplementation are covered in more than one place. The organisation also leads to extensive revisiting of the same cases studies in different parts of the book. To give some extreme examples: the greater prairie

chicken and the northern hairy-nosed wombat appear ten times, the California condor 16 times, and the golden lion tamarin 17 times. I can't help feeling that a combination of more restraint on picking examples and more thorough treatment of some of these major case studies, bringing all the information to one place, would have brought economy of effort (writing and reading).

The production style adds to the sense of scattered facts. The main text is mostly in paragraphs, but regularly breaks out into lists of bullet points and equations. Aside from the main text, there are numbered boxes, examples, figures and tables. There are also frequent line drawings of the organisms under discussion and marginal summaries of the adjacent text. The overall effect is to break up the flow somewhat, as one flits from text to one of the display items and back, being careful to remember whether one is looking for Box 12.1, Example 12.1, Figure 12.1 or Table 12.1 (they all exist).

Given these remarks, the recent publication of a much shorter primer associated with this book sounds like a very good idea indeed.

In terms of content, so few stones are left unturned that it seems churlish to point out any omissions, yet for such a thorough treatment, I do think there are a couple. Compared with the heavy working of some case histories such as the Northern elephant seal and the golden lion tamarin, I was surprised that the cheetah has such a low profile in this book. There is no thoroughgoing treatment of the arguments that have surrounded this species, which is a shame, since its relative lack of molecular variation led to several arguable inferences that need to be set straight, and these authors could have done it.

Similarly, in this determinedly apolitical book, the main omission for me is discussion of the value-for-money of conservation genetics. Despite the 617 pages here, I believe that some remarkably simple and cheap rules of thumb exist for practical genetic management of endangered populations. Indeed, the book comes up with several: To avoid inbreeding depression, keep N_e at >50; to retain evolutionary potential, keep N_e >500; N_e is likely to be around 0·1 of census population size for many organisms; captive breeding introduces problems of its own, so only use it as a last resort; minimise kinship of mates in captive breeding programmes (more can be found in the book's closing list of take-home messages). To what extent should scarce conservation funds be spent on conservation genetics, for example on molecular surveys or computer modelling of endangered populations, versus securing their habitat or understanding the ecological causes of decline? If conservation geneticists can lever additional funds specifically for their work, then great, but mainstream conservation money should surely be spent on habitat protection and understanding of ecology? I should have liked some discussion of this issue.

JOSEPHINE PEMBERTON
Institute of Cell, Animal and Population Biology
The University of Edinburgh

DOI: 10.1017/S001667230422691X

DNA: Changing Science and Society. Ed. T. KRUDE. Cambridge University Press. 2003. 193 pages. ISBN 0 521 82378 1. Price £25.00 (hardback).

To mark the 50th anniversary of the publication of Watson and Crick's model of the structure of DNA, Darwin College, Cambridge marked the event by holding a series of lectures by distinguished speakers to explore the impact of our understanding of DNA on contemporary science and society. This book comprises the eight essays based on these lectures, together with a short introductory summary by the editor, Torsten Krude. Its breadth illustrates how pervasive a topic DNA has become and how iconic its structure. Nevertheless whilst the public now accepts its use in forensic studies, for example, many remain wary of changes produced by genetic manipulation, not least because of an unwillingness by individuals, the press, and particularly the antagonists, to consider the problem at a finer level than is, for example, GM food safe or not, regardless of the insertion technique or construct.

The first chapter, by Aaron Klug, differs from the rest in dealing not with contemporary issues, but with the history of the discovery of DNA. It features in the crucial work on x-ray crystallography, particularly that of Rosalind Franklin, and how that can be interpreted, showing some of the well and less known photographs. It is not a simple read, but interesting and informative.

The remaining essays can be grouped into those which mainly consider current technical developments, perhaps with historical background, and those which mainly deal with some of the political and ethical issues that arise from developments in genetic technology. In the former are Alec Jeffreys on Genetic Fingerprinting, Svante Pääbo on Ancient DNA, Ron Laskey on DNA and cancer, Robert Winston on DNA and reproductive medicine, and Dorothy Bishop on Genes and language. They provide nice reviews, which would be useful for a person requiring background knowledge in some of the applications of DNA. The professional is likely to find most interesting that which he knows least about: in my case the genetics of language, but the discussion here is mostly limited to what determines the ability to construct language and an argument against Chomsky's view that grammatical structures are inherited.

The discussions on DNA, biotechnology and society by Malcolm Grant, who was Chair of the Agriculture and Environment Biotechnology Committee and on DNA and ethics by the philosopher Onora O'Neill seem to me well reasoned discussions. Not least both argue against the simplistic view that, for example, that decisions should not be based on a strong version of the Precautionary Principle, which are argued by some 'provides reasons for avoiding all GM technologies, indeed all new technologies, that *might* have bad consequences (O'Neill, p. 171)'. I was, however, disappointed that, in her discussion of individual's rights on their DNA she did not discuss the use of such information in life and health insurance.

The content is not, with few exceptions, highly technical and should be readily accessible to a broad audience. Also citations are not given to specific papers, but a short list of background reading is given in each chapter. Overall, this book is a diverse and enjoyable book, which I hope gets a broad readership.

WILLIAM G. HILL
Institute of Cell, Animal and Population biology
School of Biological Sciences
The University of Edinburgh

Genet. Res., Camb. (2004), **83**, *p.* 225. © 2004 Cambridge University Press
DOI: 10.1017/S0016672304006901 Printed in the United Kingdom

Books Received

Malaria Parasites: Genomes and Molecular Biology. Eds. A. P. Waters & C. J. Janse. Caister Academic Press. 2004. 546 pages. ISBN 0 9542464 6 2. Price £115 (hardback).

The Evolution of Population Biology. Eds. R. S. Singh & M. K. Uyenoyama. Cambridge University Press. 2004. 460 pages. ISBN 0 521 81437 5. Price £80.00 (hardback).

Animal Genomics. Ed. B. Chowdhary. Karger. 2003. 366 pages. ISBN 3 8055 7734 6. Price Eur113.50 (hardback).

Clone Being: Exploring the psychological and social dimensions. S. E. Levick. Rowman & Littlefield. 2004. 317 pages. ISBN 0 7425 2990 8. Price £21.95 (paperback).

Prion Biology and Diseases. 2nd Edition. Ed. S. B. Prusiner. Cold Spring Harbor Laboratory Press. 2004. 1050 pages. ISBN 0 87969 693 1. Price £100.00 (hardback).

Purifying Proteins for Proteomics: A Laboratory Manual. Ed. R. J. Simpson. Cold Spring Harbor Laboratory Press. 2004. 801 pages. ISBN 0 87969 696 6. Price £100.00 (paperback).

Ecological Genetics: Design, Analysis and Application. A. Lowe, S. Harris & P. Ashton. Blackwell Publishing. 2004. 326 pages. ISBN 1 4051 0033 8. Price £29.99 (paperback).

Mutants: On the Form, Varieties and Errors of the Human Body. A. M. Leroi. Harper Collins Publishers. 2003. 431 pages. ISBN 0 00 257113 7. Price £20.00 (hardback).